Functional outlier detection by means of h-mode depth and dynamic time warping
Alvaro Rollon de Pinedo, Mathieu Couplet, Bertrand Iooss, Nathalie Marie, Amandine Marrel, Elsa Merle, Roman Sueur

To cite this version:
Alvaro Rollon de Pinedo, Mathieu Couplet, Bertrand Iooss, Nathalie Marie, Amandine Marrel, et al.. Functional outlier detection by means of h-mode depth and dynamic time warping. Applied Sciences, 2021, 11 (23), pp.11475. 10.3390/app112311475. hal-02965504v2

HAL Id: hal-02965504
https://hal.science/hal-02965504v2
Submitted on 24 May 2021

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

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
FUNCTIONAL OUTLIER DETECTION BY MEANS OF H-MODE DEPTH AND DYNAMIC TIME WARPING

Álvaro Rollón de Pinedo\textsuperscript{1,2}, Mathieu Couplet\textsuperscript{1}, Bertrand Iooss\textsuperscript{1}, Nathalie Marie\textsuperscript{3}, Amandine Marrel\textsuperscript{3}, Elsa Merle\textsuperscript{4} and Roman Sueur\textsuperscript{1}

\textsuperscript{1} EDF R&D, 6 quai Watier, Chatou, 78400, France
\textsuperscript{2} Université Grenoble-Alpes, 621 Avenue Centrale, Saint-Martin-d’Hères, 38400, France
\textsuperscript{3} CEA, DES, IRESNE, F-13108, Saint-Paul-lez-Durance, France
\textsuperscript{4} LPSC Grenoble, 53 avenue des Martyts, Grenoble, 38000, France

contact email - alvaro.rollon-de-pinedo@edf.fr

Abstract

This paper deals with the problem of finding outliers, i.e. data that differ distinctly from other elements of the considered dataset, when they belong to functional infinite-dimensional vector spaces. Functional data are widely present in the industry and may originate from physical measurements or numerical simulations. The automatic identification of outliers can help to ensure the quality of a dataset (trimming), to validate the results of industrial simulation codes, or to detect specific phenomena or anomalies. This paper focuses on data originated from expensive simulation codes, such as nuclear thermal-hydraulic simulators, in order to take into account the realistic case where only a limited quantity of information about the studied process is available. A detection methodology based on different features, e.g. the h-mode depth or the Dynamic Time Warping, is proposed in order to evaluate the outlyingness both in the magnitude and shape senses. Theoretical examples are also used in order to identify pertinent feature combinations and showcase the quality of the detection method with respect to state-of-the-art methodologies of detection. Finally, we show the practical interest of the methodology in an industrial context thanks to a nuclear thermal-hydraulic use-case and how it can serve as a tool to perform sensitivity analysis on functional data.
Keywords: Functional Data Analysis, Functional Outlier Detection, Probabilistic Modeling, Computer experiments, Sensitivity Analysis

1 Introduction

Nowadays, the ever increasing capabilities of measurement, generation and storing of data have increased the interest of the scientific and industrial communities on what are known as functional data. In its most general form, the domain of functional data analysis deals with any object having the form of a function, regardless of its dimension. As an example, the most basic form these objects can adopt are one-dimensional real functions, which might represent the evolution of a physical parameter of interest over time. These data are normally generated through an actual empirical measure, or a simulation code.

The great interest of the functional data analysis is that it allows to take into account the intrinsic nature of the data, i.e. the underlying process that generates them is supposed to have certain restrictions of domain, regularity, continuity and so on. As the functional data are infinite-dimensional by nature, functional data analysis methods always rely on a dimension reduction technique, whether implicitly or explicitly. More precisely, it is the case in the context of classification (Chamroukhi and Nguyen 2019), clustering (Slaets et al. 2012), landmark research and registration (Ieva et al. 2011) of functional data. By providing more synthetic descriptors of functional observations, functional data analysis methods allow a more practical treatment of data thanks to the available multivariate tools.

The domain of functional data analysis can be traced back to the works of Grenander, (Grenander 1950), but the term itself was coined by Ramsay, (Ramsay 1982). Some of the most widely regarded references for the domain include the general works of Ramsay and Silverman (1997) or Ferraty and Vieu (2006), focused on non-parametric methods for functional data. Functional data can be found in a wide variety of contexts, among which we can mention: environmental sciences (Besse et al. 2005; Febrero-Bande et al. 2008),
medical sciences (Juang et al., 2017), economy (Sen and Klüppelberg, 2019) and others. In
the context of this paper, our work is motivated by the use of expensive simulation codes
(Santner et al., 2003; Roustant et al., 2010) that are commonly used in the nuclear industry
in order to complement the nuclear safety assessment reports (IAEA, 2003), and which
require specific statistical tools. These thermal-hydraulic simulators, such as the system
code CATHARE2 (Geffraye et al., 2011), are capable of simulating the evolution of the
essential physical parameters during a nuclear transient, and they are expensive to evaluate
(several hours per run) and analyze due to the complexity of the simulated phenomena
(highly non linear interactions between its parameters).

In general, in the context of nuclear safety, the main analyzed parameters are the scalar
values known as Safety Parameters, which are representative of the severity of an accidental
transient. However, the analysis of the whole evolution of the aforementioned physical pa-
rameters is a complex domain that has been subject to considerable research effort in recent
years, such as in the works of Nanty (2015) or Auder (2011). The analysis of outlying tran-
sients in these sets of simulations is useful in order to detect unexpected physical phenomena,
validate the simulations providing useful descriptors of the functional outputs of interest, or
quantify the extreme behavior or penalizing nature of specific subsets of those transients.

In spite of the diversity of domains, there are several common points of interests between
all of them. As an example, in the most usual case of one-dimensional functions, a typical
difficulty that arises in the pre-treatment phase of the data is the subject of the discretization
of the grid on which the data are observed. Without going into too much detail on the
subject, one dimensional curves are not normally observed in their entirety (since they are
intrinsically infinite-dimensional). Instead, they are indexed by some variable, normally the
time in the one-dimensional case, in a grid that can be homogeneous or not. Indeed, points
are not necessary equally spaced, as with biological measurements such as blood pressures,
or as with temporal data originated by a simulation code whose time step evolves accordingly
with the convergence impositions of the numerical solver. Naturally, the representation of the
underlying functional process is still a matter of discussion, and does not only depend on the spacing of each point in the time grid, but also on other notions such as the density of data. This quantity can differ by several orders of magnitude in the case of numerical simulators depending on the restrictions imposed to time steps in order to guarantee the convergence, as well as the presence or not of random noise etc. For a more profound analysis on the importance of these subjects the reader can refer to James et al. (2000).

In this paper we focus on a functional outlier detection technique, showing its sensitivity to both magnitude and shape outliers, and its ranking capabilities. The importance of this work is closely related to the data quality domain, since the presence of anomalous behaviors that might have been originated by measurement errors, non-convergence of algorithms or the existence of non-physical values of numerical simulators may produce spurious results when treating a dataset. As explained in Aggarwal (2017), when the underlying process that creates the data behaves unusually, it results in the creation of outliers, and therefore it may contain useful information about abnormal characteristics of the system. Finally, the particularities of functional data treatment with respect to multivariate ones require specific tools in order to have acceptable detection capabilities (Schmutz et al., 2020).

The functional outlier detection domain has gained relevance in recent years, and the methods for detecting these outliers do not cease to improve. However, the techniques can differ greatly between them. A large number of them rely on the notion of statistical depth (Febrero-Bande et al., 2008; López-Pintado et al., 2014; Nagy et al., 2017), whereas others rely on other dimensionality reduction methods, clustering or hypothesis testing on the coefficients of some functional decomposition (Barreyre et al., 2020), as well as graphical tools (Arribas-Gil and Romo, 2014). Naturally, all of these techniques showcase different detection capabilities, and they can be more adapted to the detection of a particular type of outliers.

In our study, we consider that the data correspond to independent and identically distributed (i.i.d.) realizations of a random process \( Z \), taking its values in a functional space \( \mathcal{F} \). In
practice, any realization \( z_i \) of the random variable \( Z \) can only be observed in a limited number of points in the domain, i.e., it is observed as the random vector \( (z_i(t_1), ..., z_i(t_p)) \), with \( \{t_1, ..., t_p\} \in \mathcal{T} \subset \mathbb{R} \). This data representation is not itself appropriate to the detection of outlying functions. Section 2 thus presents how to define and compute a modeling of data dedicated to the characterization of outliers. Based on this data representation, Section 3 presents our functional outlier detection methodology. Sections 4 and 5 provide some analytical and industrial application examples. The properties, capabilities and limitations of the methodology are discussed in Section 6.

### 2 Functional data representation and modeling

As mentioned in the introduction, functional data are elements of a functional space, typically functions defined on a continuous interval of \( \mathbb{R} \). Measuring, storing and analyzing these data is however realized by using numerical devices and computers, and hence impose a digital representation of data. For this reason, the value of functions is available only on a finite-dimensional sub-domain of their theoretical definition domain, that is in a discretized version. When dealing with time dependent physical quantities, this discretization basically consists in a projection on a time grid which can be regular or not. To achieve an efficient detection of outlying functions, a transformation of the dataset is required, which brings out the features of the data that discriminate outlying ones from the rest of the set.

#### 2.1 Discussion and selected approach

Functional outliers are commonly classified into magnitude and shape outliers. Magnitude outliers may be defined by functions that deviate from the bulk of curves at some point in the definition domain of the functions according to some distance metric defined in their functional space (Dai et al., 2020). There exist numerous methods to visualize and detect them, some of them being based on depth measures, as exposed by Sun and Genton (2011).

5
The other main kind of outliers are shape outliers. This type of functional outlier is significantly more difficult to detect, but several techniques able to deal with them have been developed in recent years. We can mention Slaets et al. (2012) or Arribas-Gil and Romo (2014). Some of the challenges in the use of depth measures in this case are exposed in Nagy et al. (2017). Most functional outlier detection methods can be classified into one of the three following categories (Jacques and Preda, 2014):

- Two-stage approaches: the functional data are firstly projected into the considered functional space, in what is usually called the filtering step, and then a classical multivariate clustering procedure is applied on the coefficients of the expansion. In this case, if $\Phi = \{\phi_1, ..., \phi_r\}$ is the set of functions that forms a complete orthonormal basis of $\mathcal{F}$, any function $z_i$ of the space can be reconstructed from the sampled data through an expansion of the type $z_i = \sum_{j=1}^{r} a_j \phi_j$. In practice, we work with a finite number family of functions (a subset of $\mathcal{F}$), which induces a representation error, and that is commonly obtained by truncating an actual basis of $\mathcal{F}$. This allows to perform statistical hypothesis tests on the coefficients, which provides a detection criterion. An example of this approach can be found in (Barreyre et al., 2020).

- Non-parametric approaches: they are based on measures of proximity and dissimilarity between the functions. Multivariate clustering algorithms can usually be applied on these features (Dai et al., 2020).

- Probabilistic model-based approaches: they rely on the estimation of an underlying probability model, either on some non-parametric features applied to the curves, or on the coefficients of a basis expansion. An example of this approach applied to the coefficients of a functional Principal Components basis expansion can be found in the Ph.D. works of Nanty (2015), where the coefficients of the expansion are used to perform sensitivity analysis.

In our work, the detection procedure is based on the use of non-parametric measures and
the estimation of probabilistic models in order to reconstruct the joint probability density function of those features.

2.2 Features definition of outlying functions

Depth measures (Mozharovskyi, 2016) are a set of non-parametric features that have gained relevance in the functional outlier detection field in recent years (Cuevas and Fraiman, 2009).

Generally speaking, let \( z_1, \ldots, z_n \) be a set of objects observed in \( \mathbb{R}^p \) such that a random element \( Z \) describing the population is fixed, then a depth function is a mapping \( D(\cdot, Z) : \mathbb{R}^p \to \mathbb{R}^+ \) which provides a center-outward ordering of the data. The same definition holds for the case where \( p \to \infty \) for the functional framework. This functions are widely used for central tendency estimation, outlier detection and classification.

Some of the most widely used definitions of depth measures in the functional framework are developed below.

- Band depths. Let \( z_1, \ldots, z_n \) be a sample of functional data, then the basic definition of the Band Depth of a specific function \( z_i \) takes the form (López-Pintado et al., 2014):

\[
S_{n,J}^{(j)}(z_i, Z) = \sum_{j=2}^{J} S_n^{(j)}(z_i | Z), J \geq 2
\]  

such that:

\[
S_n^{(j)}(z_i, Z) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < \ldots < i_j \leq n} \mathbb{1}_{\{G(z_i) \subset B(z_{i_1}, z_{i_2}, \ldots, z_{i_j})\}}, j \geq 2
\]

with \( \mathbb{1} \) the indicator function. In this case, \( G(z_i) \) is the graph of the function \( z_i \), i.e. \( G(z_i) = \{(t, z_i(t)) : t \in T\} \), and \( B \) represents the band delimited by the \( j \) curves \( z_1, \ldots, z_j \). The parameter \( J \) restricts the maximum number of functions that delimit the bands. López-Pintado and Romo (2009) recommend the use of \( J = 3 \).
A more flexible definition of this depth notion is the Modified Band Depth, which consists in replacing the indicator function $1$ by a measure of the subset where the analyzed function is within the limits of the band. If $A_j(x) \equiv \{ t \in \mathcal{T} : \min_{r=i_1,...,i_j} z_r(t) \leq z(t) \leq \max_{r=i_1,...,i_j} z_r(t) \}$ is the mentioned subset, then the Lebesgue measure $\lambda$ of the subset, normalized by the measure of $\mathcal{T}$ provides a measure of how much time the considered function remains within the bands. Taking this into account, the Modified Band Depth can be expressed as:

$$\text{MBD}_n^{(j)}(z_i, Z) = \sum_{j=2}^{n} \left(\begin{array}{l} n \\ j \end{array}\right)^{-1} \sum_{1 \leq i_1 < ... < i_j \leq n} \lambda_r(A(z_i; z_{i_1}, z_{i_2}, ..., z_{i_j})), 2 \leq j \leq n.$$ (3)

Contrary to the basic Band Depth, the MBD is sensitive to functions that deviate from the center of the functions even if it is only for small subsets within the domain, which is naturally essential in the outlier detection domain.

Another widely spread definition of depth is the $h$-modal depth, proposed in [Cuevas et al. (2007)]. It employs the notion of a kernel function in order to estimate the centrality of the curve by taking into account the degree of immersion of a certain curve with regard to the curves that lie closest to the analyzed one according to some distance notion defined in the considered functional space.

The $h$-mode depth of a realization $z_i \in \mathcal{F}$ with respect to the distribution of $Z \sim P \in \mathcal{P}(\mathcal{F})$ is defined as:

$$hM(z_i, Z) = \mathbb{E}\left(\frac{1}{h} K\left(\frac{||z_i - Z||}{h}\right)\right),$$ (4)
which can be substituted by its empirical version (with a sample of \( n \) functional data):

\[
hM(z_i; Z_n) = \sum_{j=1}^{K} \left( \frac{1}{h} K \left( \frac{\|z_i - z_j\|}{h} \right) \right).
\] (5)

In this context, \( \| \cdot \| \) is a norm defined on \( F \), with no \textit{a priori} imposed limitations. \( K \) is a measurable kernel function \( K : \mathbb{R} \to \mathbb{R}^+ \) with \( h \) as the bandwidth parameter. The practical implementation of this depth notion consists in substituting the actual distribution \( P \) by its empirical version \( P^* \in \mathcal{P}(F) \).

This definition strongly depends on the choice of the norm and the bandwidth parameter. The authors give some orientations regarding this subject, proposing the \( L^2 \) and \( L^\infty \) norms, and taking \( h \) as the 15th percentile of the distribution of \( \|z_i - z_j\|, \forall z_i, z_j \in F \). For some results on the consistency of the h-modal depth, the reader can refer to Gijbels and Nagy (2015).

In addition to these depth notions, some other non-parametric features can be mentioned as they will help us characterizing functional data. The Time Series framework is significantly related to the functional data analysis domain, and also provides some useful metrics that can help to quantify the degree of similarity between ordered sequences. This is the case of the Dynamic Time Warping (DTW) algorithm, whose general form is presented below.

Given two sequences \( X := (x_1, x_2, ..., x_V), V \in \mathbb{N} \) and \( Y := (y_1, y_2, ..., y_W), W \in \mathbb{N}, \) as well as a feature space \( S \), and \( x_v, y_w \in S \) for \( v \in [1 : V] \) and \( w \in [1 : W] \), we can define a local cost measure (sometimes also called local distance measure), which is an application:

\[
c : S \times S \to \mathbb{R}^+.
\] (6)

In this case, an \((V, W)\)-warping path is a sequence \( p = (p_1, ..., p_L) \) with \( p_l = (n_l, m_l) \in [1 : V] \times [1 : W], \forall l \in [1 : L] \) which also satisfies the following conditions:
• boundary condition: \( p_1 = (1, 1) \) and \( p_L = (V, W) \),

• monotonicity condition: \( v_1 \leq v_2 \leq ... \leq v_L \) and \( w_1 \leq w_2 \leq ... \leq w_L \),

• step size condition: \( p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\} \) for \( l \in [1 : L-1] \).

The total cost \( c_p(X, Y) \) of a given warping path is

\[
c_p(X, Y) := \sum_{l=1}^{L} c(x_{vl}, y_{wl}).
\]  

Finally, an _optimal warping path_ between \( X \) and \( Y \) is a warping path \( p^* \) having minimal total cost among all possible warping paths. The _DTW distance_ between \( X \) and \( Y \) is then simply defined as the total cost associated with the optimal warping path.

As the DTW might be expensive to evaluate, it is worth pointing out the existence of some accelerated versions of the algorithm reducing this cost when the number of sampling points is too high. Many of them are based on the restriction imposed to the set of acceptable \((V, W)\)-warping paths (so that the whole cost matrix is not needed), by introducing weight functions that privilege certain specific paths, or by modifying the step-size condition [Tavenard and Amsaleg, 2013].

### 2.3 Probabilistic modeling of features

The main objective of this work is to develop a novel functional outlier detection technique which is as general as possible, and sensitive to the main types of outliers that are usually found in the industrial domain (i.e. shape and magnitude outliers). The first problem that we may encounter when setting such an objective is firstly the lack of a complete and indisputable definition of what constitutes an outlier in a set of data. Considering the definition provided in [Dai et al. (2020)] as data that behave _in an abnormal way_ with respect to the other considered objects, this approach requires the definition of what an abnormality is, and it is usually quantified as the extremal values of a measure that is sensitive to the searched outliers.
A more general (but more difficult to apply) definition of what constitutes an outlier is a subset of data that has been generated by a different process that the majority of data present in the considered set (Hawkins, 1980). As an example, a set of measurements could have a small amount of incorrect data points (measurement errors) that may not be obvious at first (these data are not generated the same way as the others). This can also happen in the simulation domain. Simply changing the compilers, computers or the version of simulation codes can significantly change the outcome of any physical simulation. Finding these abnormalities is fundamental in order to ensure the quality of any dataset.

Let us suppose that a certain number of features are available to describe our functional data and are able to capture the specific characteristics of both central and abnormal observations. If $U = \{u_1, ..., u_r, ..., u_R\}$ represents this set of features, with no imposed a priori restrictions on its size, such that $\forall u_r \in U, u_r : F \rightarrow \mathbb{R}$, then it would be possible to quantify the anomalous behavior according to each measure through the extreme-value analysis theory. The generalization of this theory is based on the use of probabilistic models that can be adjusted to the data. Generally, these models are generative, i.e., they are based on the estimation of the probability of occurrence of a data point (multivariate features in our case) accordingly with an assumed underlying model. Once a parametric family has been chosen for the generative model, its values must be estimated through an optimization algorithm.

The use of joint multivariate probabilistic models also has the advantage of providing a tool able of taking into account the interaction between the different features used to evaluate the dataset, in addition to providing a score of outlyingness related to a probability of occurrence. When the underlying process that generates the data is unknown, the use of Gaussian Mixture Models (GMM) (Reynolds, 2009) is practical due to the vast existent knowledge of these models. Assuming that $R$ descriptive features are available, the form of the associated $R$-dimensional multivariate Gaussian mixture density function of the random vector $\mathbf{u} \in \mathbb{R}^R$
is

\[ p(u) = \sum_{k=1}^{K} \omega_k f_k(u; \mu_k, \Sigma_k), \tag{8} \]

where \( f_k \) represents each single Gaussian multivariate probability density function, \( \mu_k \in \mathbb{R}^R \) represents its vector of means, and \( \Sigma_k \in \mathbb{R}^R \times \mathbb{R}^R \) is the corresponding covariance matrix. The weight of each individual density among the \( K \) components is represented by \( \omega \in \mathbb{R}^K, \sum_{k=1}^{K} \omega_k = 1 \) and \( \omega_k > 0 \ \forall k \in \{1, \ldots, K\} \) and can be interpreted as the mixing probabilities of the components.

### 2.4 Adapting GMM for outlier detection

The use of probabilistic models in lower-dimensional feature spaces is useful in order to detect anomalous points assuming that they provide a good fit for the data and are able of capturing the central tendency of the data. However, their basic use suffers from well-known spurious effects in the outlier detection domain [Aggarwal (2017)]. The main problems are:

- If the probabilistic model is adjusted taking into account the presence of outliers, they may bias the estimation of the underlying model. This is especially problematic if the outliers are assumed to be generated by a different distribution than the other data and are no only considered to be extreme realizations of the same underlying process than the others. On top of that, if the sample presents a high degree of contamination or the sample is small, this bias can greatly influence the detection.

- If the multivariate sample can be classified in several different clusters but they number of components is not well-chosen, the possibility of overfitting the probabilistic model to the data becomes a real problem. In this case, some small-sized clusters may appear overly adjusted to the outliers, which will not be identified as such.

These reasons motivate the modifications of the Expectation Maximization (EM) algorithm used in order to fit the GMM to the sample of data in the considered feature space, so that
these problems may be efficiently managed and the probabilistic model can be appropriate
for outlier detection. Let \( \{z_1, \ldots, z_n\} \) be an i.i.d. sample of functional data, with \( \{u_1, \ldots, u_n\} \)
being the set of associated features for each curve, the objective is the estimation of the
set of parameters of the GMM \( \{\omega, \mu_k, \Sigma_k\}_{k=1}^K \). This optimization problem is usually solved
by maximum likelihood estimation via the EM algorithm (Moon, 1996). The form of the
problem is:

\[
\max_{\omega \in \mathbb{R}^K, \{\mu_k, \Sigma_k\}_{k=1}^K} \sum_{i=1}^n \log \left( \sum_{k=1}^K f_k(u_i; \mu_k, \Sigma_k) \right). \tag{9}
\]

This well-known algorithm consists in maximizing the log-likelihood function of the Gaussian
mixture when a certain amount of binary latent variables \( g \) such that \( g_k \in \{0, 1\} \) and
\( \sum_{k=1}^K g_k = 1 \), and which allow to represent the corresponding \( k \)th gaussian component that is
considered each time. This way, the conditional distribution of \( p(u|g_k = 1) = f_k(u|\mu_k, \Sigma_k) \)
and \( p(u) = \sum_g p(g)p(u|g) = \sum_{k=1}^K \omega_k f_k(u|\mu_k, \Sigma_k) \). The last element required for the
estimation of the GMM parameters in each step are the conditional probabilities of \( g \) given
\( u \) (usually called responsibility that component \( k \) takes in explaining the observation). Their
values are:

\[
\gamma(g_k) = p(g_k = 1|u) = \frac{p(g_k = 1)p(u|g_k = 1)}{\sum_{j=1}^K p(g_j = 1)p(u|g_j = 1)} = \frac{\omega_k f_k(u|\mu_k, \Sigma_k)}{\sum_{j=1}^K \omega_j f_j(u|\mu_j, \Sigma_j)}. \tag{10}
\]

This way, the estimation of the parameters is performed in the following steps:

1. Initialize the values of the desired parameters \( \{\omega_k, \mu_k, \Sigma_k\}_{k=1}^K \),
2. **E-step**: Evaluate the current responsibilities with the previous parameter values,
3. **M-step** Re-estimate the parameters by using the responsibilities,
4. Evaluate the log-likelihood function presented in Eq. (9).

The modifications of the algorithm consist in iteratively check and reinitialize the estimated
parameters during the procedure in order to detect the undesired and spurious effects by adding two steps after the estimation of the likelihood function:

1. Check if $u_i = \mu_k$ for any $k \in \{1, ..., K\}$. The apparition of these singularities may maximize the log-likelihood function to infinity, since it is unbounded, and cause an overfitting of the data to isolated points. This phenomenon is well described in Bishop (2006). If this kind of anomaly is detected, the point is subtracted from the sample and the EM algorithm continues without it. Naturally, these points will be subject to close analysis since they are good candidates for potential outliers in the sample.

2. For each iteration step, $\omega_k$ can be viewed as the prior probability of $g_j = 1$, and $\gamma(g_j)$ can be seen as the posterior probability given the observed sample. If this posterior probability is considered too low (in our applications we shall take a value of 0.1 as the minimum weight of the mixing coefficients), we will consider that the corresponding component is either overfitting the data, or that it has detected a small subset of points which is not representative of the central trend of the data. In this case, the other calculated parameters of the components are kept and the values of means and covariances of the small cluster are reinitialized to a random value in the space.

These modifications allow for the GMM that is fitted to the sample in the feature space to stay general, and provide good candidates of outliers during the construction of the model, all of that while avoiding overfitting.
3 Outlier detection

3.1 Test for outlyingness

Thanks to these elements, we can construct a hypothesis test by making use of the components of the GMM adjusted model. This way, for any $u_j \in \mathbb{R}^R$:

$$H_0 : u_j \text{ has been generated by } f_k \text{ with probability at least,}$$

$$H_1 : u_j \text{ is an outlier.}$$

$$p(u_j|g_k = 1) > p_{\alpha k},$$

where $p_{\alpha k} = f_k(u_\alpha|\mu_k, \Sigma_k)$ such that $\mathbb{P}(u_\alpha|g_k = 1) \geq \alpha$. The set of data points less likely than $u_\alpha$ is composed by points which verify:

$$(u - \mu_k)^T \Sigma^{-1}(u - \mu_k) \geq (u_\alpha - \mu_k)^T \Sigma^{-1}(u_\alpha - \mu_k).$$

And therefore, $\mathbb{P}[(u - \mu_k)^T \Sigma^{-1}(u - \mu_k) \geq (u_\alpha - \mu_k)^T \Sigma^{-1}(u_\alpha - \mu_k)] = 1 - \mathbb{P}[L \leq (u_\alpha - \mu_k)^T \Sigma^{-1}(u_\alpha - \mu_k)]$, where $L$ follows a Chi-squared distribution, $L \sim \chi^2(k)$. By performing this test over all of the points considered in the feature space, we have a unique criterion for outlier detection, such that the outlying points will be the ones presenting p-values under a certain threshold $\alpha, \forall f_k$.

3.2 Ordering score in the Feature Space

Once all of the parameters are fixed at their estimated values, it is possible to quantify the extremal behavior of any point in the considered multivariate space through the notion of Level Set [Ahidar-Coutrix, 2015]. Formally, given an absolutely continuous probability measure $\nu$ with density $p$, the Minimum Level Set $D^\theta$ is the set of minimal volume (according to the Lebesgue measure) whose measure is $\theta$:

$$D^\theta = \arg\min_{D \subset \mathbb{R}^H, \nu(D) \geq \theta} \lambda(D).$$
where $\lambda$ is the Lebesgue measure in $\mathbb{R}^R$. Due to the concave downward nature of the GMM (Hosseini and Sra, 2015), the set $D$ is unique and $\nu(D^\theta) = \theta$. This way, a probabilistic score of outlyingness for functional data can be obtained via the probability mass retained in the associated Level Set of each functional datum $z_i$:

$$\hat{\theta}_i = \int_{\mathbb{R}^R} \hat{p}(u) \mathbb{1}\{\hat{p}(u) \geq \hat{p}(u_i)\} \, d^R u$$

(14)

where $\mathbb{1}$ is an indicator function, and $\hat{p}$ is the adjusted probability density function of the GMM with the estimated parameters:

$$\hat{p}(u; \omega, \hat{\mu}_k, \hat{\Sigma}_k) = \sum_{k=1}^{K} \hat{\omega}_k \frac{1}{2\pi \sqrt{|\hat{\Sigma}_k|}} \exp \left( -\frac{1}{2}(u - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1}(u - \hat{\mu}_k) \right).$$

(15)

This integral must be solved by numerical methods. For $R = 1, 2$, there exists efficient software able to solve it, whereas several algorithms based on the Cholesky decomposition have been developed for the higher-dimensional cases (Genz, 2000). In the applications presented here we will limit ourselves to two components, orienting each one of them to magnitude or shape outlier detection.

The above presented level set notion gives an unambiguous way of defining to what extend an observation is unlikely to be observed when assuming it was generated in accordance to the same probability law as the rest of the dataset. The GMM provides in turn the required underlying probabilistic model allowing this definition. With the resulting outlyingness scores $\theta_i$, we have now available a properly constructed quantification tool for the degree of outlyingness of an observation. We can thus use common statistical approaches to implement a outlier detection procedure, such as considering the occurrence probability of a data point is too small when $\theta_i < \alpha$, where $\alpha$ is a significance level.

Finally, let us consider the more realistic case where the availability of data is actually limited, and the sample of functional data is too small to generate the GMM with an acceptable level of reliability (the convergence of the EM algorithm is not necessarily reached). This
is for instance the case for expensive industrial simulation codes, such as the mechanical
or thermal-hydraulic simulators. In this case, a natural extension of this idea for outlier
detection can be implemented via bootstrap resampling, [Efron and Tibshirani (1994)]. B
groups are formed by successively drawn with replacement in the original sample. This
way, the absence of data can be mitigated through the re-estimation of the GMM for each
bootstrap group. If for B bootstrap groups \( p_b(u) \) represents the GMM of the bth group, the
form of the (bootstrap) estimator of outlyingness would then be:

\[
\hat{\theta}_i = \int_{\mathbb{R}^R} \frac{1}{B} \sum_{b=1}^{B} \hat{p}_b(u) 1\{\hat{p}_b(u) \geq \hat{p}_b(u_i)\} \cdot d^R u
\]

(16)

Throughout this reasoning, the hyperparameter K (the number of components of the mixture
model) has been supposed to be fixed, but in practice, this is yet another input parameter
of the GMM that must be provided \textit{a priori} to the EM algorithm. Indeed, the actual
form of the model is significantly different depending on the number of components that
are considered. If that is the case, the use of an oversimplified mixture when modeling
complex multivariate distributions can induce incorrect conclusions about the distribution
of data, whereas an unnecessary increase in the number of components may lead to overfitting
problems, unacceptable computational costs or imprecise conclusions.

This question can be treated as a model-selection problem, and several metrics are available
in order to estimate an appropriate number of components depending on the sample. Some
examples are the Bayesian Information Criterion (BIC) ([Schwarz 1978]) or the Integrated
Completed Likelihood ([Biernacki et al. 2000]). In this paper, the selection of the number of
components is performed by means of the Bayesian Information Criterion:

\[
BIC = 2 \log(\hat{l}) - k \log(n)
\]

(17)

where \( \hat{l} \) represents the log-likelihood function for the GMM, \( k \) is the chosen number of
components and \( n \) is the sample size used for the estimation. The second term introduces a
penalty which depends on the number of components in order to mitigate overfitting effects.

3.3 Proposed detection algorithm

The practical implementation of the detection algorithm is presented below, and several clarifications will be made. Firstly, the approach does not include a projection of the considered functional data onto a functional expansion in order to homogenize the dataset. This preliminary step may be important in applications with heterogeneous time grids, if the data are observed with noise etc. Here, it will be supposed that the domain of the data is already uniform. Furthermore, this kind of projections are not necessary most times when estimating non-parametric features, such as $L^2$ norms.

Secondly, it is important to note that, in the ideal case, the estimation of the GMM should be made on the basis of a non-contaminated dataset, i.e., without outliers. These existent outliers should be separated before the estimation if the objective is to model the joint distribution of the desired features in order to avoid the introduction of a non-controlled bias. However, this knowledge is not available a priori, which justifies the trimming procedure of Algorithm 1. It is worth mentioning the iterative step. Even though it is not explicitly noted, the $K$ components that are retained depend on the chosen sample for each bootstrap group. The task of reestimating the parameters for each GMM in order to optimize the number of components is not computationally expensive for small values of $K$, which is appropriate for datasets that do not present a high degree of multimodality and to avoid overfitting the data. In our simulations, $K_{\text{max}} = 10$.

Another important remark on the algorithm is the extraction step, where the most outlying functions on the set according to the $\hat{\theta}_i$ are separated from the original dataset, and the procedure is re-applied to the remaining functions until a desired homogeneity level is reached. Naturally, in the case where a large number of samples is available, some extreme data realizations are bound to appear. So, regardless of the chosen value for $\alpha$, these significant
Algorithm 1: Functional Outlier Detection

Extraction of $\mathcal{Z} = \{z_1, \ldots, z_n\}, z_i \in \mathcal{F}$; Choose a base of features $\mathcal{U} = \{u_1, \ldots, u_R\}$ such as those from section 2.2. Set the significance level $\alpha$ and the number of bootstrap groups $B$; while $\exists \theta_i > \alpha$ do

- Estimate $u_j(z_i) \forall z_i \in \mathcal{Z}, u_j \in \mathcal{U}$; Estimate GMM parameters: $
\{\omega^b_k, \mu^b_k, \Sigma^b_k\}_{k=1}^{K}\forall b \in \{1, \ldots, B\}, \forall k \in \{1, \ldots, K_{\text{max}}\}$;
- Set $K = \arg\max_{k=\{1, \ldots, K_{\text{max}}\}} (BIC(k))$;
- Estimate $\theta_i, \forall z_i \in \mathcal{Z}$ defined by equation (16); Extract $z_i \in \mathcal{Z} | \theta_i < \alpha$ from the sample;
end

Data might be separated from the set even though they do not actually contaminate the set since, although unlikely, these data appear due to the nature of the underlying generative process. Separating (trimming) all of these data may introduce a bias in the estimation of the successive GMM. This is a known problem in the outlier detection domain. An example for the functional data case can be found in Febrero-Bande et al. (2008).

4 Analytical test cases

Following Dai et al. (2020), López-Pintado and Romo (2009) and Long and Huang (2015) the detection capabilities of the algorithm can be assessed via controlled simulation examples. Such simulations make it possible to check whether a method succeeds in recognizing different kinds of outliers, when these are purposefully inserted in the dataset. Thus, inserted outliers are constructed so as to mimic typical outlying data engineers are facing in their daily practice. Some common notations for the analytical models are summarized in Table 1.

4.1 Numerical test protocol

In all of the simulation experiments, there will be a total number of $n = 50$ curves, 49 of which are generated by a reference model, and one is the contaminating outlier. The functions are
defined in the interval $[0, 1]$, with a grid of 30 equally spaced points and $B = 10$ bootstrap groups. Using the previous notation for the DTW algorithm, $V = W = 30$ points. The description of the models for these control simulation tests is as follows:

- Model 1. $Z(t) = 4t + G(t)$ is the function generator for the reference set of curves. In this case the outliers follow the distribution $Z_o(t) = 4t + G(t) + 2I_{\{t_1 < t\}}$.

- Model 2. The reference model for the curve generation remains $Z(t) = 4t + G(t)$, whereas the outliers are now generated from the distribution $Z_o(t) = 4t + G(t) + 2I_{\{t_1 < t < t_1 + 3\}}$.

- Model 3. Here the reference model becomes $Z(t) = 30(1 - t)^{3/2} + G(t)$. The outliers are generated from $Z_o(t) = 30(1 - t)^{3/2} + G(t)$.

- Model 4. For this last case, we keep the reference model as it is for Model 1 and Model 2, but the outliers simply consist of the sole deterministic part $Z_o(t) = 4t$ (the Gaussian component is removed).

Let us remark that, compared to the latter references, the multiplicative factor of the indicator functions has been reduced in order to make the outliers less apparent. All the outliers considered in this simulation study constitute shape outliers, and in some cases, such as in the third model, magnitude outliers as well.

In all cases, a bivariate Gaussian mixture model is adjusted to a pair of selected features and the outlier detection procedure is applied thereafter. Four commonly used features in the functional data analysis framework will be considered:

- The h-mode depth [4].
- The Dynamic Time Warping (DTW) [7].
- The modified band depth (BD) [3].
Figure 1: Examples of the four analytical test cases. The blue curves correspond to the 49 ones that are generated from the main model, whereas the red one corresponds to the outlier.

- The $L^2$ norm, which is one of the most intuitive and widely used metrics that can be applied to functional data. It takes the form: $\|z(t)\|_2 = \left( \int_{\mathbb{R}} |z(t)|^2 dt \right)^{1/2}$.

The detection procedure is applied to $N = 100$ replications of each model. We shall use two scores in order to evaluate the quality of the detection procedure. The first one will naturally be the estimated $\theta$ values of the outlier in each model and replication. This parameter is directly linked to the probability of being more anomalous than the outliers if the model is correct. Therefore, the distribution of values of $\theta_i, \forall i \in \{1, ..., n\}$ constitutes an indicator of the detected outlying nature of the function.

The second score is the average ranking of the outlier with respect to the total population of data. Since the $\theta_i$ score provides an ordering of the anomalous nature of each element in the set of curves, so it is possible to rank the data accordingly to said metric. In industrial applications, this ranking can be followed by the engineer to analyze particular data (e.g. numerical simulations) from the most suspicious (potentially interesting) datum to less
suspicious ones.

The Table 2 and Figures 2 and 3 summarize the results for all the replications of the experiments for every specific couple of features, i.e., the six possible combinations of h-mode depth ($hM$), modified band depth ($BD$), Dynamic Time Warping ($DTW$) and the $L^2$ metric. The average ranks of the outlier in each model accordingly to each chosen pair of features are shown in Table 2.

![Boxplots](image)

Figure 2: Boxplots of the outlyingness score for all combinations of features in each model in the $N = 100$ replications. The Standard boxplot takes into account the whole distribution of $\hat{\theta}_i$ for all the replications of each experiment.

As one can see from the Table 2 and Figure 3, the features that show the highest detection capabilities are the ones that include at least the h-Mode depth or the DTW as a component of the considered Gaussian mixture model. In the case of the first two models, it is the combination of both features that yields the best detection results, whereas it remains close to the best result for the third and fourth models.
This result was expected, since the $L^2$ norm is a very general non-parametric measure which is probably not well suited for the direct application to the detection of anomalies in functional data, in spite of its usefulness for functional data characterization. The Modified Band Depth appears to be adapted for a quick detection of magnitude outliers, but not such a sensitive measure regarding shape outliers, which are far more complicated to define, identify and detect. That also explains why the scores for the third model are so high with respect to the others.

The presented scores can be used in order to compare different detection methods that could be based on identical features (multiple testing, use of level sets, functional boxplots...) as well as a tool to compare the usefulness of different features for a common detection on the
basis of a common detection algorithm.

In both cases (for the boxplots of the $\hat{\theta}_i$ and the rankings), it is possible to appreciate not only the absolute detection capabilities that were mentioned before, but also the relative dispersion of the data. This can also be interpreted as an indicator of robustness (which depends on the choice of features). When looking at figures 2 and 3, several aspects can be noted. The first obvious remark is that the detection capabilities for the third model are far superior to those of the others. This is explained by the fact that this is the only one that constitutes both a shape and magnitude outliers, which largely facilitates its detection, even for less sensitive measures such as the $L^2$ distance. Another interesting point is that for the first model, which is contaminated by a shape outlier, all of the best results are obtained by the combinations that employ the DTW metric. This is also coherent, since it is the feature that best takes into account the shape differences between the curves. Finally, when analyzing the results of the experiments, it can be concluded that the use of a joint model through the h-mode depth and the DTW provide not only the highest detection rates in general, but also the smallest dispersion out of all the possible combinations. This is mostly related to the fact that the DTW is the most sensitive feature when it comes to analyzing shape outliers (it is specifically designed to provide a measure of correspondence between sequences). These two features are the ones that will be retained for the industrial application in the next section.

4.2 Comparison with state-of-the-art methodologies

The detection algorithm for the selected combination of features is performed in order to compare it to other detection methods. The selected methodologies are succinctly presented below.
4.2.1 Functional Boxplots

Given a sample of functional data \( z \) defined in a time domain \( T \) indexed by the variable \( t \), the 50% central region can be defined as:

\[
C_{0.5} = \left\{ (t, z(t)) : \min_{r=1, \ldots, n/2} z_r(t) \leq z(t) \leq \max_{r=1, \ldots, n/2} z_r(t) \right\}.
\] (18)

This region can be interpreted as an analogous of the inter-quartile range for functional data, and it effectively corresponds to it pointwise. The whiskers of the functional boxplot can be computed by extending 1.5 times the pointwise extremes of the central region, such that the outliers are detected if they surpass the frontiers defined by these whiskers. The in-depth analysis of this method can be found in Sun and Genton (2011).

4.2.2 High-Density Regions

Introduced by Hyndman (2009), the method consists in regrouping the values of the functional data in the considered time steps in a matrix and performing the Karhunen-Loeve decomposition, obtaining the corresponding coefficients in a lower-dimensional feature space where the density of the components is estimated via Kernel Density Estimation (KDE). This way, the high density regions (HDR) can be defined as the regions such that:\[ \{ u : f(u) \geq f_\alpha \}, \]
and will correspond to the region with highest probability density function with a cumulative probability of \( 1 - \alpha \), which can impose a detection criterion.

4.2.3 Directional detector

As described in Dai and Genton (2019), let \( X \) be a stochastic process, \( X : T \to \mathbb{R}^p \), the Functional Directional Outlyingness is defined as:

\[
FO(X, F_X) \int_T ||O(X(t), F_{X(t)})||^2 w(t) dt
\] (19)
where \( w(t) \) is a weight function. This magnitude can be decomposed into two components, the Mean Directional Outlyingness (MO) and the Variation of Directional Outlyingness (VO). The detection algorithm is based on these quantities and the selection of cutoff values for inferred Mahalanobis distances based on standard boxplots.

### 4.2.4 Sequential Transformations

This algorithm from [Dai et al. (2020)] relies on the transformation of a wide diversity of shape outliers into magnitude outliers, much easier to detect through standard procedures. Given a sequence of operators defined in \( \mathcal{F} \) (the functional space that generates the considered data) \( \{ G_k \}, k = 0, 1, 2, \ldots \), the method consists in sorting the raw and transformed data into vectors of ranks for each observation. The vectors of ranks are sorted according to a one-side depth notion, such as the *extreme rank depth* for instance, and a global envelope is constructed, which allows the outlier identification.

### 4.2.5 Results

The results of the application of the algorithm are given for the previously used 4 models and different degrees of contamination. The experiments were simulated 500 times for a sample of curves of \( N = 100 \), and three different degrees of contamination of outlying curves: 1\%, 5\% and 10\% of outliers in the sample. The detection rates are summarized in the Table 3.

Firstly, we must note that the identification capabilities and rates are clearly reduced when the size of the outlying sample is increased. This reduction of the performance of any detection algorithm is logical, since higher degrees of contamination naturally pollute the functional sample, which increases the bias of the score that is used for outlier detection. In the same line, if the size of the outlying sample is considerable (10\% of outliers for instance), an argument can be made to defend that this sample might not be outlying, and that it simply corresponds to another mode in a hypothetical multimodal functional sample. This kind of phenomenon, as well as masking effects, are described in detail in [Aggarwal (2017)].
Looking at the results, we can appreciate that the performance of the proposed algorithm is indeed competitive and on par with existent methods, even for complex sets of functional data, such as Model 4. In this case, we can clearly appreciate how the inclusion of a measure specifically dedicated to the detection of shape differences allows the consistent detection of the outlier. This capability is especially significant when we compare it with the other methods, which prove to be unable to detect this kind of shape outlier. In the case of the widely used Functional Boxplots, this is to be expected since they are intended to detect magnitude outliers. Regarding the HDR method, its low detection capabilities in this case are due to the fact that the low-dimensional representation through robust Functional Principal Component Analysis is not sufficiently precise to capture the outlying nature of the straight line. It is indeed possible that retaining a higher number of modes in this case could allow better detection capabilities, but this procedure greatly increases the curse of dimensionality problem (even if this subject is not treated in the paper by Hyndman (2009)), and it does not allow visualization purposes.

It is clear that Model 3 (being the only pure magnitude outlier amongst the considered models) is the most simple and easy to detect and virtually any method can consistently detect this kind of outlier when the sample is not overly polluted. Methods which rely the most on the density of curves in the functional space and their trends is more vulnerable to the bias induced in the sample by the curves, as they tend to identify the proportion of curves that behave unusually as belonging to a different mode of curves instead of genuine outliers. In the case of the functional boxplots, this is to be expected since by construction they are dedicated to the detection of magnitude outliers, which is useful if the contamination of the sample is made by a wide variety of magnitude outliers, but not so much if those outliers have all been generated by a homogeneous family of curves. In the case of the HDR plots, the existence of a homogeneous sample of outliers generates a set of points in their two-dimensional feature space of principal component scores with a high density of data. In Models 1 and 2 the conclusions are similar (both models present a combination of slight
magnitude and shape outliers). Most methods do not showcase any robustness for such slight magnitude outliers, contrary to the presented algorithm. The main conclusion that can be extracted from these tests is that most methods struggle to find outliers when they are not apparent, as it is the case of the models presented here.

Finally, it must be mentioned that the Directional Detector is the most robust method when it comes to detecting the pure magnitude outlier presented in Model 3, as it is the least sensitive method to more contaminated samples. The main advantage of this methodology is its capability of finding outliers in multivariate functional data sets.

4.3 Ranking results

Finally, another advantage of the methodology presented in this paper is the ability to provide a scalar ranking criterion \( \theta_i \in [0, 1], \forall i \in \{1, ..., n\} \) for a sample of functional data. This is not only useful from a clustering or outlying detection perspective, but also from an exploratory analysis perspective. For instance, this kind of score can be used in order to perform sensitivity analysis on the functional data.

Depth measures are widely used in this setting, but some advanced techniques that have developed in recent years, such as the method presented in 4.2.4, which provides an efficient ordering of the data. Naturally, as it happened in the outlier detection setting, a good order measure should be capable of identifying a potential multimodality in the set of data, as well as handling the existence of magnitude, shape, or mixed outliers.

The ranking experiments are the same ones as the ones performed for outlier detection testing, with one outlier in the sample of 100 curves. The results are presented in Table 4. In this case, we can appreciate that more advanced ranking methods such as the one presented here and the one presented in Dai et al. (2020) provide a consistent ordering of the functional data. The results provided by these up to date methods show that for the sample of 100 curves, the introduced outlier is always found to belong to the 5% more outlying curves of the sample, and is frequently found to be the most outlying in simpler cases like Model 3.
Usual ranking techniques for functional data such as depth definitions fail to clearly identify the more outlying nature of the outlier in the sample, and cannot be reliably used as order measures in such homogeneous packages of univariate curves.

Finally, we can mention that the better ranking results provided by the Sequential Transformations algorithm with respect to the method presented here can be explained due to the nature of the chosen transformations. In particular, the use of what they call $D_1$ transformation in their paper, which consists in taking the first order derivative of the functional data (in their notation, $D_1[X(t)] = \frac{d(t)X(t)}{dt}$), is obviously appropriate in the case of Model 4, where the pure shape outlier differs mainly by the values of derivatives. This means that its superior ranking capabilities for this specific model cannot be generalized for any type of shape outlier, and both methods provide comparable ranking results.

5 Industrial test-case study

5.1 Presentation

In this section the outlier detection methodology is applied to a real industrial dataset of time-dependent numerical simulations. We consider a Intermediate Break Loss of Coolant Accident (commonly called IBLOCA or simply LOCA) in a nuclear power plant, simulated with the CATHARE2 code. CATHARE2 is a best estimate computer code capable of recreating the main physical phenomena that may occur in the different systems involved in nuclear reactors, in particular in the 900 MW French Pressurized Water Reactors (PWR). It embeds two-phase modeling to calculate the thermal-hydraulic behavior of the coolant fluid in the reactor.

A LOCA accident is originated by a breach in the primary circuit, which is designed to evacuate the heat generated by the nuclear core. The sudden loss of large quantities of coolant implies a fast increase of the water temperature nearby the nuclear fuel rods, due to the residual power generated by the core during the accidental transient. This power discharge
and the subsequent temperature elevation must be compensated by the injection of water through a dedicated safety system. This is supposed to ensure that fuel rods temperature would remain below the fusion point at all times. Hence, the main safety criterion in LOCA with regard to the confinement of the fuel concerns its Peak Cladding Temperature (PCT), that is its maximum cladding temperature over the duration of the LOCA (here, the time-dependent cladding temperature is the maximum cladding temperature for all the fuel rods, whatever its localization in the core).

The particular statistical model under study involves a large number of scalar input parameters \( P = 97 \), which specify all sorts of physical phenomena whose relative influences in the PCT are difficult to assess a priori. These input variables can be classified into various categories, such as i) initial conditions and limit states for the system (Primary pressure, starting thermal power, the primary pumps inertia...), ii) some parameters of specific physical models and correlations that are used (thermal exchanges between the components, friction between fluid phases or some geometric parameters of the installation), as well as iii) some scenario variables (existence of blockages in the heat exchangers, fuel use in its life-cycle or initial temperature of the safety water injection).

All of these scalar inputs of the simulation code are uncertain, and are hence represented by the random vector \( X = (X_1, X_2, ..., X_{97}) \), whereas the output of interest would be \( Y \), which is normally a critical safety parameter in this case, such as the aforementioned PCT.

The total number of simulations that can be performed is relatively limited for such a high-dimensional input vector, since each run of the code takes around one hour to finish. For this reason, the use of classical multivariate statistical techniques is not straightforward. This explains the widely spread use of space-filling design methods in order to maximize the coverage of the input space (Iooss and Marrel, 2019), as well as metamodeling techniques to better exploit the number of code runs available for the physical model. Briefly explained, space-filling designs try to optimally explore the space of input variables i.e., they establish criteria in order to better choose the analyzed points of this high-dimensional space.
In the case of nuclear transients, there can easily exist more than a hundred input variables. In the case of metamodels, they are mathematical approximations of more complex physical models that, despite showing a higher precision in their calculations, take a much higher computation time. The use of metamodels helps to improve (increase) the total number of available simulations, so that the results that are finally obtained are more statistically relevant.

In this context, the consideration of the whole functional output (the evolution of the maximum cladding temperature, whatever its location) is expected to provide a better insight on the physical phenomena that govern in the transient than the scalar value of the PCT alone. In our case, 1000 Monte Carlo runs of the code were launched, generating the set of curves that is presented in Figure 4 for the evolution of the maximum cladding temperature during the transient.

Figure 4: Examples of IBLOCA transients simulated with CATHARE2. Only 40 curves are displayed for clarity.
5.2 "Functional outlier detection"

The previously presented outlier detection technique is applied to this set of curves. Both
the h-mode depth and the DTW are the selected features in order to obtain the degree of
outlyingness of each functional datum. The curves presenting a degree $\hat{\theta}$ of outlyingness over
95% are shown in Figure 5.

Figure 5: IBLOCA transient curves presenting the highest degree of outlyingness (red) and
the least outlying curve (blue).

The first apparent result is that the main magnitude outlier is easily detected, since the
curve that acts as the upper envelope of temperature in most points of the domain is the
one presenting the highest value of $\theta$ ($\theta = 1.0$ actually in this case). This curve is not only
anomalous in the magnitude sense, but also in the shape one, much like the Model 4 that
was presented in the previous section (these are sometimes called phase outliers).

Two other magnitude outliers have been identified, and one of them is also a shape outlier,
presenting an anomalous peak of temperature after about 120 seconds of simulation (physical
time). The final main outlier is a pure shape one, remaining zones with high density of data
during the whole domain of simulation, but presenting two peaks of temperature. Specially
notorious is the first peak, which occurs around the 100 seconds of transient, in a time
interval that does not match the vast majority of curves.

5.3 Sensitivity analysis on outliers

In this kind of numerical simulations, the detection of outputs that present a globally anomalous behavior is of critical importance, and characterizing what are the physical phenomena which have an actual influence on it can have the same importance, if not more. A way of performing this analysis is to establish some kind of dependence measure between the inputs of the simulation code and the outlying score $\theta$. However, the high dimensionality of the problem and the possible correlations between the input variables of the code can make this a difficult task. The Hilbert Schmidt Independence Criterion (Da Veiga, 2015) can be a useful tool in this context in order to test the dependence between the scalar input variables of the code and the outlying score. This is a first step in order to understand which physical variables are the ones that actually influence the anomalous behavior of the outputs.

By performing statistical tests on the HSIC values of the couples $(X_i, Y)$ in the design of experiments it is possible to quantify their dependence. Without going into the technical details of the procedure (see De Lozzo and Marrel (2016)), the HSIC represents a dependence measure between both variables, and can be used in order to build a statistical test with null hypothesis: $H_0$: the variable $X_i$ and $Y$ are independent. The hypothesis is rejected if the associated p-value of the test is inferior to a significance level threshold $\alpha$. If $H_0$ is rejected, then the existence of a dependence structure between the input variable $X_i$ and the output exists.

In this case, we apply this measure in order to perform a Target Sensitivity Analysis, i.e., sensitivity analysis but quantifying the influence of the considered parameters in a restricted domain $\mathcal{M}$ of the possible output values ($\mathcal{M} \subset \Theta$ such that $\mathcal{M} = \{ \theta \in \Theta | \theta > 0.9 \}$). This application to the set of input data and the obtained values of $\theta$ yields a number of influential variables that are shown in Table 5.

All of these input variables can be considered to have an influence on the outlyingness
of the resulting simulated curve, which represent the evolution of the maximum cladding temperature during the transient. Variable $X_{16}$ represents the friction between the injected water in the primary circuit by the accumulators and the injection line. This parameters has been found to be influential in other similar studies since the compensation of the lost water during the transient is mainly guaranteed by this system (the accumulators), and therefore the line connecting it to the primary is of crucial importance. If this value of friction increases, the water flow will be reduced, with the consequent increase in the average temperatures of the fuel.

Regarding variables $38$, $45$ and $64$, they are representative of physical phenomena occurring in the Reactor Pressure Vessel (RPV) during the transient. They model respectively: the heat transfer coefficient between the nuclear fuel and the surrounding coolant; the increase in pressure drop due to the deformation of the nuclear fuel due to the thermal-mechanical stress, prevents the coolant from ascending easily to the top of the RPV; and the friction between the steam and water in the core during the reflood phase of the transient. These elements are relevant since their evolution greatly influences the rewetting dynamic of the fuel, and the heat extraction in the short-term phase of the transient.

Finally, variables $X_{62}$ and $X_{68}$ are representative of phenomena which occur between the steam and coolant water in the downcomer (the annular part which links the injection line of the accumulators and the nuclear core). This element is critical during the reflood process, which is why increases in the friction between the ascending steam and the descending water in this element are penalizing from a safety point of view. This is due to the fact that if the friction coefficient increases, it means that the momentum of the injected water will be reduced, and the core will take longer to be filled. Similar conclusions can be obtained for the heat exchange coefficient between both phases in the downcomer, since low values for this variable will imply lower heat extraction rates in the vessel.

The study of two-dimensional scatter plots between these input parameters and the outlying score $\theta$ could already prove to be useful in order to visualize how the inputs affect the outlying
nature of the functional outputs. An example that illustrates the idea is presented in Figure 6:

Figure 6: Scatter plots of two input variables of CATHARE2 and the outlyingness score $\theta$. The points correspond to the bivariate plot of the values of the selected variable and the corresponding $\theta$. The red dots correspond to the simulations that have been retained as outliers.

As it can be seen, outlying transients concentrate around specific subsets of the domain of the identified influential variables. These plots are useful in order to evaluate if the physical values that originate outlying simulations are physically coherent with their expected influence. In this particular case, for instance, lower values of friction should correspond to less penalizing and outlying configurations when the safety criterion is the Peak cladding temperature of the nuclear fuel. Therefore, the observed effect of this variable is actually not expected, and it corresponds to an anomalous effect in the coding of this particular transient that was later corrected by the engineers. In other words, the methodology and the analysis technique was capable of capturing not only extreme effects in the analyzed physical time-dependent variable (the Maximum Local Cladding temperature), but it was also capable of finding actual outliers, in the sense that those simulations showcase non-physical events in the particular modelling used in this study.
6 Discussion and conclusions

This paper has dealt with a fundamental branch of functional data analysis: the outlier detection problem. The main aspects to take into account when dealing with functional data or high-dimensional objects in general have also been developed, exposing its main challenges and advantages.

A new time-dependent outlier detection methodology based on the use of non-parametric features has been proposed, assessed with synthetic data, and illustrated on thermal-hydraulic simulations, aiming at capturing the outlyingness of the elements of any considered functional dataset both in the magnitude and the shape senses. This is done via the joint utilization of the h-mode depth and the Dynamic Time Warping, and by defining outliers as data which do not belong to the minimum volume set of a chosen probability. The maximal probability for which a datum is not regarded as an outlier anymore is used as a score \( \theta \) of its outlyingness.

An original detection algorithm has been proposed, effectively allowing the trimming of functional data. This methodology, based on the use of two features, benefits from the notion of level set in order to treat real industrial problems based on time-dependent data even if the available data are scarce. Several features have been compared in this framework on the basis of some toy examples and two scores related to the outlyingness of functional data. Based on the results of these application cases, both the notions of Dynamic Time Warping and the h-mode depth have proven their efficiency when compared to other features such as the \( \ell^2 \) norm and the Modified Band Depth.

Finally, the analysis of simulations of the thermal-hydraulic behavior of a nuclear reactor during a Loss Of Coolant Accident has been carried out to illustrate the benefits of the method. This was achieved thanks to the use of sensitivity analysis tools capable of accounting for the dependence of the input variables of the numerical simulator, Hilbert Schmidt Independence Criterion (HSIC).

Regarding the perspectives of this work, a primary objective would be the in-depth quantification of the causes for the detected anomalous characteristics of certain functions in
real physical cases. In the case of numerical simulators, the identification of the inputs of
the code that actually present an influence on the anomalous outputs can help engineers to
detect possible defects of the code or finding physical phenomena of interest. This is also
relevant in order to ensure the quality of the datasets that are used in the assessment reports
of critical systems such as nuclear power plants.

References


thesis, Université de Toulouse.


Auder, B. (2011). *Classification and modelling of computer codes functional outputs: appli-
cation to accidental thermo-hydraulic computations in pressurized water reactors (PWR).*

testing for outlier detection in space telemetries. *IEEE Transactions on Big Data*, 6:443–
451.


7 Appendix

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G(t)$</td>
<td>Centered Gaussian process of covariance function $\Sigma(t_1, t_2) = 0.3 \exp^{-\frac{</td>
</tr>
<tr>
<td>$Z(t)$</td>
<td>Functional random variable generating the main model</td>
</tr>
<tr>
<td>$Z_0$</td>
<td>Functional random variable generating the outliers</td>
</tr>
<tr>
<td>$T_I$</td>
<td>Random point uniformly generated in the definition domain of the function</td>
</tr>
</tbody>
</table>

Table 1: Description of the common parameters of the models

<table>
<thead>
<tr>
<th>Pairs of features</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD-DTW</td>
<td>48.663</td>
<td>41.272</td>
<td>49.621</td>
<td>42.376</td>
</tr>
<tr>
<td>BD-hM</td>
<td>41.342</td>
<td>39.067</td>
<td>49.833</td>
<td>43.643</td>
</tr>
<tr>
<td>DTW-L2</td>
<td>44.551</td>
<td>42.660</td>
<td>50</td>
<td>43.842</td>
</tr>
<tr>
<td>hM-L2</td>
<td>48.937</td>
<td>44.133</td>
<td>49.968</td>
<td>41.929</td>
</tr>
<tr>
<td>hM-DTW</td>
<td>49.225</td>
<td>45.154</td>
<td>49.852</td>
<td>42.343</td>
</tr>
<tr>
<td>BD-L2</td>
<td>44.254</td>
<td>41.418</td>
<td>49.944</td>
<td>43.672</td>
</tr>
</tbody>
</table>

Table 2: Average rankings of the outlier for each analytical model and combination of features.
Table 3: Performances of the different algorithms on the test models. The results are expressed as a percentage (detection rates). DO: Directional Detector; FB: Functional Boxplots; HDR: High-Density Regions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100, p=1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DO</td>
<td>100.00</td>
<td>96.94</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>FB</td>
<td>59.26</td>
<td>39.51</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HDR</td>
<td>2.33</td>
<td>0.00</td>
<td>100.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100, p=5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DO</td>
<td>91.14</td>
<td>96.79</td>
<td>99.17</td>
<td>97.50</td>
</tr>
<tr>
<td>FB</td>
<td>58.23</td>
<td>54.40</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HDR</td>
<td>2.53</td>
<td>4.18</td>
<td>11.95</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100, p=10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DO</td>
<td>81.50</td>
<td>75.49</td>
<td>86.67</td>
<td>92.37</td>
</tr>
<tr>
<td>FB</td>
<td>47.25</td>
<td>45.97</td>
<td>99.63</td>
<td>0.00</td>
</tr>
<tr>
<td>HDR</td>
<td>0.75</td>
<td>1.71</td>
<td>7.41</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4: Average ranking of the outlier curve across the 100 replications of the experiments for the selected models. The Sequential transformations procedure is presented in Dai et al. (2020). The Modified Band Depth is presented in López-Pintado and Romo (2009), and the standard Integrated Depth appears in Cuevas and Fraiman (2009). In this case, the closer the value of a method is to 100, the more outlying it will be according to the corresponding ranking measure.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{16}$</td>
<td>Friction between the water and the discharge line of the accumulators</td>
</tr>
<tr>
<td>$X_{38}$</td>
<td>Global Heat Transfer Coefficient (HTC) in reflood fuel/coolant</td>
</tr>
<tr>
<td>$X_{45}$</td>
<td>Pressure drop to model the constrained flow due to the deformation of the fuel.</td>
</tr>
<tr>
<td>$X_{62}$</td>
<td>Friction coefficient between steam and water in the Downcomer during reflood</td>
</tr>
<tr>
<td>$X_{64}$</td>
<td>Friction coefficient between water and steam in core during the reflood phase</td>
</tr>
<tr>
<td>$X_{68}$</td>
<td>HTC between steam and water in the Downcomer</td>
</tr>
</tbody>
</table>

Table 5: Detected influential variables for $\theta \in \mathcal{M}$. The variables are not the actual values of the physical parameters, but **multiplicative coefficients** that increase or decrease their importance in a scenario.