Proceedings of Machine Learning Research 266:1–21, 2025 Conformal and Probabilistic Prediction with Applications

Conformal Anomaly Detection for Functional Data with Elastic Distance Metrics

Jason Adams JRADAMS@SANDIA.GOV
Sandia National Laboratories, Albuquerque, NM, USA
Brandon Berman BJBERMA@GMAIL.COM
Sandia National Laboratories, Albuquerque, NM, USA
Joshua Michalenko JJMICH@SANDIA.GOV
Sandia National Laboratories, Albuquerque, NM, USA
J. Derek Tucker JDTUCK@SANDIA.GOV
Sandia National Laboratories, Albuquerque, NM, USA

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

This paper considers the problem of outlier detection in functional data analysis with a particular focus on the more difficult case of shape outliers. We present an inductive conformal anomaly detection method based on elastic functional distance metrics. This method is evaluated and compared to similar conformal anomaly detection methods for functional data using simulation experiments. The method is also used in the analysis of a real exemplar data set that shows its utility in a practical application. The results demonstrate the efficacy of the proposed method for detecting both magnitude and shape outliers.

Keywords: Anomaly detection, conformal prediction, elastic functional data analysis

1. Introduction

Functional data are prevalent in many applications across a wide range of scientific domains (Ullah and Finch, 2013). As such, methods enabling the principled statistical analysis of functional data are important. The development and application of such methods have been rich areas of research for many years, and functional data analysis (FDA) is now a well-established branch of statistics (Ramsay and Silverman, 2005). This paper introduces a novel approach to detecting anomalous functional data. Our method leverages both the conformal prediction (CP) framework (Vovk et al., 2005) and the elastic functional data analysis (EFDA) framework (Srivastava and Klassen, 2016). Technical details regarding these frameworks are provided in Section 2.

Functional data vary continuously over some independent variable or variables. In this work, we only consider the case where all functional data are observed over a single independent variable. Within a given set of n observed functions, we assume that each function is observed over the same values of the independent variable. While this assumption is often not satisfied in practice, smoothing and interpolation methods can be used to adjust the observed data for this purpose. Although this is far from a trivial process, such methods for preprocessing functional data are outside the scope of this paper. We refer readers to Ramsay and Silverman (2005) as a starting point for smoothing and interpolation methods.

Throughout this paper, the terms *anomaly* and *outlier* are used interchangeably to describe an observation that does not fit well within a distribution. This is in line with the definition used by Aggarwal (2016) which says that "[a]n outlier is a data point that is significantly different from the remaining data." Within this definition, there are two distinct scenarios to consider. In the first scenario, one has access to an outlier-free dataset and seeks to compare a potentially contaminated data set with the pristine one. In the second only a single data set is available and it is desired to identify the most outlying observations. The method we introduce is most appropriately used in the first scenario, so that is the focus of this paper. Consideration of the second case is reserved for future work.

1.1. Related Work

Much previous work has been done in both functional outlier detection and conformal anomaly detection. This section reviews some of the most relevant methods and compares them to the present work.

1.1.1. Functional Outlier Detection

Numerous methods for functional outlier detection have been proposed in the literature. Many papers distinguish between *magnitude* and *shape* outliers. As described by Hyndman and Shang (2010), magnitude outliers "lie outside the range of the vast majority of the data" while shape outliers "may be within the range of the rest of the data but have a very different shape from other curves." Visualization methods are often sufficient to identify magnitude outliers, but shape outliers are typically much more difficult to identify.

Several approaches to functional outlier detection depend on visualization, aiming to extend standard univariate and multivariate techniques (e.g., boxplots) to the functional case. The primary challenge with such an extension is to determine a reasonable method for ordering functional data. A robust principal component (PC) analysis is used by Hyndman and Shang (2010) to reduce the dimensionality of the functional data. Using the first two PCs, depth and density measures are used to order the PC scores. From these orderings, three functional visualization tools — the rainbow plot, the bagplot, and the boxplot - are constructed and used to detect both magnitude and shape outliers. Similarly, Sun and Genton (2011) use the concept of band depth from López-Pintado and Romo (2009) to construct functional boxplots, and Huang and Sun (2019) introduce the notion of total variation depth to order functional data and construct visualizations for detecting both magnitude and shape outliers. Arribas-Gil and Romo (2014) propose visualization tools called outliergrams, which are constructed based on metrics from López-Pintado and Romo (2011). Noting that magnitude outliers are much easier to detect than shape outliers, Dai et al. (2020) apply several transformations to functional data so that shape outliers can be detected as magnitude outliers.

Other methods of functional outlier detection do not rely on visualization. Sawant et al. (2012) develop a robust principal component method and apply a multivariate outlier detection method from Hubert et al. (2005) to the PC scores. Similarly, Yu et al. (2017) represent functional data with a B-spline basis (Ramsay and Silverman, 2005) and use the minimum covariance determinant method (Rousseeuw and Driessen, 1999) on the basis coefficients to detect outliers. Both of these methods can be viewed as using standard

outlier detection methods on a reduced number of features extracted algorithmically from a set of functional data. In a different approach by Azcorra et al. (2018), three features are selected to measure the degree of outlyingness in terms of magnitude, shape, and amplitude respectively (note that this paper further divides what other papers call shape outliers into both shape and amplitude outliers). Thresholds are then set on these features to identify outliers.

Additionally, several methods for functional outlier detection within the EFDA framework have also been proposed. In the method proposed by Xie et al. (2017), elastic distance metrics (discussed in more detail in Section 2 in the present work) are used to construct functional boxplots. Measures of elastic depths are introduced by Harris et al. (2021) and used to identify shape outliers. While not explicitly presented as an outlier detection method, Tucker et al. (2020) introduced functional tolerance bounds that are constructed by bootstrapping and using the EFDA boxplots of Xie et al. (2017), and these tolerance bounds can be used to identify outliers.

While the present work does make use of the EFDA framework and can be used for detecting both magnitude and shape outliers, it differs from the above-mentioned methods in its reliance on the conformal prediction framework, which none of the others use. It also does not rely on either visualization or dimension reduction/featuriziation techniques.

1.1.2. Conformal Anomaly Detection

Within the conformal prediction literature, a number of papers have focused on the problem of detecting outliers. The method of conformal anomaly detection (CAD) was introduced by Laxhammar (2014). In a follow up work (Laxhammar and Falkman, 2015), the method of inductive conformal anomaly detection (ICAD) was proposed, and both of these focused on trajectory data. Since we will frame our proposed method as ICAD for functional data, more details on ICAD are given in Section 2. In the work of Cai and Koutsoukos (2022), ICAD is used to detect outliers in cyber-physical systems, such as autonomous vehicles, for the purpose of improving the control mechanisms of such systems. An adaptation of ICAD for univariate time series is proposed by Ishimtsev et al. (2017). A method for adjusting the outputs of ICAD algorithms to reduce the number of false positives (i.e., incorrectly labeling an observation as an outlier) is presented by Bates et al. (2023). The clearest distinction between these papers and the present work is that none of them are concerned with functional data.

1.1.3. Functional Conformal Anomaly Detection

To date, there have only been a small number of papers that use conformal prediction within a functional data context. In the work of Wang et al. (2025), EFDA methods are used with CP to provide bounds for partially observed functional data. Both Diana et al. (2023) and De Magistris et al. (2024) use CP for spatial functional data. CP has been applied to multivariate functional data and functional time series data by Diquigiovanni et al. (2022) and Ajroldi et al. (2023), respectively. However, two previous works focus on the functional outlier detection problem. First, Lei et al. (2015) introduce a CP approach to visualize classes of functional data. While the main focus of the paper is on visualization and data exploration, the method is also useful for outlier detection. In the work of Diquigiovanni

et al. (2021), a CP method is proposed specifically for the purpose of detecting outliers in functional data.

Given the close relationship between the methods presented by Lei et al. (2015) and Diquigiovanni et al. (2021) and the one we are proposing, we give a detailed explanation of them in Section 3 and employ them as benchmark methods in Section 4. We note that these methods were not originally presented as ICAD methods, but it is natural to frame them as such. Throughout, we will refer to the ICAD version of the method from Lei et al. (2015) as GMD (for Gaussian Mixture Density) and the ICAD version of the method from Diquigiovanni et al. (2021) as SNCM (for Supremum Non-conformity Measure). The primary difference between our proposed method and GMD and SNCM is that our method is based on the EFDA framework.

1.2. Contribution and Outline

Our primary contribution in this work is to introduce a novel ICAD method, based on the EFDA framework, for detecting functional data outliers. We conduct empirical evaluations of this method using simulated data and compare it to both GMD and SNCM. Our results demonstrate the efficacy of our method for functional anomaly detection.

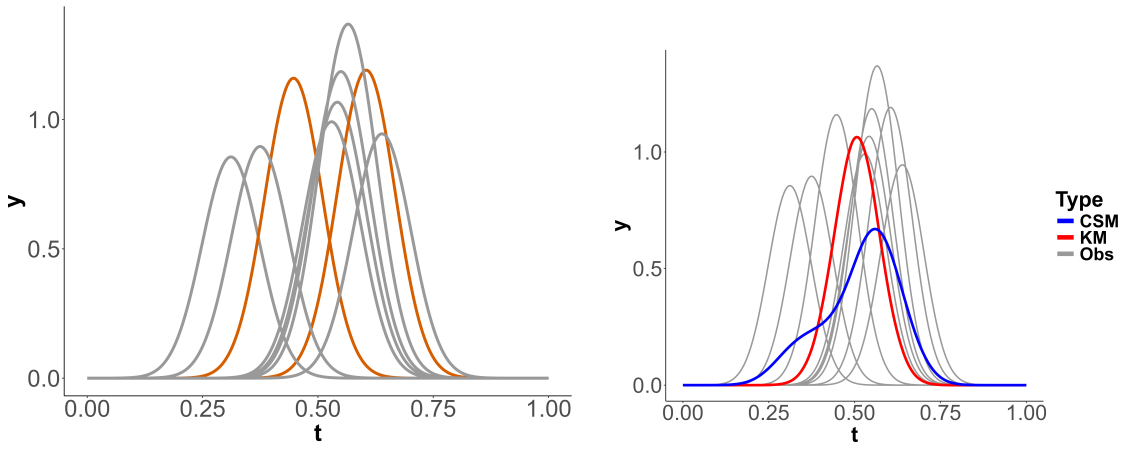
The outline of the paper is as follows: Section 2 provides necessary details regarding EFDA, CP, and ICAD to understand our method. In Section 3, we introduce our proposed method. Section 4 describes our empirical method comparison. Our method is demonstrated on a real data exemplar in Section 5, and we conclude in Section 6.

2. Preliminaries

2.1. Elastic Functional Data Analysis

In the analysis of functional data, there are two aspects of variability that must be considered. These are *phase* (or x-axis) variability and *amplitude* (or y-axis) variability. Figure 1 shows a sample of nine functional data observations that display both phase and amplitude variability. For instance, the two highlighted functions in Figure 1(a) vary considerably in phase but only very little in amplitude. Traditional functional data analysis often utilizes dimension reductions which can lead to a loss of fidelity and inability to detect outliers. The advantage of the EFDA framework is that it utilizes the entire function to create distance metrics that quantify the degree of separation between two functions with regard to both amplitude and phase. It also allows for estimating a Karcher mean, which is a better measure of center than a standard cross-sectional mean (Ramsay and Silverman, 2005) when functional data contain phase variability. In Figure 1(b), the grey curves are the same functional data as in panel (a), the red function is the Karcher mean, and the blue function is the cross-sectional mean. Note that in panel (b), the cross sectional mean, CSM bears little resemblance to the functional data whereas the Karcher mean is much more representative. Thus, the Karcher mean provides a better foundation for inference than the CSM.

The math underpinning the EFDA framework is quite involved, therefore we provide only the essential details for computing amplitude and phase distances as well as the Karcher mean for a set of functional data. We refer readers who are interested in more detail on the EFDA framework to Srivastava and Klassen (2016).



(a) (b)

Figure 1: (a) A small sample of example functional data. (b) The same functional data observations (OBS) with the Karcher mean (KM) and cross-sectional mean (CSM) overlaid.

Let f be a real-valued and absolutely continuous function over the domain $[0,1]$ and $\mathcal{F}$ be the set of all such functions[1]. As mentioned elsewhere (Tucker et al., 2013), the absolutely continuous assumption is not a restriction in practice because the observed data are always observed discretely. Let $\gamma : [0,1] \to [0,1]$ be a boundary preserving diffeomorphism with $\gamma(0) = 0$ and $\gamma(1) = 1$, and let Γ be the set of all such diffeomorphisms. We call γ a *warping function* as the composition of f and γ, $f \circ \gamma$, effectively warps the function f but maintains the original domain over which f is defined. Also, we denote the *square root slope function* (SRSF) of a function f by $q : [0,1] \to \mathbb{R}$ such that $q(t) = \text{sign}(\dot{f}(t))\sqrt{|\dot{f}(t)|}$ where the dot over a function indicates its derivative with respect to t. Similarly, we denote the SRSF of the warping function to be ψ. However, since $\gamma > 0$ and $\dot{\gamma} > 0$ for all $t \in [0,1]$, then the SRSF of γ is simplified to $\psi = \sqrt{\dot{\gamma}}$.

2.1.1. Elastic Distance Metrics

Let f_1, $f_2 \in \mathcal{F}$ and q_1, q_2 be their corresponding SRSFs. The amplitude distance between f_1 and f_2 is defined as

$$d_a(f_1, f_2) = \inf_{\gamma \in \Gamma} ||q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}}|| \tag{1}$$

where $||\cdot||$ represents the functional $\mathbb{L}^2$ metric[2]. The warping function, γ, which optimizes the distance in Equation 1 is typically identified in practice through the Dynamic Programming algorithm (Bertsekas, 2012). Intuitively, the optimal warping function, γ, not only minimizes the distance from q_2 to q_1, but also effectively aligns f_2 to f_1 as $(q_2 \circ \gamma)\sqrt{\dot{\gamma}}$ is the

1. In practice, the function f can be defined over any closed interval of the form $[a,b] \subset \mathbb{R}$. This is typically done by simply rescaling the interval $[0,1]$ to $[a,b]$.
2. Note that this is the Fisher-Rao metric and a proper distance. If we compute this distance using f directly, then it is not a proper distance and hence the reason that the SRSF transformation is used.

SRSF of $f_2 \circ \gamma$. For all EFDA computations herein, we use the `fdasrvf` package (version 2.3.4) (Tucker, 2017) within the `R` programming language (R Core Team, 2023).

For two warping functions, γ_1, $\gamma_2 \in \Gamma$, the distance between them is computed as

$$d_\gamma(\gamma_1, \gamma_2) = \cos^{-1}\left(\int_0^1 \psi_1(t)\psi_2(t)dt\right) \tag{2}$$

where ψ_1 and ψ_2 are the SRSFs of γ_1 and γ_2, respectively. In order to find the phase distance between two functions, $f_1, f_2 \in \mathcal{F}$, we first find the optimal warping function from f_2 to f_1. Again, denote this function as γ. Next the warping function from f_1 to itself is simply the identity function ($\gamma_I(t) = t$), and the SRSF of the identity function is the constant function equal to one ($\sqrt{\dot{\gamma}_I(t)} = 1$). Thus, the phase distance between f_2 and f_1 simplifies to

$$d_p(f_1, f_2) = d_\gamma(\gamma_I, \gamma) = \cos^{-1}\left(\int_0^1 \psi(t)dt\right) \tag{3}$$

where ψ is the SRSF of the optimal warping function γ that aligns f_2 to f_1. This simplification occurs because the space of all ψs forms a Hilbert Sphere (see Tucker et al., 2013).

2.1.2. Karcher Mean

Let $f_1, \ldots, f_n$ represent a set of functions from the function space $\mathcal{F}$ and $q_1, \ldots, q_n$ be their respective SRSFs. The Karcher mean of $f_1, \ldots, f_n$ is given by

$$\mu_f = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} d_a(f, f_i)^2. \tag{4}$$

Equivalently, we can define the Karcher mean of the SRSFs, $q_1, \ldots, q_n$, as

$$\mu_q = \underset{q \in \mathbb{L}^2}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\inf_{\gamma_i \in \Gamma} ||q - (q_i \circ \gamma_i)||^2\right). \tag{5}$$

Note that μ_q is the SRSF of μ_f. The algorithm used in the `fdasrvf R` package finds μ_q and then transforms it to μ_f.

The transformation from SRSF space back to the original function space, assuming a generic function f and its SRSF q, is

$$f(t) = f(t_0) + \int_{t_0}^{t} q(s)|q(s)|ds \tag{6}$$

where $f(t_0)$ is the function value at the initial time point, t_0. When obtaining μ_f from μ_q, the initial value is computed as $n^{-1}\sum_{i=1}^{n} f_i(t_0)$.

2.2. Conformal Prediction and ICAD

Conformal prediction was first introduced by Vovk et al. (2005) as a distribution-free approach to obtaining valid uncertainty quantification (UQ). It has recently been gaining popularity in the machine learning literature as a computationally efficient means to obtain high quality UQ with many data types and different classes of models (Angelopoulos and Bates, 2021; Zhou et al., 2024). Let $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ be an *i.i.d* random sample and X_{n+1} be a test point with Y_{n+1} the unobserved label associated with the test point X_{n+1}. For simplicity, we use the notation $Z_i = (X_i, Y_i)$. CP methods produce a *prediction set*, $C(X_{n+1})$, for the test point X_{n+1}, such that $P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$ for any level of significance, $\alpha \in (0, 1)$. This property is known as the marginal coverage guarantee, and we note that it also holds under the weaker assumption of exchangeability of $Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})$.

To produce a prediction set, a *non-conformity measure* (NCM) must first be defined. The NCM is a function that measures how well Z_i conforms with the other observations. In the original formulation of CP, NCM values are computed for all Z_i, $i \in \{1, \ldots, n, n+1\}$, relative to the set $\mathcal{Z}_{(-i)} = \{Z_j, : j = 1, \ldots, i-1, i+1, \ldots, n, n+1\}$. For NCMs that incorporate information from the entire set $\mathcal{Z}_{(-i)}$, this process can become computationally expensive.

Inductive CP (ICP) is an alternative approach that has a lower computational burden at the cost of less efficient use of available data. ICP randomly splits all available observations for training into two sets: a *training data set* and a *calibration data set*. The purpose of the training data set is to fit a base model which relates the object X to the label Y. Meanwhile, the purpose of the calibration data set is to evaluate the NCM. For clarity, we use the term *full training data set* to refer to all observations available for training and denote this set as $D^{full} = \{Z_{(1)}, \ldots, Z_{(n)}\}$. Similarly, we denote the training and calibration sets, respectively, as $D^{tr} = \{Z_1, \ldots, Z_{n_1}\}$, and $D^{cal} = \{Z_{n_1+1}, \ldots, Z_{n_1+n_2}\}$ (note the parentheses in subscripts of D^{full} are intended to emphasize that the observations in D^{full} are indexed differently than those in D^{tr} and D^{cal}, due to random splitting). Let $\mathcal{I}^{tr} = \{1, \ldots, n_1\}$ and $\mathcal{I}^{cal} = \{n_1 + 1, \ldots, n_1 + n_2\}$ represent the indices of the observations in D^{tr} and D^{cal}, respectively. Throughout this work, we take $n_1 = \lceil \frac{2}{3}n \rceil$ and $n_2 = n - n_1$ where $\lceil \cdot \rceil$ is the ceiling function. To construct a prediction set for a test observation, Z_{n+1}, the NCM is computed and compared to the NCM values obtained by evaluating the NCM on the calibration data set.

In this work, we use the p-value approach to constructing prediction sets. Let $s_{n_1+i} = s(Z_{n_1+i}; D^{tr})$ represent the NCM value of the i^{th} calibration observation and $s^y_{n+1} = s((X_{n+1}, y); D^{tr})$ be the NCM value of the test point where y is the assumed value of Y_{n+1}, the label associated with X_{n+1}. The p-value corresponding to X_{n+1} is then computed as

$$p^y_{n+1} = \frac{|\{i \in \mathcal{I}^{cal} : s_i \geq s^y_{n+1}\}| + 1}{n_2 + 1} \tag{7}$$

In traditional CP, when $p^y_{n+1} \geq \alpha$, the assumed value, y, of the label Y_{n+1}, is one element that defines the prediction set, $\mathcal{C}(X_{n+1})$. In the case of ICAD, there is no reliance upon the label, and to simplify the notation we drop the superscript y and denote the p-value as p_{n+1}. If exchangeability assumptions hold and $p_{n+1} \geq \alpha$, then Z_{n+1} is labeled as an

inlier with $(1-\alpha)\cdot 100\%$ confidence. Otherwise, Z_{n+1} is labeled an outlier. Note that the marginal coverage guarantee is only applicable to inliers and not to outliers. Also, we must use $\alpha \geq \frac{1}{n_2+1}$ or all test points will automatically be labeled as inliers.

3. Elastic Functional Distance Metrics ICAD

Our proposed method is ICAD where the NCM is based on elastic functional distances and the Karcher mean. We call this method *elastic functional distance metrics* ICAD, or EFDM. Let $D^{full} = \{f_{(1)}, \ldots, f_{(n)}\}$ be the full training data set of functional data observations. Further assume $f_{(1)}, \ldots, f_{(n)} \sim \mathcal{P}_{\mathcal{F}}$ are exchangeable, where $\mathcal{P}_{\mathcal{F}}$ is a probability distribution over the function space $\mathcal{F}$. Using random assignment we create $D^{tr} = \{f_1, \ldots, f_{n_1}\}$ and $D^{cal} = \{f_{n_1+1}, \ldots, f_{n_1+n_2}\}$. As described in Section 1, we also assume that all functional observations are measured at the same points in the domain. We denote these domain points as $\mathcal{T} = \{t_0, t_1, \ldots, t_M\}$.

To label a new function f_{n+1} as an inlier or outlier using a level of significance α, EFDM proceeds as follows:

1. Compute the Karcher mean of the training data set D^{tr}, denoted as μ_{tr}.

2. Compute the sets $d_a^{tr} = \{d_a(\mu_{tr}, f_i) \colon i \in \mathcal{I}^{tr}\}$ and $d_p^{tr} = \{d_p(\mu_{tr}, f_i) \colon i \in \mathcal{I}^{tr}\}$. Let $min_a = \min d_a^{tr}$, $max_a = \max d_a^{tr}$, $min_p = \min d_p^{tr}$, and $max_p = \max d_p^{tr}$.

3. Compute $d_{a\,i}^{cal} = d_a(\mu_{tr}, f_{n_1+i})$ and $d_{p\,i}^{cal} = d_p(\mu_{tr}, f_{n_1+i})$, the amplitude and phase distances from the Karcher mean to each function in the calibration data.

4. Compute the NCM for each calibration observation as

$$s_i = \frac{1}{2}\left[\left(\frac{d_{a\,i}^{cal} - min_a}{max_a - min_a}\right) + \left(\frac{d_{p\,i}^{cal} - min_p}{max_p - min_p}\right)\right]$$

5. Compute $d_a^{ts} = d_a(\mu_{tr}, f_{n+1})$ and $d_p^{ts} = d_p(\mu_{tr}, f_{n+1})$, the amplitude and phase distances from the Karcher mean to the test function.

6. Compute the NCM for the test function as

$$s_{n+1} = \frac{1}{2}\left[\left(\frac{d_a^{ts} - min_a}{max_a - min_a}\right) + \left(\frac{d_p^{ts} - min_p}{max_p - min_p}\right)\right]$$

7. Compute the p-value, p_{n+1}, as in equation (7). If $p_{n+1} < \alpha$, f_{n+1} is labeled as an outlier; else it is labeled as an inlier.

As seen in steps 4 and 6, the NCM used for EFDM is the average of the standardized amplitude and phase distances from an observation to the Karcher mean of the training data. This standardization is important so that both the amplitude and phase components are unitless and can contribute equally to the NCM. We now describe several potential adjustments to this NCM.

The basic NCM given above is primarily intended to detect shape outliers. This approach will also detect magnitude outliers if they are outlying *on the x-axis* (phase distance will properly account for these). However, magnitude outliers that are outlying predominately in the y direction may not be accounted for by amplitude distance, due to the differentiation when transforming a function to SRSF space. Thus, magnitude outliers that are essentially vertical shifts of observations in the data will not be detected. To enable the detection of such magnitude outliers, *translation distance* can be incorporated into the NCM. To do this we compute $|\mu_{tr}(t_k) - f_i(t_k)|$ where t_k is a point in the domain where the vertical shift is prominent. Once the translation distance is standardized it can be included alongside the standardized phase and amplitude distances as well. The final NCM is then the average of the three standardized distances.

Another possible adjustment is to allow for different weightings of the amplitude and phase (and potentially translation) distance components so that the NCM is a weighted average rather than a simple arithmetic mean. While a functional data set that displays more amplitude than phase variability, or vice versa, is not a problem because we standardize the distances, there may still be cases when it is desirable to weight one component higher than the other. Taking this to an extreme, it is also possible to use only one of the components if, for instance, we know *a prioi* that there is only one type of variability in a data set. We caution against this, however. Suppose a functional data set that contains only amplitude variability is used in the EFDM algorithm with only amplitude distance in the NCM. If a new observation looks similar to the training data in amplitude but varies in phase, it will not be marked as an outlier.

Finally, smoothed conformal prediction (Vovk et al., 2005) is often used to guarantee exact asymptotic validity. This is carried out by adjusting the p-value computation to

$$\tilde{p}_{n+1} = \frac{|\{i \in \mathcal{I}^{cal} : s_i > s_{n+1}\}| + \tau|\{i \in \mathcal{I}^{cal} \cup \{n+1\} : s_i = s_{n+1}\}|}{n_2 + 1} \tag{8}$$

where τ is a random draw from the $U(0,1)$ distribution. In all of our computations, we use these smoothed conformal p-values. Other p-value adjustments are possible and may be desirable, depending on the use case (see, e.g., Bates et al., 2023).

3.1. Other Functional ICAD Methods

In this section, we frame the methods of Lei et al. (2015) and Diquigiovanni et al. (2021) as functional ICAD methods for comparison to EFDM.

3.1.1. GMD

The GMD ICAD method begins by projecting the functional data to a lower-dimensional space. In our case, we use functional principal component analysis (FPCA) (Ramsay and Silverman, 2005) to achieve the projection, but other projection methods could be used. Let $\xi_{ij} = \langle f_i, \theta_j \rangle$ be the j^{th} component of the projected functional observation, f_i for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Here, θ_j represents the j^{th} functional principal component and $\langle \cdot, \cdot \rangle$ is the functional inner product (Ramsay and Silverman, 2005). The projection of f_i can then be written as $\xi_i = (\xi_{i1}, \ldots, \xi_{ip})$. Note that the FPCs are estimated using only the training data, and all training, calibration, and test data use the same FPCs for projection to ξ_i.

The projected data are modeled with a Gaussian mixture model (GMM) (Hastie et al., 2009) with K components. From the projected version of the training data, $D_\xi^{tr} = \{\xi_1, \ldots, \xi_{n_1}\}$, the mean, covariance and mixture proportions are learned. We denote the fitted GMM as

$$G(\xi) = \sum_{k=1}^{K} \hat{\pi}_k \phi(\xi; \hat{\mu}_k, \hat{\Sigma}_k)$$

where $\phi(\cdot; \mu, \Sigma)$ represents the multivariate normal density with mean vector μ and covariance matrix Σ, and the $\hat{\pi}_k$ are the estimated mixing proportions. The NCM is then computed as $s_i(\xi_i) = -1 \cdot G(\xi_i)$, hence the name Gaussian mixture density (GMD) for this approach[3]. We also implement a refinement that was introduced by Lei et al. (2015) to provide a narrower prediction band. With the GMD maximum (GMDM) refinement, the NCM is instead computed as

$$s_i(\xi_i) = -1 \cdot \left(\max_{\{1,\ldots,K\}} \hat{\pi}_k \phi(\xi_i; \hat{\mu}_k, \hat{\Sigma}_k) \right).$$

We wrote our own implementations of GMD and GMDM in `R` that can be found on Github. Our code uses the `mclust` package (version 6.1.1) (Scrucca et al., 2023) and the `mvtnorm` package (version 1.3.1) (Genz et al., 2021).

3.1.2. SNCM

In the work of Diquigiovanni et al. (2021), the proposed NCM is computed as

$$s_i(f_i) = \sup_{t \in \mathcal{T}} \left| \frac{f_i(t) - m(t)}{r(t)} \right|$$

where $m(t)$ is a mean function and $r(t)$ is a modulation function, and both of these are estimated from only the training data. Because this NCM uses the supremum, we refer to this method as supremum non-conformity measure (SNCM) ICAD. For the mean function, only the cross-sectional mean function is used by Diquigiovanni et al. (2021). For the modulation function, the authors compare three options: $r(t) = 1$; the cross-sectional standard deviation function; and what is referred to as the optimal modulation function. The optimal modulation function is defined as

$$r(t) = \frac{\max_{i \in H_1} |f_i(t) - m(t)|}{\int_{\mathcal{T}} \max_{i \in H_1} |f_i(t) - m(t)| dt}$$

where

$$H_1 = \begin{cases} \mathcal{I}^{tr} & \text{if } \lceil (n_1+1)(1-\alpha) \rceil > n_1 \\ \{i \in \mathcal{I}^{tr} : \sum_{t \in \mathcal{T}} |f_i(t) - m(t)| \leq \nu\} & \text{if } \lceil (n_1+1)(1-\alpha) \rceil \leq n_1 \end{cases}$$

where ν is the $\lceil (n_1+1)(1-\alpha) \rceil^{th}$ smallest value of the set $\{\sup_{t \in \mathcal{T}} |f_i(t) - m(t)| : i \in D^{tr}\}$.

We wrote our own implementation of SNCM in `R` that can be found on Github. Our code made use of the `fdasrvf` package (version 2.3.4) and the `caTools` package (version 1.18.3) (Tuszynski and Dietze, 2024).

3. Note that Lei et al. (2015) use a *conformity* measure rather than a non-conformity measure, so they simply compute the density rather than the negative density. The two approaches are equivalent.

4. Simulation Experiments

In this section, we empirically evaluate the performance of the methods described in Section 3. Magnitude outliers, as mentioned in Section 1.1.1, are the easier case, so these experiments focus instead on detecting shape outliers.

4.1. Experiment Data

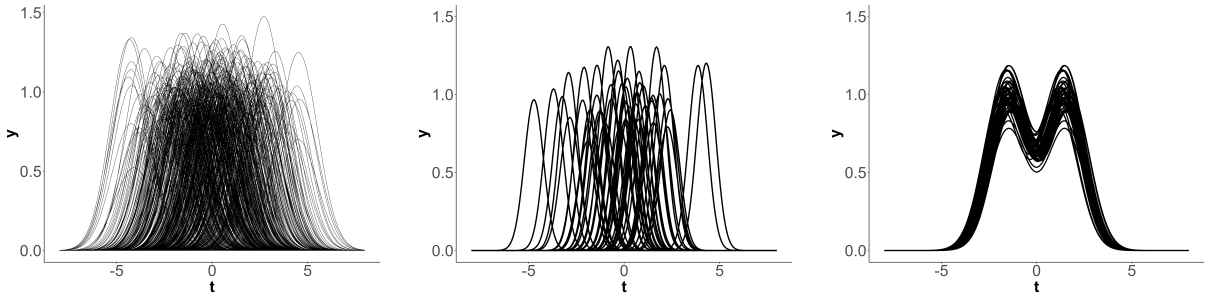
These experiments use functional data generated from three different templates. These are the *standard*, *narrow*, and *double peaked* templates. Respectively, they are given as:

$$F_{st}(t) = a \exp\left[-\frac{1}{2}(t-p)^2\right]$$

$$F_{nr}(t) = a \exp\left[-2(t-p)^2\right]$$

$$F_{dp}(t) = a \exp\left[-\frac{1}{2}(t-1.5-p)^2\right] + a \exp\left[-\frac{1}{2}(t+1.5-p)^2\right].$$

To generate functional data from these templates, we randomly generate values for the a and p parameters. Figure 2 shows randomly generated data from each of these templates. Panel (a) contains 500 standard functions, panel (b) contains 50 narrow functions, and panel (c) contains 50 double peaked functions. The standard and narrow data sets were generated using $a \sim N(\mu = 1, \sigma = 0.15)$ and $p \sim N(0, 1.75)$. The double peaked functions were generated with $a \sim N(1, 0.10)$ and $p \sim N(0, 0.15)$. The domain points for all generated functions are 250 equally spaced points over the interval $[-8, 8]$. Treating the standard data as inliers, the key thing to note about both the narrow and double peaked data is that these are *shape-only* outliers. In terms of magnitude, they fit well within the distribution of the standard data.



(a) (b) (c)

Figure 2: Randomly generated data sets. (a) 500 standard functions, (b) 50 narrow functions, and (c) 50 double peaked functions. These are also the test sets for the first experiment.

4.2. Experiment 1

The goal of the first experiment is two-fold: first, we investigate the coverage of the ICAD methods over inlier and outlier functions separately. Second, we consider the effect of

sample size on the coverage. Using full training data set sizes of $n = 50$, 100, and 250, we randomly generate 500 data sets using the standard template. That is, there are 1500 total simulated full training data sets for this experiment. Each of the 1500 data sets are run through the ICAD methods, and inlier/outlier labels are assigned to each function in three different test sets. The first test set consists of 500 standard functions generated with the same parameters as the full training sets (these are the inlier functions). The second test set contains 50 narrow functions, and the third contains 50 double peaked functions. Both the second and third test sets are outliers relative to the training data. These test sets are shown in Figure 2.

Method Name	Details
EFDM	NCM uses amplitude and phase distances only
SNCM1	Uses the cross-sectional mean function and the optimal modulation function
SNCM2	Uses the Karcher mean and the optimal modulation function
GMD1	Uses $p = 2$, $K = 1$
GMD2	Uses $p = 3$, $K = 2$
GMD3	Uses $p = 5$, $K = 2$
GMDM	Uses $p = 5$, $K = 2$

Table 1: Description of methods used for experiment 1.

Table 1 details each method used for this experiment. For each simulated data set, inlier/outlier labels were obtained for every functional observation in the three test sets using a significance level of $\alpha = 0.10$. The coverage for a single simulated data set is simply the proportion of test functions marked as inliers (this can be viewed as the coverage averaged over the test functions). These proportions are then averaged over all 500 simulated data sets that have the same sample size to provide the final mean coverage estimate for a method. These values, along with the standard deviation over the simulated data sets, are given in Tables 2 and 3. Table 2 contains results from the standard test set where the mean coverage values should be close to $1 - \alpha = 0.90$. Table 3 contains results from the narrow and double peaked test sets where mean coverage values close to 0 are desirable. Note that the values for GMD3 and GMDM are missing because, with some of the full training sets of size $n = 50$ and $n = 100$, the GMMs could not be fit. In all tables, bolded values represent the best value(s) in each column.

	Standard Test Data		
Method	$n = 50$	$n = 100$	$n = 250$
EFDM	0.893(0.07)	0.895(0.05)	**0.901 (0.03)**
SNCM1	**0.900 (0.06)**	**0.899 (0.05)**	**0.899 (0.03)**
SNCM2	0.901(0.06)	**0.899 (0.05)**	0.903(0.03)
GMD1	0.901(0.07)	**0.899 (0.05)**	0.905(0.04)
GMD2	**0.900 (0.06)**	**0.899 (0.05)**	**0.901 (0.03)**
GMD3	–	–	**0.901 (0.03)**
GMDM	–	–	0.884(0.03)

Table 2: Mean(SD) coverage results from experiment 1 for the standard test data.

	Narrow Test Data			Double Peaked Test Data		
Method	$n = 50$	$n = 100$	$n = 250$	$n = 50$	$n = 100$	$n = 250$
EFDM	**0.052(0.13)**	**0.001(0.02)**	**0(0)**	**0.001(0.01)**	**0(0)**	**0(0)**
SNCM1	0.927(0.08)	0.911(0.06)	0.888(0.04)	0.972(0.08)	0.982(0.04)	0.991(0.03)
SNCM2	0.927(0.09)	0.908(0.07)	0.882(0.08)	0.966(0.08)	0.978(0.05)	0.992(0.03)
GMD1	0.992(0.02)	0.997(0.01)	0.999(0.002)	0.999(0.002)	1(0)	1(0)
GMD2	0.889(0.13)	0.900(0.09)	0.912(0.04)	0.389(0.34)	0.241(0.26)	0.131(0.17)
GMD3	–	–	0.542(0.19)	–	–	0.170(0.21)
GMDM	–	–	0.487(0.19)	–	–	0.104(0.15)

Table 3: Mean(SD) coverage results from experiment 1 for the narrow and double peaked test data and standard functions as training/calibration data.

The results of Table 2 utilize full training and testing data sets that are comprised of standard functions. Thus, the mean coverage values, as seen in Table 2, are all very close to 0.90 for all methods. The only effect of sample size seems to be to reduce the variability in the coverage estimates for each data set, as would also be expected. From this table, no methods stand out as outperforming the others. In the online supplementary material, we include plots that show that different methods achieve the correct average coverage in distinct ways.

The performance of the methods is much more variable in Table 3. Here, EFDM has the best performance in every case. With enough training data ($n = 250$ for the narrow data and $n = 100$ for the double peaked), it never labels outliers as inliers in any of the 500 runs. Both SNCM methods perform poorly, but the coverage does decrease somewhat as sample size increases for the narrow test data. This pattern, however, does not hold for the double peaked data. GMD1 has the worst performance of any method in every case. GMD2 performs poorly on the narrow test data but much better on the double peaked data. GMD3 and GMDM perform better on the narrow test data than the other GMD and SNCM methods, and they even outperform GMD2 on the double peaked data.

4.3. Experiment 2

In experiment 2, our goal is to simulate the first outlier detection scenario where the training and calibration data contain no outliers and external data, which may contain outliers, are labeled using the ICAD methods. In this experiment, we reuse the same 1500 full training data sets from experiment 1 (so there are 500 data sets for each sample size, $n = 50$, $n = 100$, and $n = 250$). Each full training set has two corresponding test sets containing 100 functional observations each. One test set contains 95% inliers (i.e., they are simulated using the standard template function and the same normal distribution parameters as the full training data) and 5% double peaked outliers, simulated with the same parameters as the double peaked test data in experiment 1. The other test set contains 90% inliers and 10% double peaked outliers.

Each full training data set is randomly split into training and calibration data, and each of the methods from Table 1 are used to assign inlier/outlier status to all functions in both test data sets. For each test data set, we examine the effects of using two different levels of significance, $\alpha = 0.05$ and $\alpha = 0.10$. Method performance is assessed using the Matthew's

correlation coefficient (MCC) (Matthews, 1975), which has been shown to perform well in cases of extreme class imbalance (Chicco et al., 2021). MCC values close to 1 indicate better performance (or higher positive correlation between predictions and ground truth labels) while values close to 0 indicate poor performance (little correlation between predictions and ground turth labels). Since it is a correlation, MCC can also be negative (indicating negative correlation between predictions and ground truth labels). For this experiment, the double peaked outliers are treated as the positive class. MCC is computed for each test set, and the mean and standard deviation over the 500 full training sets (per sample size) are given in Tables 4 and 5. As in experiment 1, values are missing for GMD3 and GMDM when $n = 50$ and $n = 100$. Additional metrics from this experiment are given in the online supplementary material.

	5% Outliers in Test Data			10% Outliers in Test Data		
Method	$n = 50$	$n = 100$	$n = 250$	$n = 50$	$n = 100$	$n = 250$
EFDM	**0.668(0.19)**	**0.734(0.14)**	**0.728(0.12)**	**0.735(0.15)**	**0.818(0.11)**	**0.817(0.10)**
SNCM1	−0.033(0.05)	−0.037(0.04)	−0.044(0.03)	−0.049(0.06)	−0.058(0.04)	−0.063(0.04)
SNCM2	−0.027(0.06)	−0.034(0.05)	−0.044(0.03)	−0.042(0.07)	−0.053(0.05)	−0.062(0.04)
GMD1	−0.043(0.03)	−0.048(0.03)	−0.049(0.02)	−0.062(0.04)	−0.067(0.03)	−0.068(0.03)
GMD2	0.276(0.27)	0.353(0.24)	0.426(0.22)	0.342(0.28)	0.428(0.25)	0.508(0.21)
GMD3	–	–	0.403(0.27)	–	–	0.464(0.28)
GMDM	–	–	0.435(0.24)	–	–	0.522(0.24)

Table 4: Mean(SD) MCC values for experiment 2 using $\alpha = 0.05$.

	5% Outliers in Test Data			10% Outliers in Test Data		
Method	$n = 50$	$n = 100$	$n = 250$	$n = 50$	$n = 100$	$n = 250$
EFDM	**0.601(0.16)**	**0.589(0.14)**	**0.578(0.10)**	**0.710(0.14)**	**0.702(0.11)**	**0.697(0.09)**
SNCM1	−0.046(0.07)	−0.051(0.06)	−0.059(0.05)	−0.066(0.08)	−0.081(0.06)	−0.085(0.05)
SNCM2	−0.042(0.07)	−0.048(0.07)	−0.059(0.05)	−0.059(0.08)	−0.070(0.06)	−0.085(0.05)
GMD1	−0.070(0.03)	−0.073(0.02)	−0.072(0.02)	−0.098(0.04)	−0.103(0.04)	−0.102(0.03)
GMD2	0.359(0.23)	0.444(0.17)	0.495(0.12)	0.440(0.26)	0.546(0.17)	0.609(0.13)
GMD3	–	–	0.479(0.16)	–	–	0.588(0.15)
GMDM	–	–	0.473(0.11)	–	–	0.596(0.11)

Table 5: Mean(SD) MCC values for experiment 2 using $\alpha = 0.10$.

Both Tables 4 and 5 exhibit similar patterns to those seen in Table 3. As in the previous experiment, EFDM outperforms all other methods in each case. The performance gap between EFDM and the other methods is larger when $\alpha = 0.05$ than when $\alpha = 0.10$. In fact, when $\alpha = 0.10$ the mean MCC for EFDM is within one standard deviation of the mean MCC for GMD2 (and GMD3 and GMDM when applicable). In both tables, SNCM1, SNCM2, and GMD1 consistently perform poorly, registering negative MCC values in every case.

4.4. Experiment Conclusions

There are several conclusions to be drawn from these results. As discussed, shape outliers are harder to detect than magnitude outliers, and we intended the narrow and double peaked outliers to be a particularly difficult case of shape outliers relative to our standard simulated data. When plotted, both narrow and double peaked outliers fit well within the standard data. The narrow data are also only very slightly different in shape than the standard data. We note that Lei et al. (2015) is primarily focused on visualization of functional data, and the method of Diquigiovanni et al. (2021) is best suited for identifying magnitude outliers. Thus, it should be emphasized that we are considering a case not explicitly mentioned by either of the other two works.

Regarding SNCM, there does not seem to be any benefit of using the Karcher mean rather than a cross-sectional mean function as SNCM1 and SNCM2 performed similarly in all experiments. SNCM does not appear to be well-suited for detecting the kind of shape outliers that we used in these experiments.

There did not seem to be much of a difference in performance between GMD and GMDM methods when they used the same parameters. In both experiments, some of the GMD/GMDM methods perform reasonably well detecting these difficult shape outliers. We believe that the poor performance of GMD1 is due to its rigidity. That is, the projected functional data are being represented by a single bivariate Gaussian distribution. It is likely not flexible enough to provide a good representation of the functional data we used. As flexibility increases with larger values of p and K, these methods do much better, even coming fairly close to EFDM performance on experiment 2. With even greater flexibility, it is possible that GMD/GMDM methods could perform even better. One downside of this method is that more data are needed to enable more flexible implementations. On the other hand, GMD/GMDM methods are trained and produce labels with relatively low computational expense, so they may be a good choice with large functional data sets.

We have already noted that EFDM performs well in each experiment. We also emphasize here that EFDM can perform well on relatively small sample sizes. When $n = 50$, $n_1 = 34$ and $n_2 = 16$. Providing the correct coverage and being useful as an outlier detector with such a small amount of data is an important feature of EFDM.

Finally, experiment 2 clearly demonstrates that results depend heavily on the selection of the significance level, α. To select an appropriate value, the user needs to understand and compare the costs of false positives and false negatives. This decision is further influenced by the fact that we must choose $\alpha \geq \frac{1}{n_2+1}$. Because selecting α forces a determination about outlier status, it may be wise to analyze p-values rather than settling on a single value of α for a given application.

5. Analysis of Exemplar Data

In this section, we analyze a real data set using EFDM. Our purpose is to demonstrate a use case of EFDM and not to provide thorough analysis of the data.

5.1. Temperature Data

For our exemplar, we use data downloaded from the National Centers for Environmental Information at the National Oceanic and Atmospheric Administration (NOAA). We downloaded daily high and low temperatures and calculated the daily mean temperature from the years 1981-2024 at five different airports in the United States: Atlanta (Hartsfield-Jackson), Dallas-Love Field, Dallas-Fort Worth International, Miami International, and San Francisco International. In this case, each year is taken as a single functional observation. Thus, we train on the data from each site and then label the years at all other sites as inliers or outliers. With this analysis, we hope to see most years labeled as inliers for similar sites (e.g., the two sites in Dallas) and most years labeled as outliers when the sites are dissimilar (e.g., Miami and San Francisco). To account for potential magnitude outliers, we included translation distance in two runs of EFDM. The first run used the initial point for computing translation distance while the second run used the midpoint, where the Atlanta data seems to differ most from the Dallas data sets.

EFDA methods work best on smooth data, and since our goal is to demonstrate EFDM capabilities on real data rather than provide thorough analysis, the temperature data went through overly simplistic preprocessing before training. We stress that the analysis here should not be used to make general conclusions about the temperature data themselves but only to shed light on EFDM as a method. To produce the data used for analysis, we first computed a daily mean temperature, $d_{mean} = 0.5(d_{max} + d_{min})$ for every day over the 44 years at each site, where d_{max} and d_{min} are the daily high and daily low temperatures, respectively. We then averaged all d_{mean} temperatures in a given month to produce m_{mean}. At this point, each functional observation (year) was observed over 12 time points (months). We then smoothed the data using a Fourier basis, as described by Ramsay and Silverman (2005), and interpolated the smoothed functions so that we again had daily mean temperatures.

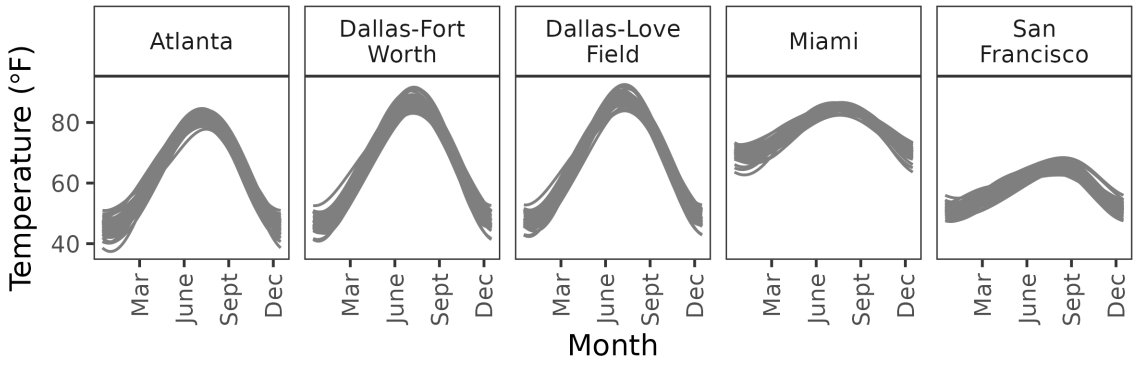


Figure 3: The monthly averaged daily mean temperature data for five sites in the United States.

Because each site has $n = 44$ years of temperature data, we have $n_1 = 30$ training functions and $n_2 = 14$ calibration functions. Since we need $\alpha \geq \frac{1}{n_2+1}$, we use $\alpha = 0.10$. Table 6 shows the percentage of years at the site in the column that are labeled as outliers when training on the site in the row. All years at all other sites are labeled as outliers when

training on either Miami or San Francisco. When training and testing on Atlanta and the two Dallas sites, the proportion of outliers is very different depending on which point is used for computing translation distance. This highlights the impact the point selection can have when detecting magnitude outliers. Similar tables are presented for GMD and SNCM in the supplementary material.

	EFDM with Initial Point					EFDM with Midpoint				
	Testing Site					Testing Site				
Training Site	ATL	DFW	DAL	MIA	SFO	ATL	DFW	DAL	MIA	SFO
ATL	–	0.114	0.091	1	1	–	0.841	0.682	1	1
DFW	0.5	–	0.205	1	1	0.864	–	0.25	1	1
DAL	0.318	0.182	–	1	1	0.5	0.091	–	1	1
MIA	1	1	1	–	1	1	1	1	–	1
SFO	1	1	1	1	–	1	1	1	1	–

Table 6: Percentage of outliers when training on the site in the row and labeling the site in the column for all five sites: Atlanta (ATL), Dallas-Forth Worth (DFW), Dallas-Love Field (DAL), Miami (MIA), and San Francisco (SFO). The first five columns show results of using EFDM with translation distance computed at the initial point; the last five columns are obtained using the midpoint instead.

6. Conclusion

In this paper, we have introduced the elastic functional distance metrics ICAD method, evaluated it alongside two competing ICAD methods for functional data, and demonstrated its effectiveness on a real data set. We argue that our results show EFDM to be effective at detecting both shape and magnitude outliers.

As argued by Bates et al. (2023), users of conformal prediction methods, including ICAD, need to be aware of potential multiple testing pitfalls. In particular, the coverage guarantee of CP methods holds for a single test point, $P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$. However, the probability that a large number of test points will all belong to their respective prediction sets can become much smaller than $1 - \alpha$ as the number of test points increases. We did not implement any corrections for this, and this could be potentially increasing the number of false positives for some of our results. In future work, we will implement corrections and compare the outcomes with our current results.

Code, Data, and Supplementary Materials

The R code for implementing the methods and analyses in this work is available at https://github.com/sandialabs/conformal-functional-outliers. This repository also contains data sets used in both the exemplar analysis and the simulation experiments. Additional results from our analyses can also be found in the supplementary material document within the respository.

Acknowledgments

The authors express gratitude to Dr. Jing Lei who graciously shared `R` code implementing his functional conformal prediction methods.

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

Charu C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2nd edition, 2016. ISBN 3319475770.

Niccolò Ajroldi, Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for two-dimensional functional time series. *Computational Statistics & Data Analysis*, 187:107821, 2023.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

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

Arturo Azcorra, Luis F Chiroque, Rubén Cuevas, Antonio Fernández Anta, Henry Laniado, Rosa Elvira Lillo, Juan Romo, and Carlo Sguera. Unsupervised scalable statistical method for identifying influential users in online social networks. *Scientific reports*, 8(1):6955, 2018.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

Feiyang Cai and Xenofon Koutsoukos. Real-time out-of-distribution detection in cyber-physical systems with learning-enabled components. *IET Cyber-Physical Systems: Theory & Applications*, 7(4):212–234, 2022.

Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *Ieee Access*, 9:78368–78381, 2021.

Wenlin Dai, Tomáš Mrkvička, Ying Sun, and Marc G Genton. Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, 149:106960, 2020.

Anna De Magistris, Andrea Diana, and Elvira Romano. Conformal prediction for functional ordinary kriging. *arXiv preprint arXiv:2409.20084*, 2024.

Andrea Diana, Elvira Romano, and Antonio Irpino. Distribution free prediction for geographically weighted functional regression models. *Spatial Statistics*, 57:100765, 2023.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. *arXiv preprint arXiv:2102.06746*, 2021.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Bjoern Bornkamp, Martin Maechler, Torsten Hothorn, and Maintainer Torsten Hothorn. Package 'mvtnorm'. *Journal of Computational and Graphical Statistics*, 11(950-971):155, 2021.

Trevor Harris, J Derek Tucker, Bo Li, and Lyndsay Shand. Elastic depths for detecting shape anomalies in functional data. *Technometrics*, 63(4):466–476, 2021.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Huang Huang and Ying Sun. A decomposition of total variation depth for understanding functional outliers. *Technometrics*, 61(4):445–458, 2019.

Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

Rob J Hyndman and Han Lin Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010.

Vladislav Ishimtsev, Alexander Bernstein, Evgeny Burnaev, and Ivan Nazarov. Conformal k-nn anomaly detector for univariate data streams. In *Conformal and Probabilistic Prediction and Applications*, pages 213–227. PMLR, 2017.

Rikard Laxhammar. *Conformal anomaly detection: Detecting abnormal trajectories in surveillance applications*. PhD thesis, University of Skövde, 2014.

Rikard Laxhammar and Göran Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74:67–94, 2015.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

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.

Sara López-Pintado and Juan Romo. A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55(4):1679–1695, 2011.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis, 2nd*. Springer, New York, 2005. doi: 10.1002/0471667196.ess3138.

Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Pallavi Sawant, Nedret Billor, and Hyejin Shin. Functional outlier detection with robust functional principal component analysis. *Computational Statistics*, 27:83–102, 2012.

Luca Scrucca, Chris Fraley, T Brendan Murphy, and Adrian E Raftery. *Model-based clustering, classification, and density estimation using mclust in R*. Chapman and Hall/CRC, 2023.

Anuj Srivastava and Eric P Klassen. *Functional and shape data analysis*, volume 1. Springer, 2016.

Ying Sun and Marc G Genton. Functional boxplots. *Journal of computational and graphical statistics*, 20(2):316–334, 2011.

J. D. Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics and Data Analysis*, 61:50–66, 2013.

J Derek Tucker. Package 'fdasrvf', 2017.

J Derek Tucker, John R Lewis, Caleb King, and Sebastian Kurtek. A geometric approach for computing tolerance bounds for elastic functional data. *Journal of applied statistics*, 47(3):481–505, 2020.

Jarek Tuszynski and Michael Dietze. Package 'catools', 2024.

Shahid Ullah and Caroline F Finch. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13:1–12, 2013.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Fangyi Wang, Sebastian Kurtek, and Yuan Zhang. Joint registration and conformal prediction for partially observed functional data. *arXiv preprint arXiv:2502.15000*, 2025.

Weiyi Xie, Sebastian Kurtek, Karthik Bharath, and Ying Sun. A geometric approach to visualization of variability in functional data. *Journal of the American Statistical Association*, 112(519):979–993, 2017.

Fengmin Yu, Liming Liu, Liying Jin, Nanxiang Yu, and Hua Shang. A method for detecting outliers in functional data. In *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, pages 7405–7410. IEEE, 2017.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. Conformal prediction: A data perspective. *arXiv preprint arXiv:2410.06494*, 2024.