Density kernel depth for outlier detection in functional data

Autor ditelliano: *Martos Venturini, Gabriel (Universidad Torcuato Di Tella. Departamento de Matemáticas y Estadística)*

Fecha de publicación: *2023*

Publicado originalmente en: *International Journal of Data Science and Analytics*

International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00420-w

REGULAR PAPER

Density kernel depth for outlier detection in functional data

Nicolás Hernández[1] · Alberto Muñoz[2] · Gabriel Martos[3]

Received: 31 January 2023 / Accepted: 15 June 2023

Abstract

In this paper, we propose a novel approach to address the problem of functional outlier detection. Our method leverages a low-dimensional and stable representation of functions using Reproducing Kernel Hilbert Spaces (RKHS). We define a depth measure based on density kernels that satisfy desirable properties. We also address the challenges associated with estimating the density kernel depth. Throughout a Monte Carlo simulation we assess the performance of our functional depth measure in the outlier detection task under different scenarios. To illustrate the effectiveness of our method, we showcase the proposed method in action studying outliers in mortality rate curves.

Keywords Functional data · Depth measures · Outlier detection · Mortality curves

1 Introduction

Advances in technology are providing data scientists with an unprecedented amount of high-dimensional data. Electrocardiogram signals, fMRI images or Mortality curves are relevant examples of what is nowadays called *Functional Data* [1]. In Functional Data Analysis (FDA), each observation is a function that constitutes an infinite dimensional object. Analysing functional outliers is critical in several contexts, including functional regression [2], robust functional principal component analysis [3], functional outlier visualisation [4], robust functional data clustering [5]; and also in several applied context where FD is involved [6].

Functional outliers are commonly classified into two categories: *magnitude outliers* and *shape outliers* [7, 8]. The contribution of this paper, is to propose a novel depth measure for functional data to handle both type of functional outliers simultaneously. To achieve this goal, we introduce a density kernel depth that relies on a finite dimensional and stable representation of functions. The kernel depth measure satisfies desirable properties, as we formally discuss in Sect. 3.1, and induces a centre—outward ordering on functions. We also propose suitable estimation methods that are based on a One-Class-Neighbour-Machine, a non parametric estimator of density level sets. Furthermore, we discuss a statistically sound bootstrap approach to infer functional outliers in data.

The remainder of the paper is organised as follows: Sect. 2 propose a suitable representation model for functional data. In Sect. 3 we introduce a density kernel depth measure, and also discuss its properties and suitable estimation methods. In Sect. 4 we provide an algorithm for functional outlier detection based on bootstrap procedure. In Sect. 5, we present a Monte Carlo simulation study and a real data application to mortality curves. Finally, in Sect. 6 we conclude the paper.

2 An RKHS framework for functional data

In what follows we consider a square integrable stochastic processes $\mathcal{X}(t) \in \mathcal{H}$ in a separable Hilbert space of functions $\mathcal{H} \subset L_2(T)$, where $T \subset \mathbb{R}$ is a compact and convex set. As usual in practice, we also assume that curves are sampled over a discrete grid of points $\mathbf{t} = \{t_1, \dots, t_p\}$, being $p \gg 0$, in a signal plus noise fashion as follows:

$$\mathbf{x} = x(\mathbf{t}) + \mathbf{e}, \tag{1}$$

where $\mathbf{x} = \{x(t_1) + e_1, \dots, x(t_p) + e_p\}$ is the vector with the observed data and $\mathbf{e} = (e_1, \dots, e_p)$ is an independent and zero-mean residual term.

✉ Nicolás Hernández
n.hernandez@ucl.ac.uk

[1] Department of Statistical Science, University College London, 1-19 Torrington Place, London W1T 4TJ, UK

[2] Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid, 126, Getafe, Madrid 28903, Spain

[3] Departamento de Matemática y Estadística, Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, Ciudad de Buenos Aires 1428BCW, Buenos Aires, Argentina

Published online: 04 August 2023

Most functional data analysis approaches for preprocessing raw data as in Eq. (1) suggest to proceed as follows: Choose an orthogonal basis of functions $B = \{\phi_i\}_{i\geq 1}$, where each $\phi_i \in \mathcal{H}$, and then represent each functional datum by means of a linear combination in the Span(B) [9, 10]. A usual choice is to consider $\mathcal{H}$ as a Reproducing Kernel Hilbert Space (RKHS) of functions [11]. In this case, the elements in the spanning set B are the eigenfunctions associated to the positive-definite and symmetric kernel $K : T \times T \to \mathbb{R}$ that span $\mathcal{H}$. In our setting, the functional representation problem can be framed as follows: We observe each curve on p sample points and the corresponding functional data estimator is obtained solving the following regularization problem:

$$\widehat{x}(t) := \arg\min_{h\in\mathcal{H}} \sum_{j=1}^{p} L(x(t_j), h(t_j))^2 + \gamma\Omega(h), \tag{2}$$

where L is a strictly convex functional with respect to the second argument, $\gamma > 0$ is a regularization parameter (chosen by cross–validation), and $\Omega(h)$ is a regularization term. By the Representer Theorem [[12], Th. 5.2,p. 91] [[13], Pr. 8,p. 51] the solution of the problem stated in Eq. (2) exists, is unique, and admits a representation of the form:

$$\widehat{x}(t) = \sum_{j=1}^{p} \alpha_j K(t, t_j). \tag{3}$$

In the particular case of a squared loss function $L(w, z) = (w - z)^2$ and considering $\Omega(h) = \int_T h^2(t)\, \mathrm{d}t$, the coefficients of the linear combination in Eq. (3) are obtained solving the following linear system:

$$(\gamma p\mathbf{I} + \mathbf{K})\boldsymbol{\alpha} = \mathbf{x}, \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$, $\mathbf{I}$ is the identity matrix of order p, and $\mathbf{K}$ the is the $p \times p$ Gram matrix with the kernel evaluations, $[\mathbf{K}]_{k,l} = K(t_k, t_l)$, for $k = 1, \ldots, p$ and $l = 1, \ldots, p$. A main drawback on the estimation entitled in Eq. (3) is the instability of $\boldsymbol{\alpha}$, which change substantially under small perturbations in data. To avoid such problem, we resort on Mercer theorem [11] and consider an alternative functional estimator based on the projection of $\widehat{x}(t)$ onto the functional space generated by the first $d \ll p$ eigenfunctions of K:

$$\widehat{x}(t) \approx \tilde{x}(t) = \sum_{j=1}^{d} \lambda_j \phi_j(t) = \boldsymbol{\lambda}^T \boldsymbol{\Phi}(t), \tag{5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ are the projection coefficients onto the functional space generated by the first d eigenfunctions of K (i.e. $\lambda_j \equiv l_j(\boldsymbol{\alpha}^T \mathbf{v}_j)/\sqrt{p}$, where $(l_j, \mathbf{v}_j)$ is the j^{th} eigen pair of Gram matrix $\mathbf{K}$), $\boldsymbol{\Phi}(t) = (\phi_1(t), \ldots, \phi_d(t))$ is a vector function with the first d eigenfunctions associated to K, and $d \ll p$ is a resolution parameter such that for a small ε_d it holds that $\sup_{t\in T} \mid \widehat{x}(t) - \tilde{x}(t) \mid \leq \varepsilon_d$, see [14] for further details. The proposed depth measures for functional data relies on the computation of $\boldsymbol{\lambda}$, as we discuss in next Section.

3 Depth measures for functional data

There are several notions of depth measures in Statistics, all of them involve the computation of a quantity that represent the centrality of a given point $\mathbf{z} \in \mathbb{R}^d$ with respect to a probability distribution f. In this way, depth measures induce an order in data, and are a natural tool to identify outliers.

Some remarkable examples of depth measures for functional data are in order. In Fraiman and Muniz [15], the authors propose the *Integrated Depth* that resort on a trimmed functional mean estimator to ranks the functions. The *Random Tuckey Depth* [16] and the *Random Projection Depth* [17], rely on random projections of the functional data and the computation of the deepest function using univariate statistics. Another well known example is the case of the *h-Mode Depth* [17], that considers the expected kernel distances for each curve using the L_2 norm, see Example 1.

Example 1 *h-Mode depth.*

$$\text{h-MD}(x(t)) = E(K_h(\|x(t) - \mathcal{X}(t)\|_{L_2}),$$

where K is a kernel function and h is the bandwidth parameter.

In FDA, the graphical analysis is always a complementary approach in terms of visualisation and interpretation of outliers. In this sense, the *Band Depth* and *Modified Band depth* introduced in López-Pintado and Romo [18] are suitable methods. An interesting review on topological functional depth measures can be found in [19] and [20]. We introduce next our depth measure that rely on density kernels.

3.1 Depth induced by density kernels

Let $Z \in \mathbb{R}^d$ be a random vector with density function f, the function $g : \mathbb{R}^d \to \mathbb{R}$ is f-monotone if satisfies the following condition:

$$f(\mathbf{z}) \geq f(\mathbf{y}) \Rightarrow g(\mathbf{z}, f) \geq g(\mathbf{y}, f). \tag{6}$$

As an example, consider the parametric model $Z \sim N(\mu, \sigma^2)$; then $g(x, (\mu, \sigma^2)) = -(x - \mu)^2/\sigma^2$ is f-

monotone. A density kernel $K_f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as the product of two f-monotone functions as follows:

$$K_f(\mathbf{z}, \mathbf{y}) = g(\mathbf{z}, f)g(\mathbf{y}, f). \tag{7}$$

Notice that K_f depends on the density function f that is unknown or intractable in practice. We address the estimation of K_f using asymptotically f-monotone functions as discussed in next subsection. The density kernel depth measure is then obtained by combining a density kernel with a deepest (central) curve. To define the later, consider a statistical model f (a bounded density function) for the projection coefficients (i.e. $\boldsymbol{\lambda} \sim f$), the deepest curve $\tilde{x}_{\text{mo}}(t) = \sum_{j=1}^{d} m_j \phi_j(t)$ is in correspondence to the parameter $\mathbf{m} \equiv (m_1, \ldots, m_d) = \text{argmax}_{\boldsymbol{\lambda} \in \mathbb{R}^d} f(\boldsymbol{\lambda})$ (i.e. $\mathbf{m}$ is the mode of f). Finally, the density kernel depth (DKD) is defined as follows:

Definition 1 **Density Kernel Depth.** Let f be the density of the projection coefficients $\boldsymbol{\lambda}$, and let $\tilde{x}_{\text{mo}}(t) = \sum_{j=1}^{d} m_j \phi_j(t)$ be the deepest curve where $\mathbf{m} \equiv (m_1, \ldots, m_d)$ represent the mode of f, the DKD of the curve $\tilde{x}(t) = \boldsymbol{\lambda}^T \boldsymbol{\Phi}(t)$ is defined as follows:

$$\text{DKD}(\tilde{x}(t), f, K_f) \equiv g(\boldsymbol{\lambda}, f)g(\mathbf{m}, f), \tag{8}$$

Proposition 1 *The DKD satisfies the following desirable properties [21] for depth measures:*

P1. **Maximality at center**: *The DKD take the largest value evaluated at the deepest curve, i.e.* $\sup_{x(t) \in \mathcal{H}} DKD(x(t), f, K_f) = DKD(\tilde{x}_{mo}(t), f, K_f)$.

P2. **Monotonicity relative to the deepest function**: *For two curves* $\tilde{x}_1(t)$ *and* $\tilde{x}_2(t)$ *such that* $f(\boldsymbol{\lambda}_1) \leq f(\boldsymbol{\lambda}_2)$ *(i.e.* $\tilde{x}_1(t)$ *is further apart from the central curve* $\tilde{x}_{mo}(t)$ *than* $\tilde{x}_2(t)$*), it holds that* $DKD(\tilde{x}_1(t), f, K_f) \leq DKD(\tilde{x}_2(t), f, K_f)$.

P3. **Vanishes at infinity**: $DKD(\tilde{x}(t), f, K_f) \to 0$ *as* $\|\boldsymbol{\lambda}\| \to \infty$.

P4. **Invariant under affine transformations**: *Let* $\mathcal{T}$ *be the class of affine transformations in* $\mathcal{H}$ *and let* $\tau \in \mathcal{T}$ *be an affine map, then* $DKD(\tilde{x}(t), f, K_f) = DKD(\tau \circ \tilde{x}(t), f, K_f)$.

In addition to properties P1–P4, the order induced by the DKD is also invariant under changes in the basis function $B = \{\phi_1(t), \ldots, \phi_d(t)\}$ that span the linear subspace in $\mathcal{H}$ where we project the curves. We give more details and a formal proof in the supplementary material. In the following subsection we address the details involved in the estimation of DKD from data.

3.2 Estimating the density kernel depth

Notice that $g(\cdot, f)$ depends on the density function f that is unknown or intractable in practice; this leads to the following related concept: Given a random sample $S_n = \{Z_1, \ldots, Z_n\} \overset{iid}{\sim} f$, then $g(\mathbf{z}, S_n)$ is asymptotically f-monotone if the following relation holds: $f(\mathbf{z}) \geq f(\mathbf{y}) \Rightarrow \lim_{n \to \infty} P(g(\mathbf{z}, S_n) \geq g(\mathbf{y}, S_n)) = 1$. As an example of asymptotic monotone function consider $g(\mathbf{z}, S_n) = 1/d_k(\mathbf{z}, S_n)$; where $d_k(\mathbf{z}, S_n)$ is the distance from $\mathbf{z}$ to its (random) k–nearest neighbour in S_n (for $1 \leq k < n$). Therefore, a suitable estimator for K_f is given by the product of two asymptotically f-monotone functions.

The estimation of the DKD is based on an asymptotically f-monotone function $g(\cdot, S_n)$ which is proportional to a non–parametric consistent estimator of f –the density function that corresponds to the projection coefficients. Our density estimation method is based on the One–Class Neighbor Machine (OCNM), a well known nonparametric density level set estimator [22, 23]. The ν-level set of f is defined as $V_\nu(f) = \{\boldsymbol{\lambda} \in \mathbb{R}^d : f(\boldsymbol{\lambda}) \geq \alpha_\nu\}$, such that $P(V_\nu(f)) = 1 - \nu$ for $0 < \nu < 1$. Some comments about the density estimation method are in order. The OCNM is a consistent estimator of ν-level sets under mild conditions on f. In addition, the OCNM relies on a linear–convex optimization problem, entailing important computational advantages in comparison to other standard nonparametric density estimation approaches such as kernel method. For more details about the OCNM we refer to [23].

The estimation of DKD via the OCNM is straightforward. Given a sample of n discretised curves as in Eq. (1), $\mathcal{D}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and its corresponding projection coefficients $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n\}$ according to Eq. (5); we use the OCNM to estimate density contour clusters $\widehat{V}_\nu$ around the estimated mode $\widehat{\mathbf{m}}$ for an increasing sequence $\boldsymbol{\nu} \equiv \{\nu_1, \ldots, \nu_m\}$, such that $0 \leq \nu_1 < \cdots < \nu_m \leq 1$. Notice that $\widehat{\mathbf{m}} \in \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n\}$ corresponds to the sample curve which belongs to the highest ν–density level set. We consider the following asymptotic f–monotone function: $g(\boldsymbol{\lambda}, \mathcal{D}_n) = \sum_{i=1}^{m} i \mathbb{I}_{\widehat{V}_i}(\boldsymbol{\lambda})$, where $\mathbb{I}_{\widehat{V}_i}(\boldsymbol{\lambda})$ is the indicator function that take value 1 if $\boldsymbol{\lambda} \in \widehat{V}_{\nu_i}$ and 0 otherwise. In this sense, the estimated $\text{DKD}(\tilde{x}(t), \mathcal{D}_n, K_f) \equiv g(\boldsymbol{\lambda}, \mathcal{D}_n)g(\widehat{\mathbf{m}}, \mathcal{D}_n)$ order the sampled curves around the estimated center $\widehat{x}_{\text{mo}}(t) = \widehat{\mathbf{m}}^T \boldsymbol{\Phi}(t)$; i.e. high values of DKD corresponds to estimated central curves and vice-verse.

4 Functional outlier detection

Depth measures induces a centre-outward ordering, and constitute a natural tool to assess the presence of outlying curves in the sample. From an statistical perspective, the analysis of atypical curves relies on the distribution of $\mathrm{DKD}(\widetilde{X}(t), f, K_f)$. Nevertheless, nither the exact nor the asymptotic distribution of DKD is known. For this reason, we employ bootstrap methods to approximate such distribution and determine which curves in the sample are more likely to be outliers. In Algorithm 1, we present our bootstrap method that resembles the procedure given in [24]. For the analysis of outlier functional data, a suitable threshold $q \in (0, 1)$ (q is intended to be the type I error on the identification method) needs to be specified in advance. In the case of Australian mortality curves discussed in Sect. 5, we choose $q = 0.01$; nevertheless in practice we recommend to conduct a sensitivity analysis regarding the value given to this sensible parameter.

```
Algorithm 1 Bootstrap based outlier identification method.
Given data $\mathcal{D}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and the corresponding coefficients $\{\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n\}$:
  1. Compute $\mathrm{DKD}_1 \equiv \mathrm{DKD}(\tilde{x}_1(t), \mathcal{D}_n, K_f), \dots, \mathrm{DKD}_n \equiv \mathrm{DKD}(\tilde{x}_n(t), \mathcal{D}_n, K_f)$ and the empirical quantile $F_n^{-1}(q) = \mathrm{DKD}_{n(i)}$ where $\mathrm{DKD}_{n(i)}$ is the order $i$ statistic such that $(i-1)/n < q \leq i/n$.
  2. Resampling: Generate $B \gg 0$ bootstrap samples $\{\mathcal{D}_n^{(b)}\}_{b=1}^B$ where on each sample $\mathcal{D}_n^{(b)}$ the vector $\mathbf{x}_l$ have a probability of being sampled equal to $\mathrm{DKD}_l / \sum_{i=1}^n \mathrm{DKD}_i$ for $l = 1, \dots, n$ and $b = 1, \dots, B$.
  3. For each $\mathcal{D}_n^{(b)}$, produce a small perturbations on the raw bootstraped sample data as follows: $\mathbf{x}_l^{(b)} \mapsfrom \mathbf{x}_l^{(b)} + \mathbf{z}_l^{(b)}$ for $l = 1, \dots, n$; where $\mathbf{z}$ is sampled from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{S}_\epsilon = \frac{\epsilon}{n} \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$, being $\epsilon > 0$ is a smoothing parameter.
  4. Bootstrap empirical percentile estimation:
     for (b in 1 to B): Compute $F_n^{-1,(b)}(q)$ from $\mathcal{D}_n^{(b)}$ as in Step 1.
     Set $\mathrm{DKD}_{\mathrm{me}}(q)$ as the median of $\{F_n^{-1,(1)}(q), \dots, F_n^{-1,(B)}(q)\}$ and identify as outlying curves the elements in the sample such that $\mathrm{DKD}_i < \mathrm{DKD}_{\mathrm{me}}(q)$.
  5. Remove the curves identified as outliers in Step 4 and return to Step 1 until no more outliers are found in the remaining data.
```

Some comments on Algorithm 1 are in order. Step 1 and 2 entails standard statistical procedures. Step 3 involves some parametric perturbation on data in order to produce a robust estimation of the true DKD quantile.

5 Experiments

In this section we develop numerical experiments to assess the performance of the proposed depth measure in the task of functional outlier detection. The implementation of DKD is available in the `'bigdatadist'` R-package, and the R code to reproduce the experiments are provided as supplementary material.

Hereafter, for the functional data representation, we consider a Gaussian kernel function $K(t, s) = \exp(-\sigma \|t - s\|^2)$. The bandwidth parameter σ and the dimension of the basis function system d given in Eq. (5) where cross-validated through grid search. To compare our method we consider several functional depth methods: the modified band depth (MBD) [18] already implemented in the R-package `'depthTools'` [25], the random Tukey depth (RTD) and the h-mode depth (HMD), see [16, 17], implemented in the R-package `'fda-usc'` [26], and the functional spatial depth (FSD), see [27].

5.1 Monte Carlo simulation study

Simulation setting We simulated a sample of $n = 400$ curves, where a small proportion $q \in [0, 1]$, known a priori, presents an atypical pattern. The remaining $n(1 - q)$ curves are considered the main data. We study the performance of DKD over three data configurations (scenarios a, b and c) and for three different values of the contamination parameter $q \in \{1\%, 5\%, 10\%\}$. Specifically, we consider the following generating processes:

$$X_l(t) = \sum_{j=1}^{4} \xi_j \sin(j\pi t) + \varepsilon_l(t), \ \text{ for } l = 1, \dots, (1 - q)n,$$

$$Y_l(t) = \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \varepsilon_l(t), \ \text{ for } l = 1, \dots, qn/2,$$

$$Z_l(t) = \sum_{j=1}^{4} \eta_j \sin(j\pi t) + \varepsilon_l(t), \ \text{ for } l = 1, \dots, qn/2,$$

where $t \in T \equiv [0, 1]$ and $\varepsilon_\cdot(t)$ are independent and zero mean random error functions.

(a) **Symmetric scenario** $(\xi_1, \dots, \xi_4)$ follows a multivariate normal distribution (MND) with mean $\boldsymbol{\mu}_\xi = (4, 2, 4, 1)$ and diagonal co-variance matrix $\boldsymbol{\Sigma}_\xi = \mathrm{diag}(5, 2, 2, 1)$. To generate magnitude outliers, we consider $(\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ following a MND with parameters $\boldsymbol{\mu}_\zeta = 2.5\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\zeta = (2.5)^2\boldsymbol{\Sigma}_\xi$. To generate shape outliers, we specify $(\eta_1, \eta_2, \eta_3, \eta_4)$ as MND with parameters $\boldsymbol{\mu}_\eta = (4, -2, 1, 3)$ and $\boldsymbol{\Sigma}_\eta = \boldsymbol{\Sigma}_\xi$.

(b) **Asymmetric scenario** In this case, $\{\xi_1, \ldots, \xi_4\}$ are independent Chi-square distributed random variables (ICrv) with 16, 16, 12, 12 degrees of freedom respectively; while $\{\zeta_1, \zeta_2, \zeta_3, \zeta_4\}$ are ICrv with 40, 40, 30, 30 degrees of freedom respectively; and $(\eta_1, \eta_2, \eta_3, \eta_4)$ follows a MND with mean $\boldsymbol{\mu}_\eta = (18, 16, 8, -10)$ and variance $\boldsymbol{\Sigma}_\eta = \mathrm{diag}(15, 12, 12, 15)$.

(c) **Bi-modal scenario:** $\boldsymbol{\xi}$ follows a mixture of two MND as follows: $(\xi_1, \ldots, \xi_4) \sim 0.5N_4(\boldsymbol{\mu}_{\xi,1}, \boldsymbol{\Sigma}_\xi) + 0.5N_4(\boldsymbol{\mu}_{\xi,2}, \boldsymbol{\Sigma}_\xi)$, where $\boldsymbol{\mu}_{\xi,1} = (1, 1, 1, 1)$, $\boldsymbol{\mu}_{\xi,2} = (9, 9, 9, 9)$, and $\boldsymbol{\Sigma}_\xi = \mathrm{diag}(5, 2, 2, 1)$ is a diagonal covariance matrix. Outliers are generated as follows: $(\zeta_1, \zeta_2, \zeta_3, \zeta_4) \sim N_4(\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$, where $\boldsymbol{\mu}_\zeta = (8, 4, 8, 2)$ and $\boldsymbol{\Sigma}_\zeta = 4\boldsymbol{\Sigma}_\xi$; and $(\eta_1, \eta_2, \eta_3, \eta_4) \sim 0.5N_4(\boldsymbol{\mu}_{\eta,1}, \boldsymbol{\Sigma}_\eta) + 0.5N_4(\boldsymbol{\mu}_{\eta,2}, \boldsymbol{\Sigma}_\eta)$ with parameters $\boldsymbol{\mu}_{\eta,1} = (-4, 4, -1, -3)$, $\boldsymbol{\mu}_{\eta,2} = (5, 5, 5, 5)$ and $\boldsymbol{\Sigma}_\eta = 0.5\,\mathrm{diag}(1, 1, 1, 1)$.

To illustrate the generating process, in Fig. 1, we show one instance of the simulated curves in the scenarios (a) to (c) for $q = 10\%$.

Results
To evaluate the outlier detection performance of the DKD we develop a Monte Carlo simulation study. To this end, for each scenario and contamination level, we generate $M = 1000$ data replications and report the following average metrics: True Positive Rate TPR $= \frac{\mathrm{TP}}{q \times n}$ (sensitivity); True Negative Rate TNR $= \frac{\mathrm{TN}}{(1-q) \times n}$ (specificity), and the area under the ROC curve aROC.

The Monte Carlo simulation results are presented in Table 1. The DKD resents a remarkable performance among other functional depths measures in the three scenarios considered for different levels of contamination $q \in \{1\%, 5\%, 10\%\}$. However in the case where $q = 1\%$ the difference with respect to other depth measures is not statistically significant. Among the three different scenarios the performance of the DKD increases with respect to the rest of the competitors in scenario (b) and (c), where me move away from the Gaussian setting.

5.2 Detecting outlying curves in the Australian mortality database

For this experiment we consider age-specific log-mortality rates of Australian males. The data is publicly available in the R-package `'fds'`[1] [28]. The data set consists of 103 curves that are registered over a range of 0-100 age cohorts, with each curve corresponding to one year between 1901 and 2003. As shown in Fig. 2, for low-age cohorts (until 12 years approximately), the mortality rates present a decreasing trend and then start to grow until late ages, where all cohorts achieve a 100% mortality rate.

Results Outlier detection is an unsupervised learning problem. Therefore, we do not know a-priori if there are any outlying curves (years) in this data set. Following the procedure detailed in Algorithm 1, we estimate the cut-off point q such that mislabelling of correct observations as outliers is about 1%. The results are presented in Table 2 and Fig. 2. The outliers detected by the DKD are years: *1919*, *1943-1945*. The year 1919 corresponds to the influenza pandemic episode that caused around 15, 000 casualties, as the virus spread through Australia. Given the pattern in the curve corresponding to year 1919, this curve can be visually identified and labelled as a *magnitude outlier*. In a related study that rely on the same data [29], the authors identified 1919 as an anomalous observation as well.

Moreover, these curves exhibit a distinctive shape compared to the others, lacking any prominent points (age-cohorts) that would indicate their dissimilarity. As a result, visually detecting them is exceedingly difficult. Specifically, for the age-cohorts between 15 to 40, one can observe a discrepancy in the curve patterns, so they can be considered as shape outliers. It is particularly relevant to identify outliers in this dataset to achieve reliable predictions of mortality curves. With the onset of the COVID-19 pandemic, improved predictions can aid governments in making better-informed decisions.

The DKD handle both type of functional outliers, as follows from the simulations in the previous section. In Table 2 we compare our findings against other standard benchmarks introduced in the Monte Carlo simulation study as well as the *Outliergram* -particularly devised to capture *shape outliers*- [7] and the *Functional Boxplot* [8] (see Supplementary Material). Both devices are already implemented in the R-packages `'roahd'` and `'fda'` respectively. The Outliergram identifies the year '1914' as an outlier, while the Functional Boxplot has not pinpointed any anomalous observation. These tools are based on the Modified Band Depth measure. This approach is very efficient identifying magnitude outliers or shape outliers that present a sharp difference with respect to the rest of the sample. In this case the shape outliers are hidden within the rest of the data and this could explain the poor performance of the method. Finally, is also interesting to notice that the mortality curve corresponding to year 2003 is identified as an outlier for all the competitors. Since year 2003 is located on the outwards of the data with respect to the deepest curve, we tend to see these find as a type-II-error in the analysis.

[1] Sourced primary from the Australian Demographic Data Bank

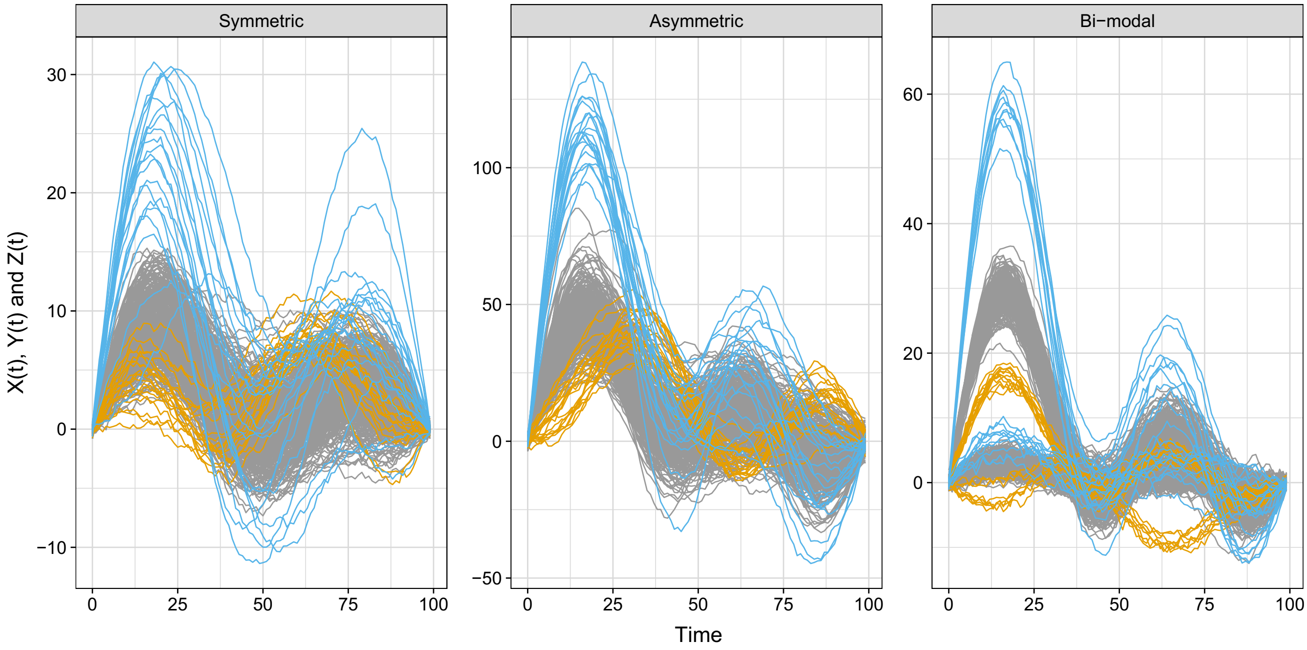

Fig. 1 Functional data: 400 curves corresponding to $q = 10\%$ In grey (—), we represent instances of regular curves $X(t)$, abnormal curves $Y(t)$ depicted in amber (—), and $Z(t)$ in light-blue (—). (color figure online)

Fig. 2 Australian Mortality data: regular curves in gray (—), deepest curve black-dashed (- - - -) and outliers detected by the DKD in red (—) (color figure online)

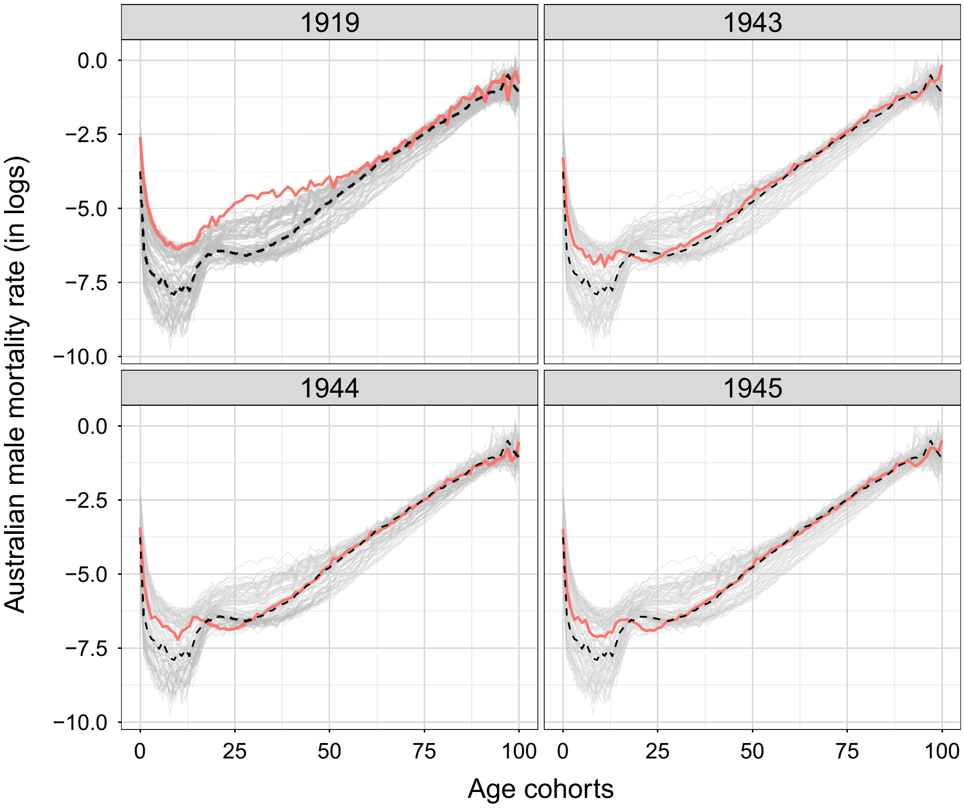

Table 1 Scenarios and contamination percentages q in columns

Method	Metric	Scenario A			Scenario B			Scenario C		
		10%	5%	1%	10%	5%	1%	10%	5%	1%
MBD	TPR	67.96	58.57	36.12	81.20	74.38	52.20	30.54	27.58	24.425
		(5.27)	(7.97)	(18.43)	(4.27)	(7.076)	(21.28)	(5.20)	(7.41)	(17.37)
	TNR	96.44	97.82	99.35	97.91	98.65	99.52	92.28	96.18	99.23
		(0.58)	(0.42)	(0.18)	(0.47)	(0.37)	(0.21)	(0.57)	(0.39)	(0.17)
	aROC	95.31	95.40	95.50	98.44	98.49	98.54	58.49	58.84	57.96
		(1.37)	(1.84)	(3.84)	(0.57)	(0.73)	(1.44)	(3.16)	(4.38)	(10.41)
HMD	TPR	80.07	76.77	67.00	83.04	79.84	70.47	79.08	79.24	67.17
		(4.41)	(6.86)	(18.30)	(4.27)	(6.14)	(16.09)	(5.75)	(6.99)	(18.66)
	TNR	97.78	98.77	99.66	98.11	98.93	99.70	97.67	98.90	**99.68**
		(0.58)	(0.42)	(0.18)	(0.47)	(0.37)	(0.21)	(0.57)	(0.39)	(0.17)
	aROC	97.67	97.90	98.00	98.90	99.20	99.40	96.24	96.80	**97.19**
		(1.08)	(1.47)	(3.29)	(0.44)	(0.47)	(0.74)	(1.83)	(2.35)	(5.49)
RTD	TPR	72.07	64.87	49.87	83.79	78.02	63.70	28.27	26.22	26.67
		(6.95)	(9.45)	(17.98)	(5.55)	(8.66)	(15.97)	(5.48)	(7.67)	(17.76)
	TNR	96.89	98.15	99.49	98.19	98.84	99.63	92.02	96.11	99.25
		(0.77)	(0.49)	(0.18)	(0.61)	(0.45)	(0.16)	(0.61)	(0.40)	(0.18)
	aROC	96.13	96.30	96.46	98.84	98.96	99.10	62.85	63.50	62.83
		(1.65)	(1.98)	(3.85)	(0.73)	(0.77)	(1.01)	(3.95)	(5.54)	(12.55)
FSD	TPR	74.89	69.59	54.65	86.18	82.80	69.97	29.57	27.10	26.90
		(4.47)	(6.84)	(17.21)	(3.31)	(5.19)	(15.63)	(5.45)	(7.60)	(17.76)
	TNR	97.21	98.40	99.54	98.46	99.09	99.69	92.17	96.16	99.26
		(0.49)	(0.36)	(0.17)	(0.36)	(0.27)	(0.15)	(0.60)	(0.40)	(0.17)
	aROC	96.79	97.13	97.29	99.13	99.35	99.47	60.81	61.42	60.63
		(1.15)	(1.54)	(3.43)	(0.38)	(0.41)	(0.63)	(3.56)	(4.98)	(12.05)
DKD	TPR	**88.39**	**86.37**	**79.35**	**90.86**	**87.51**	**74.85**	**86.26**	**82.83**	**67.52**
		(3.94)	(6.10)	(17.54)	(3.37)	(5.66)	(18.01)	(4.03)	(6.11)	(19.52)
	TNR	**98.71**	**99.28**	**99.79**	**98.98**	**99.34**	**99.746**	**98.47**	**99.09**	99.67
		(0.43)	(0.32)	(0.17)	(0.37)	(0.29)	(0.18)	(0.44)	(0.32)	(0.19)
	aROC	98.18	98.58	**98.80**	98.95	99.28	**99.58**	**96.57**	**96.84**	97.14
		(0.91)	(1.21)	(2.83)	(0.41)	(0.37)	(0.47)	(1.71)	(2.23)	(5.40)

In rows, different methods and average TPR, TNR and aROC (in a scale of 10^2). The corresponding standard-error are reported in parenthesis. Largest value per column per metric (in bold)

Table 2 Outliers curves detected by the different methods ($q = 1\%$).

	MBD	HMD	RTD	FSD	DKD
1st iteration	2001-2003	1919; 2003	2003	1919; 2003	1919; 1944-1945
2nd iteration	2000	–	–	–	1943

6 Conclusions

In this work, we propose a density kernel depth (DKD) measure as a tool for detecting outliers in functional data. The DKD method relies on a stable low-dimensional representation of curves in a linear subspace of a Reproducing Kernel Hilbert Space. We discuss interesting properties associated with the DKD and address the subtleties involved in its estimation. We showcase the performance of our method through a Monte Carlo simulation study using three different scenarios: symmetric, asymmetric, and bi-modal models for functional data. The DKD is able to identify the presence of shape and magnitude outliers not only in simulated data but also in the analysis of Australian male mortality rate curves. Although the DKD has a remarkable performance in identifying outliers that are hidden within the rest of the curves and are extremely difficult to spot visually, we suggest combining DKD analysis with the outliergram [7] and functional boxplot methods [8] to achieve more robust results.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41060-023-00420-w.

Author contributions Authors contributed equally to this work.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability Data is available in the `'fds'` R-package.

Declarations

Conflicts of interest None of the authors has a conflict of interest.

References

1. Ramsay, J.O.: When the data are functions. Psychometrika **47**(4), 379–396 (1982)
2. Beyaztas, U., Lin Shang, H.: A robust functional partial least squares for scalar-on-multiple-function regression. J. Chemom. **36**(4), 3394 (2022)
3. Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., Boente, G., Fraiman, R., Brumback, B., Croux, C., et al.: Robust principal component analysis for functional data. TEST **8**, 1–73 (1999)
4. Dai, W., Genton, M.G.: Multivariate functional data visualization and outlier detection. J. Comput. Graph. Stat. **27**(4), 923–934 (2018)
5. García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A.: A review of robust clustering methods. Adv. Data Anal. Classif. **4**, 89–109 (2010)
6. Staerman, G., Adjakossa, E., Mozharovskyi, P., Hofer, V., Sen Gupta, J., Clémençon, S.: Functional anomaly detection: a benchmark study. Int. J. Data Sci. Anal. **16**, 101–117 (2022)
7. Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. Biostatistics **15**(4), 603–619 (2014)
8. Sun, Y., Genton, M.G.: Functional boxplots. J. Comput. Graph. Stat. **20**(2), 316–334 (2011)
9. Ramsay, J.O.: Functional data analysis (2006)
10. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer, Berlin (2006)
11. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer, Berlin (2011)
12. Kimeldorf, G., Wahba, G.: Some results on tchebycheffian spline functions. J. Math. Anal. Appl. **33**(1), 82–95 (1971)
13. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. **39**(1), 1–49 (2002)
14. Muñoz, A., González, J.: Representing functional data using support vector machines. Pattern Recogn. Lett. **31**(6), 511–516 (2010)
15. Fraiman, R., Muniz, G.: Trimmed means for functional data. TEST **10**(2), 419–440 (2001)
16. Cuesta-Albertos, J.A., Nieto-Reyes, A.: The random Tukey depth. Comput. Stat. Data Anal. **52**(11), 4979–4988 (2008)
17. Cuevas, A., Febrero, M., Fraiman, R.: Robust estimation and classification for functional data via projection-based depth notions. Comput. Stat. **22**(3), 481–496 (2007)
18. López-Pintado, S., Romo, J.: On the concept of depth for functional data. J. Am. Stat. Assoc. **104**(486), 718–734 (2009)
19. Nagy, S.: Statistical Depth for Functional Data. Univerzita Karlova, Matematicko-fyzikální fakulta (2016)
20. Nieto-Reyes, A., Battey, H.: A topologically valid definition of depth for functional data. Stat. Sci. **31**(1), 61–79 (2016)
21. Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Stat. **28**, 461–482 (2000)
22. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)
23. Muñoz, A., Moguerza, J.M.: Estimation of high-density regions using one-class neighbor machines. IEEE Trans. Pattern Anal. Mach. Intell. **28**(3), 476–480 (2006)
24. Febrero, M., Galeano, P., González-Manteiga, W.: Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. Environmetrics **19**(4), 331–345 (2008)
25. Lopez-Pintado, S., Torrente, A.: depthtools: Depth tools package. R package version 0.4. http://CRAN.R-project.org/package=depthTools (2013)
26. Febrero-Bande, M., Oviedo de la Fuente, M.: Package "fda. usc": Functional data analysis and utilities for statistical computing. R package version 1.4.0 (2013)
27. Sguera, C., Galeano, P., Lillo, R.: Spatial depth-based classification for functional data. TEST **23**(4), 725–750 (2014)
28. Shang, H., Hyndman, R.: FDS: functional data sets. R package version 1.7 (2013)
29. Hyndman, R.J.: Computing and graphing highest density regions. Am. Stat. **50**(2), 120–126 (1996)