Contents

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 [6], or whose *scale* [9] differs. There exist numerous methods to visualize and detect them, some of them being based on depth measures, as exposed by [19].

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 [18] or [1]. Some of the challenges in the use of depth measures in this case are exposed in [15]. Most functional outlier detection methods can be classified into one of the three following categories in the functional clustering domain [10]:

- 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 [2].
- 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 [6].
- 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 [16], 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 *Features definition of outlying functions*

Depth measures [14] are a set of non-parametric features that have gained relevance in the functional outlier detection field in recent years [5]. Generally speaking, let $z_1, ..., z_n$ 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 $z_1, ..., z_n$ be a sample of functional data, then the basic definition of the Band Depth of a specific function z_i takes the form [11]:

$$S_{n,J}(z_i, Z) = \sum_{j=2}^{J} S_n^{(j)}(z_i|Z), J \geq 2 \tag{S1}$$

  such that:

$$S_n^{(j)}(z_i, Z) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < ... < i_j \leq n} \mathbb{1}_{\{G(z_i) \subset B(z_{i_1}, z_{i_2}, ..., z_{i_j})\}}, j \geq 2 \tag{S2}$$

  with $\mathbb{1}$ the indicator function. In this case, $G(z_i)$ is the graph of the function z_i, i.e. $G(z_i) = \{(t, z_i(t)) : t \in \mathcal{T}\}$, and B represents the band delimited by the j curves $z_1, ..., z_j$. The parameter J restricts the maximum number of functions that delimit the bands. [12] 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 $\mathbb{1}$ by a measure of the subset where the analyzed function is within the limits of the band. If $A_j(x) \equiv \left\{t \in \mathcal{T} : \min_{r=i_1,...,i_j} z_r(t) \leq z(t) \leq \max_{r=i_1,...,i_j} z_r(t)\right\}$ is the mentioned subset, then the Lebesgue measure λ 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:

$$MBD_n^{(j)}(z_i, Z) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \leq i_1 < ... < i_j \leq n} \lambda_r(A(z_i; z_{i_1}, z_{i_2}, ..., z_{i_j})), 2 \leq j \leq n. \tag{S3}$$

  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 [4]. 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:

$$hM(z_i, Z) = \mathbb{E}\Big(\frac{1}{h} K\Big(\frac{\|z_i - Z\|}{h}\Big)\Big), \tag{S4}$$

  which can be substituted by its empirical version (with a sample of n functional data):

$$hM(z_i; Z_n) = \sum_{j=1}^{K} \Big(\frac{1}{\hat{h}} K\Big(\frac{\|z_i - z_j\|}{\hat{h}}\Big)\Big). \tag{S5}$$

In this context,$\|\cdot\|$ is a norm defined on $\mathcal{F}$, with no *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 $P^* \in \mathcal{P}(\mathcal{F})$.

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^∞ norms, and taking h as the 15th percentile of the distribution of $\|z_i - z_j\|, \forall z_i, z_j \in \mathcal{F}$. For some results on the consistency of the *h-modal* depth, the reader can refer to [7].

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 [3].

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 $\mathcal{Q}$, and $x_v, y_w \in \mathcal{Q}$ 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:

$$c : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}^+. \tag{S6}$$

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

$$c_p(X, Y) := \sum_{l=1}^{L} c(x_{v_l}, y_{w_l}). \tag{S7}$$

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 [20].

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 [6] 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 [8]. 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 $\mathcal{U} = \{u_1, ..., u_r, ..., u_R\}$ represents this set of features, with no imposed a priori restrictions on its size, such that $\forall u_r \in \mathcal{U}, u_r : \mathcal{F} \to \mathbb{R}$, 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) [17] 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

$$p(\mathbf{u}) = \sum_{k=1}^{K} \omega_k f_k(\mathbf{u}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{S8}$$

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

4 Pairwise comparisons of measures

In this section we showcase several tests performed on Models 1 to 4 through the combination of the considered couples of measures in the paper. The validation is performed on the basis of the outlyingness score and the ranking measures.

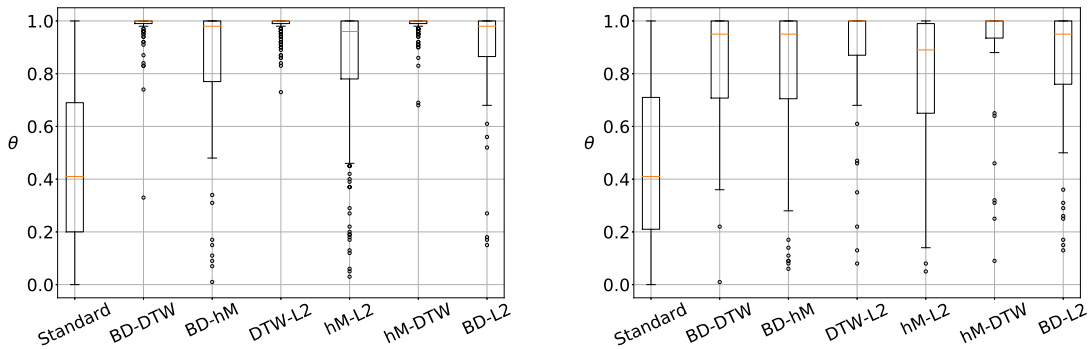
The Table S1 and Figures S1 and S2 summarize the results for all the replications of the experiments for every specific couple of features, i.e., the six possible combinations of h-mode depth (hM), modified band depth, dynamic time warping and the L^2 metric. The average ranks of the outlier in each model accordingly to each chosen pair of features are shown in Table S1.

As one can see from the Table S1 and Figure S2, 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.

Pairs of features	Model 1	Model 2	Model 3	Model 4
BD-DTW	48.663	41.272	49.621	42.376
BD-hM	41.342	39.067	49.833	43.643
DTW-L2	44.551	42.660	**50**	**43.842**
hM-L2	48.937	44.133	49.968	41.929
hM-DTW	**49.225**	**45.154**	49.852	42.343
BD-L2	44.254	41.418	49.944	43.672

Table S1. Average rankings of the outlier for each analytical model and combination of features.

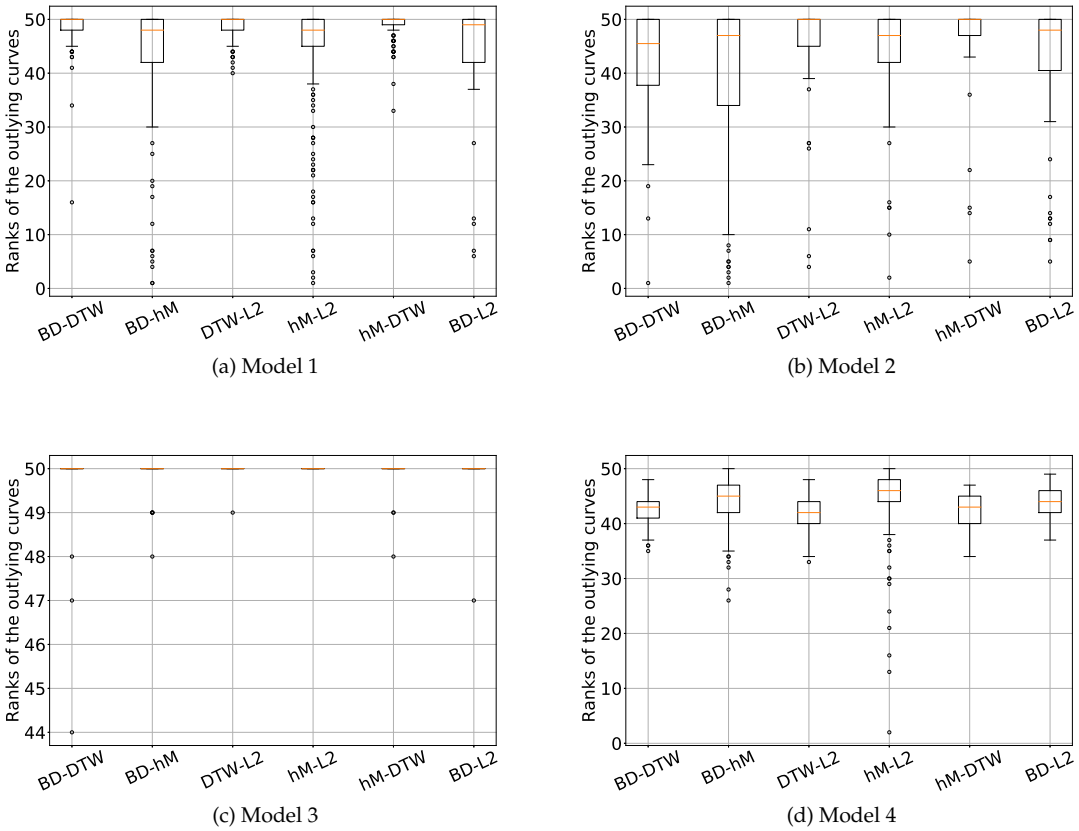


(a) Model 1 (b) Model 2

(c) Model 3 (d) Model 4

Figure S1. 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.

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 S1 and S2, 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



(a) Model 1 (b) Model 2

(c) Model 3 (d) Model 4

Figure S2. Boxplots of the ranking score of the outlier for all models over the $N = 100$ replications.

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).

5 Complementary comparisons with other models

In this section we compare the detection capabilities of our algorithm with those of the paper [13], where the authors also consider a sample of $n = 400$ realizations of a stochastic process. Different models of contamination (ranging from 1% to 10% of outliers), with magnitude and shape outliers are proposed. They also apply similar techniques of detection, of which the parametric one is similar to ours, but with the notable difference of considering the contamination rate (noted ν) already known. Slightly adapting their models to our notations, the $n = 400(1 - \nu)$ inliers (i.e., not outliers) are generated from the following model:

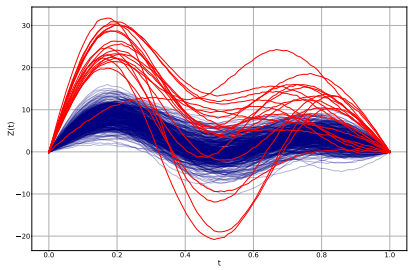
$$Z_l(t) = \sum_{j=1}^{4} \xi_j \sin(j\pi t) + \epsilon_l(t), \text{for } l = 1, ..., (1-\nu)n, \text{and } t \in [0,1]$$

where $\boldsymbol{\xi} = (\xi_1, ..., \xi_4)$ represents a multivariate normal random variable of mean $\mu_{\xi} = (4, 2, 4, 1)$ and diagonal covariance matrix Σ_{ξ} whose elements in the diagonal are $(5, 2, 2, 1)$, and $\epsilon_l(t)$ are independent autocorrelated random error functions.

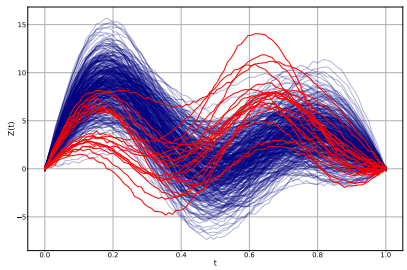
The outlying sample of data of size $n\nu$ with $\nu \in 1\%, 5\%, 10\%$, is generated from outlying function generators according to one of the following three scenarios:

- **Scenario A: magnitude outliers**.$Z_o(t) = \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \epsilon_l(t)$, for $l = 1, ..., n\nu$, and $t \in [0,1]$, where $\boldsymbol{\zeta}$ is a normally distributed random variable of mean $\mu_{\zeta} = 2.5\mu_x i$ and covariance matrix $\Sigma_{\zeta} = (2.5)^2 \Sigma_x i$.
- **Scenario B: shape outliers**. $Z_o(t) = \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \epsilon_l(t)$, for $l = 1, ..., n\nu$, and $t \in [0,1]$, where $\boldsymbol{\zeta}$ is a normally distributed random variable of mean $\mu_{\zeta} = (4, -2, 1, 3)$ and covariance matrix $\Sigma_{\zeta} = \Sigma_x i$.
- **Scenario C: magnitude and shape outliers**. The outliers are generated considering a proportion of $n\nu/2$ outliers from Scenario A, and $n\nu/2$ outliers from Scenario B.

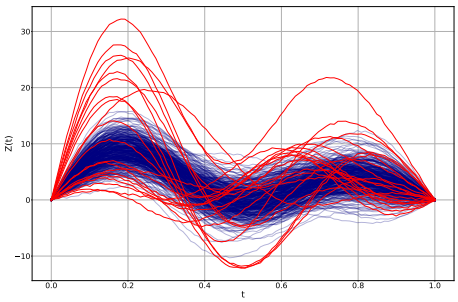
An illustration of the scenarios is shown in Figure S3, where the inlying curves are displayed with a higher degree of transparency to make the outliers more visible.



(a) Scenario A



(b) Scenario B



(c) Scenario C

Figure S3. Examples of the three considered scenarios for a sample of $n = 400$ curves with different degrees of contamination by the outliers (shown in red).

The detection rates (DR) and True Negative Rates (TNR), or specificity, of the algorithm presented in the present paper, as well as those of the considered methodologies (parametric and non-parametric) in [13] are displayed in Table S2. To be precise, let P be the total amount of outliers in the sample, and I the total amount of inliers, and let TP be the True Positive and TN the True Negative amounts detected by any considered method, then $DR = TP/P$ and $TNR = TN/N$ averaged over the replications.

		Scenario A			**Scenario B**			**Scenario C**		
Method	Metric	10%	5%	1%	10%	5%	1%	10%	5%	1%
Our algorithm	DR	93.41	98.54	96.00	51.63	56.21	93.04	74.71	96.67	98.04
	TNR	90.28	91.87	90.31	92.78	93.74	83.23	91.7	91.55	90.12
Entropy-parametric	DR	94.15	93.21	91.72	80.74	77.39	66.92	87.55	84.93	77.65
	TNR	99.35	99.45	99.92	97.86	98.81	99.66	98.62	99.21	99.77
Entropy-NonParametric	DR	92.72	91.50	89.05	74.21	77.14	71.25	87.22	85.81	79.77
	TNR	99.19	99.55	99.89	97.13	98.79	99.71	98.59	99.25	99.79

Table S2. Detection rates and true negative rates of the considered methods and our algorithm of detection.

As it can be seen, the detection rates of our algorithm are considerably higher on average when the sample is only slightly contaminated, whereas the detection rates become considerably lower for high levels of contamination. This trend is common in most unsupervised methods (see the comparisons with state-of-the-art methods in the main paper for example). This effect is partly due to the fact that if the number of outliers which follow a particular trend is too large, they might not be identified as abnormal realizations of the process, but rather as a different mode from the main set of curves. Shape outliers are particularly likely to suffer this effect, since in the considered scenarios they display similar shapes across the sample of outliers (see Figure S3 (b) or (c) for an illustration).

References

[1] Ana Arribas-Gil and Juan Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 03 2014.

[2] Clementine Barreyre, Beatrice Laurent, Jean-Michel Loubes, Loic Boussouf, and Bertrand Cabon. Multiple testing for outlier detection in space telemetries. *IEEE Transactions on Big Data*, 6:443–451, 2020.

[3] R. Bellman and R. Kalaba. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9, 1959.

[4] Antonio Cuevas, Manuel Febrero-Bande, and Ricardo Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22:481–496, 02 2007.

[5] Antonio Cuevas and Ricardo Fraiman. On depth measures and dual statistics. a methodology for dealing with general data. *Journal of Multivariate Analysis*, 100:753–766, 04 2009.

[6] Wenlin Dai, Tomas Mrkvicka, Ying Sun, and Marc Genton. Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics and Data Analysis*, 149, 04 2020.

[7] I. Gijbels and Stanislav Nagy. Consistency of non-integrated depths for functional data. *Journal of Multivariate Analysis*, 140:259–282, 05 2015.

[8] D.M. Hawkins. *Identification of Outliers*. Monographs on applied probability and statistics. Chapman and Hall, 1980.

[9] Mia Hubert, Peter J. Rousseeuw, and Pieter Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202, Jul 2015.

[10] Julien Jacques and Cristian Preda. Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, 71:92–106, 03 2014.

[11] Sara López-Pintado, Ying Sun, Juan K. Lin, and Marc G. Genton. Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8(3):321–338, Sep 2014.

[12] Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

[13] G. Martos, N. Hernández, A. Muñoz, and J.M. Moguerza. Entropy measures for stochastic processes with applications in functional anomaly detection. *Entropy*, 20(1), 2018.

[14] Pavlo Mozharovskyi. Tukey depth: linear programming and applications. *Preprint, arXiv:1603.00069*, 02 2016.

[15] Stanislav Nagy, Irène Gijbels, and Daniel Hlubinka. Depth-based recognition of shape outlying functions. *Journal of Computational and Graphical Statistics*, 26(4):883–893, 2017.

[16] Simon Nanty. *Stochastic methods for uncertainty treatment of functional variables in computer codes : application to safety studies*. PhD thesis, Université Grenoble Alpes, 2015.

[17] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA, 2009.

[18] Leen Slaets, Gerda Claeskens, and Mia Hubert. Phase and amplitude-based clustering for functional data. *Computational Statistics and Data Analysis*, 56:2360–2374, 07 2012.

[19] Ying Sun and Marc G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.

[20] Romain Tavenard and Laurent Amsaleg. Improving the efficiency of traditional DTW accelerators. *Knowledge and Information Systems*, 42:215–243, 01 2013.