entropy

MDPI

Article

# Entropy Measures for Stochastic Processes with Applications in Functional Anomaly Detection

**Gabriel Martos [1,*], Nicolás Hernández [2], Alberto Muñoz [2,*] and Javier M. Moguerza [3]**

[1] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Buenos Aires C1428EGA, Argentina
[2] Department of Statistics, Universidad Carlos III de Madrid, 28903 Getafe, Spain; nihernan@est-econ.uc3m.es
[3] Department of Computer Science and Statistics, University Rey Juan Carlos, 28933 Móstoles, Spain; javier.moguerza@urjc.es
* Correspondence: gabriel.martos@ic.fcen.uba.ar (G.M.); albmun@est-econ.uc3m.es (A.M.); Tel.: +54-11-45763375 (G.M.); +34-91-6249579 (A.M.)

Received: 5 December 2017; Accepted: 2 January 2018; Published: 11 January 2018

**Abstract:** We propose a definition of entropy for stochastic processes. We provide a reproducing kernel Hilbert space model to estimate entropy from a random sample of realizations of a stochastic process, namely functional data, and introduce two approaches to estimate minimum entropy sets. These sets are relevant to detect anomalous or outlier functional data. A numerical experiment illustrates the performance of the proposed method; in addition, we conduct an analysis of mortality rate curves as an interesting application in a real-data context to explore functional anomaly detection.

**Keywords:** entropy; stochastic process; minimum-entropy sets; anomaly detection; functional data

## 1. Introduction

The family of $\alpha$-entropies, originally proposed by Rényi [1], plays an important role in information theory and statistics. Consider a random variable $Z$ distributed according to a measure $F$ that admits a probability density function $f$. Then, for $\alpha \geqslant 0$ and $\alpha \neq 1$, the $\alpha$-entropy of $Z$ is computed as follows:

$$H_\alpha(Z) = \frac{1}{1-\alpha} \log\left(V_\alpha(Z)\right) \tag{1}$$

where $V_\alpha(Z) = \mathbb{E}_F\{f^{\alpha-1}\}$, and $\mathbb{E}_F$ stands for the expected value with respect to the $F$ measure. Several renowned entropy measures in the statistical literature are particular cases in the family of $\alpha$-entropies. For instance, when $\alpha = 0$, we obtain the Hartley entropy; when $\alpha \to 1$, then $H_\alpha$ converges to the Shannon entropy; and when $\alpha \to \infty$, then $H_\alpha$ converges to the Min-entropy measure. The contribution of this paper is two-fold. Firstly, we propose a natural definition of entropy for stochastic processes that extends the previous one and a suitable sample estimator for the observation of partial realizations of the process, the typical framework when dealing with functional data. We also show that Minimal Entropy Sets (MES), as formally defined in Section 3, are useful to solve anomaly detection problems, a common task in almost all data analysis contexts.

The paper is structured as follows: In Section 2, we introduce a definition of entropy for a stochastic process and suitable sample estimators for this measure. In Section 3, we show how to estimate minimum-entropy sets of a stochastic process in order to discover atypical functional data in a sample. Section 4 illustrates the theory with simulations and examples, and Section 5 concludes the work.

## 2. Entropy of a Stochastic Process

In this section, we extend the definition of entropy to a stochastic process. For the sequel, let $(\Omega, \mathcal{F}, P)$ be a probability space, where $\mathcal{F}$ is the $\sigma$-algebra in $\Omega$ and $P$ a $\sigma$-finite measure. We

consider random elements (functions) $X(\omega, t) : \Omega \times T \to \mathbb{R}$ in a metric space $(T, \tau)$. As usual in the case of functional data, the realizations of the random elements $X(\omega, \cdot)$ are assumed in $C(T)$, the space of real continuous functions in a compact domain $T \subset \mathbb{R}$ endowed with the uniform metric.

The first step is to consider a suitable representation for the stochastic process. We make use of the well-known Karhunen–Loève expansion [2] (p. 25, Theorem 1.5). Let $X(\omega, t)$ be a centered (zero-mean) stochastic process with continuous covariance function $K_X(s, t) = \mathbb{E}(X(\omega, s)X(\omega, t))$, then there exists a basis $\{e_i\}_{i \geqslant 1}$ of $C(T)$ such that for all $t \in T$:

$$X(\omega, t) = \sum_{i=1}^{\infty} \xi_i(\omega) e_i(t), \tag{2}$$

where the sequence of random coefficients $\xi_i(\omega) = \int_T X(\omega, t) e_i(t) dt$ comprises zero mean random variables with (co)variance $\mathbb{E}(\xi_i \xi_j) = \delta_{ij} \lambda_j$, being $\delta_{ij}$ the Kronecker delta and $\{\lambda\}_{j \geqslant 1}$ the sequence of eigenvalues associated with the eigenfunctions of $K_X(s, t)$.

The equality in Equation (2) must be understood in the mean square sense, that is:

$$\lim_{d \to \infty} \mathbb{E}\{(X(\omega, t) - \sum_{i=1}^{d} \xi_i(\omega) e_i(t))^2\} = 0, \tag{3}$$

uniformly in $T$. Therefore, we can always consider a $\varepsilon$-near representation $X_d(\omega, t) = \sum_{i=1}^{d} \xi_i(\omega) e_i(t)$ such that for all $\varepsilon$ arbitrarily small, there exists an integer $D$ such that for $d \geqslant D$, then $\tau(X, X_d) = \sup_{t \in T} |X(\omega, t) - X_d(\omega, t)| \leqslant \varepsilon$. From this result, it is possible to establish a suitable way to approximate the entropy of a random element $X(\omega, t)$ according to the distribution of the "representation coefficients" $\{\xi_i(\omega)\}_i^d$ obtained from $X_d(\omega, t)$.

**Definition 1** ($d$-truncated entropy for stochastic processes)**.** *Let $X$ be a centered stochastic process with a continuous covariance function. Consider the truncation $X_d(\omega, t) = \sum_{i=1}^{d} \xi_i(\omega) e_i(t)$ and the random vector $Z = (\xi_1, \ldots, \xi_d)$; then, the $d$-truncated entropy of $X$ is defined as $H_\alpha(X, d) = H_\alpha(Z)$.*

The "approximation error" when computing the entropy of the stochastic process $X$ with Definition 1 decreases monotonically with the number of terms retained in the Karhunen–Loève expansion, at a rate that depends on the decay of the spectrum of the covariance function $K_X(s, t)$. In general, the more autocorrelated the process is, the more quickly the eigenvalues of $K_X(s, t)$ converge to zero. In practical functional data applications (see for instance the mortality-rate curves in Section 4), the autocorrelation is usually strong, and the truncation parameter $d$ will be small when approximating the entropy of the process. The next example illustrates the definition.

**Example 1.** *[Gaussian process] When $X$ is a Gaussian Process (GP), the coefficients in the Karhunen–Loève expansion have the further property that they are independent and zero-mean normally distributed random variables. Therefore, the Shannon entropy ($\alpha = 1$) of $X$ can be approximated with the truncated version of the GP as follows:*

$$H_1(X, d) = \frac{1}{2} \log(2\pi e)^d \det(\Sigma),$$

*where $\Sigma$ is the diagonal covariance matrix with elements $[\Sigma]_{i,j} = \mathbb{E}(\xi_i \xi_j)$ for $i, j = 1, \ldots, d$.*

In practice, we can only observe some realizations of the stochastic process $X$, and these observations are sparsely registered. Therefore, to estimate the entropy of $X(\omega, t)$ from a random sample of discrete realizations of a stochastic process, a first task is the representation of these paths by means of continuous functions. To this end, we consider a reproducing kernel Hilbert space $\mathcal{H}$ of functions, associated with a positive definite and symmetric kernel function $K : T \times T \to \mathbb{R}$.

*Estimating Entropy in a Reproducing Kernel Hilbert Space*

Most functional data analysis approaches for representing raw data suggest proceeding as follows: (i) choose an orthogonal basis of functions $\Phi = \{\phi_1, \ldots, \phi_N\}$, where each $\phi_i$ belongs to a general function space $\mathcal{H}$; and (ii) represent each functional datum by means of a linear combination in the Span($\Phi$) [3,4]. Our choice is to consider $\mathcal{H}$ as a Reproducing Kernel Hilbert Space (RKHS) of functions [5]. In this case, the elements in the spanning set $\Phi$ are the eigenfunctions associated with the positive-definite and symmetric kernel function $K : T \times T \to \mathbb{R}$ that span $\mathcal{H}$ [5] (Moore-Aronszajn Theorem p. 19).

In our setting, the functional representation problem can be framed as follows: We have available $m$ discrete observations, that is a realization path $x(t_1), \ldots, x(t_m)$ of the stochastic element $X(\omega, t)$. We also assume that the discrete path $\{x(t_i), t_i\}_{i=1}^m$, as usual when dealing with real data, contains zero mean *iid* error measurements. Then, the functional data estimator, denoted onwards as $\tilde{x}(t)$, is obtained solving the following regularization problem:

$$\tilde{x}(t) := \arg\min_{g \in \mathcal{H}} \sum_{i=1}^{m} V(x(t_i), g(t_i))^2 + \gamma \Omega(g), \tag{4}$$

where $V$ is a strictly convex functional with respect to the second argument, $\gamma > 0$ is a regularization parameter, frequently chosen by cross-validation, and $\Omega(g)$ is a regularization term. By the representer theorem [6,7] (Theorem 5.2, p. 91, Proposition 8, p. 51), the solution of the problem stated in Equation (4) exists, is unique and admits a representation of the form:

$$\tilde{x}(t) = \sum_{i=1}^{m} a_i K(t, t_i). \tag{5}$$

In the particular case of a squared loss function $V(w, z) = (w - z)^2$ and considering $\Omega(g) = \int_T g^2(t)\, dt$, the coefficients of the linear combination in Equation (5) are obtained solving the following system:

$$(\gamma m \mathbf{I} + \mathbf{K})\mathbf{a} = \mathbf{y}, \tag{6}$$

where $\mathbf{a} = (a_1, \ldots, a_m)^T$, $\mathbf{y} = (x(t_1), \ldots, x(t_m))^T$, $\mathbf{I}$ is the identity matrix of order $m$ and $\mathbf{K}$ is the Gram matrix with the kernel evaluations, $[\mathbf{K}]_{k,l} = K(t_k, t_l)$, for $k = 1, \ldots, m$ and $l = 1, \ldots, m$. To relate the Karhunen–Loève expansion in Equation (2) to the RKHS representation, we make use of Mercer's theorem [2] (Lemma 1.3, p. 24), then $K_X(s, t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(t)$, where $\lambda_j$ is the eigenvalue associated with the orthonormal eigenfunction $\phi_j$ for $j \geqslant 1$, and invoking the reproducing property, then:

$$\begin{array}{rcl} X(\omega, t) & = & \langle X(\omega, s), K_X(s, t) \rangle \\ & = & \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \int_T X(\omega, s) \phi_j(s) ds. \end{array} \tag{7}$$

Therefore, following Equation (2), $\xi_j(\omega) := \sqrt{\lambda_j} \int_T X(\omega, s) \phi_j(s) ds$ and $e_j(t) = \sqrt{\lambda_j} \phi_j(t)$; and the connection is clearly established. When working with discrete realizations of a stochastic process, we must solve two sequential tasks. First, we need to represent raw data as functional data and later find a truncated representation of the function. To this end, when combining Equation (5) with Mercer's theorem and the reproducing property, we obtain:

$$\tilde{x}_d(t) = \sum_{j=1}^{d} \sqrt{\lambda_j} \phi_j(t) \sqrt{\lambda_j} \Big( \sum_{i=1}^{m} a_i \phi_j(t_i) \Big),$$

and now, $z_j := \sqrt{\lambda_j} \sum_{i=1}^{m} a_i \phi_j(t_i)$ is the realization of the random variable $\xi_j$ for $j = 1, \ldots, d$; see [8] for further details. For some kernel functions, for instance the Gaussian kernel, the associated sequence

of eigen-pairs $(\lambda_j, \phi_j)$ for $j \geqslant 1$ is known [9] (pp. 10), and we can obtain an explicit value for all $z_j$. If not, let $(\lambda_j, \mathbf{v}_j)$ be the $j$-eigenpair associated with the kernel matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$, then $z_j = \sqrt{\lambda_j} \sum_{i=1}^{m} a_i v_{i,j}$ for $j = 1, \ldots, d$.

In practice, given a sample of $n$ discrete paths (realizations) of the stochastic process $X$, say $\{x_l(t_1), \ldots, x_l(t_m)\}$ for $l = 1, \ldots, n$, a suitable input to estimate entropy in Definition 1 is to consider the set of multivariate vectors $\mathbf{z}_l = (z_{l,1}, \ldots, z_{l,d})$ for $l = 1, \ldots, n$, as formally proposed in the next definition.

**Definition 2** (*K*-entropy estimation of a stochastic process)**.** *Let $\{x_1(t_i), \ldots, x_n(t_i)\}$ for $i = 1, \ldots, m$ be a discrete random sample of $X$, and let $\{(\lambda_j, \mathbf{v}_j)\}_{j=1}^{d}$ be the eigen-pairs of the kernel matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$, where $d = \text{rank}(\mathbf{K})$. Consider the corresponding finite dimensional representation $S_n := \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, where $\mathbf{z}_l = (z_{l,1}, \ldots, z_{l,d}) \in \mathbb{R}^d$ for $l = 1, \ldots, n$ and $z_{l,j} = \sqrt{\lambda_{l,j}} \sum_{i=1}^{m} a_{l,i} v_{i,j}$ for $j = 1, \ldots, d$. Then, the estimated kernel entropy of $X$ is defined as $\widehat{H}_\alpha(X, K) = \widehat{H}_\alpha(Z)$.*

In Definition 2, $\widehat{H}_\alpha(Z)$ denotes the estimated entropy using the (finite dimensional) representation coefficients $S_n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. In Section 3, we formally introduce two approaches to estimate entropy departing from $S_n$. The next example illustrates the estimation procedure in the context of GPs in Example 1.

**Illustration with Example 1:** Consider 100 realizations of a GP as follows: 50 curves from $X(t) = \sum_{i=1}^{3} \xi_i e_i(t)$ and another 50 curves from $Y(t) = \sum_{i=1}^{3} \zeta_i e_i(t)$; where $e_i(t)$ is a Fourier basis in $T = [0, 1]$, $\xi_i \sim N(\mu = 0, \sigma^2 = 0.5)$, and $\zeta_i \sim N(\mu = 0, \sigma^2 = 2)$ are independent normally distributed random variables (r.v.) for $i = 1, 2, 3$.

In Figure 1 (left), we illustrate the realizations of the stochastic processes, in black ("—") the sample paths of $X(t)$ and in red ("—") the paths corresponding to $Y(t)$. In Figure 1 (right), we show the distribution of the linear combination coefficients $\{(z_1, z_2, z_3)_l, (w_1, w_2, w_3)_l\}_{l=1}^{50}$ corresponding to these paths. Following Example 1, we estimate the covariance functions $\widehat{\Sigma}_X$ and $\widehat{\Sigma}_Y$ using the respective coefficients and plug this covariance matrix into the Shannon entropy expression to obtain the estimated entropies $\widehat{H}_1(X) = 1.402$ and $\widehat{H}_2(Y) = 99.552$, similar to the true entropies $H_1(X) = 1.428$ and $H_2(Y) = 91.420$, respectively. We formally propose the estimation procedure in Algorithm 1.

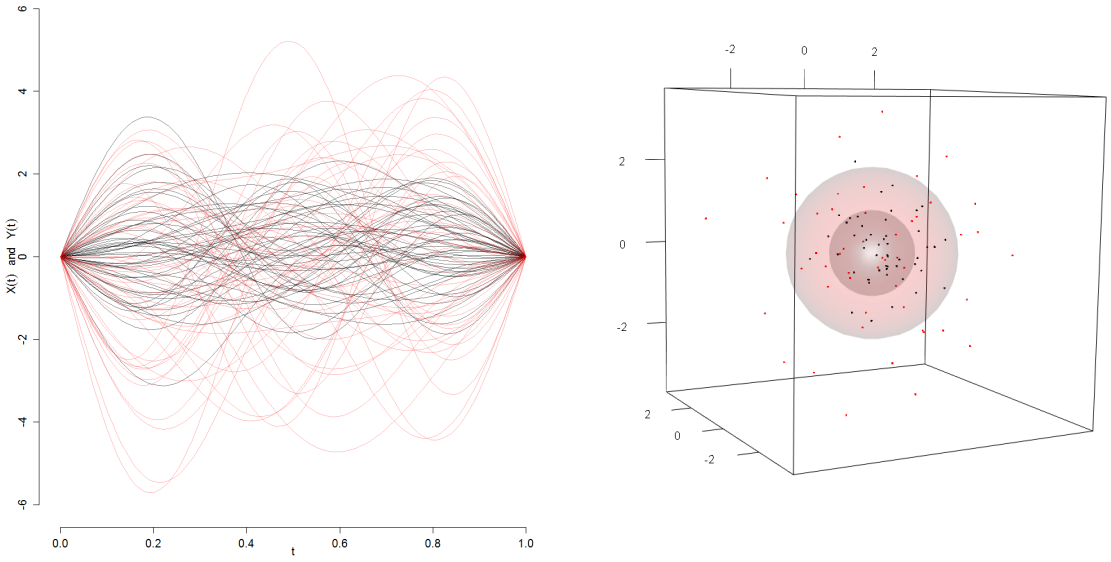


**Figure 1.** Gaussian processes realizations on the left and coefficients for entropy estimation on the right. The sizes of the balls on the right are proportional to the determinants of $\widehat{\Sigma}_X$ (in black) and $\widehat{\Sigma}_Y$ (in red).

**Algorithm 1:** Estimation of $H_\alpha(X,K)$ from a sample of random paths.

1 Functional $K$-entropy: $(\mathbf{X}, K, \alpha, \gamma, d, density)$;
**Input :** The raw-data matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ (paths in rows), the kernel function $K$, the entropy parameter $\alpha$, the regularization parameter $\gamma$, the truncation parameter $d \leqslant \text{rank}(\mathbf{K})$ and a predefined density estimation procedure.
**Output:** $\widehat{H}_\alpha(X,K)$
2 **for** *l in 1 to n* **do**
3 compute $\boldsymbol{a}_l = (\gamma m\mathbf{I} + \mathbf{K})^{-1}\mathbf{y}_l$;
4 **for** *j in 1 to d* **do**
5 $z_{l,j} = \sqrt{\lambda_j}\sum_{i=1}^{m} a_{l,i}v_{i,j}$
6 **end**
7 store $\mathbf{z}_l = (z_{1,l},\ldots,z_{d,l})$
8 **end**
9 Consider $S_n = (\mathbf{z}_1,\ldots,\mathbf{z}_n)$ an *iid* sample from the random vector $Z$. Estimate $\widehat{F}_Z$ with a predefined density estimation procedure and compute $\widehat{V}_\alpha(Z) = \mathbb{E}_{\widehat{F}}\{f^{\alpha-1}\}$;
10 Return $\widehat{H}_\alpha(X,K) = \widehat{H}_\alpha(Z)$.

The choice of kernel parameters in Algorithm 1 is made by cross-validation. This ensures that the curve fitting method is asymptotically optimal. Nonetheless, although the selection of the kernel parameters affects the scale of the estimated entropy, the center-outward ordering induced by $H_\alpha(X,K)$, as formally proposed in the next section, is unaffected. In the Supplementary Material, we present relevant experimental results to illustrate this property, which make the method robust in terms of the selection of the kernel and regularization parameters.

## 3. Minimum Entropy for Anomaly Detection

Anomaly detection is a common task in almost all data analysis context. The unsupervised approach considers a sample $X_1,\ldots,X_n$ of random elements where most instances follow a well-defined pattern and a small proportion, here denoted as $\nu \in [0,1]$, present an abnormal pattern. In recent works (see for instance [10–13]), the authors propose depth measures and related methods, to deal with functional outliers. In this section, we propose a novel criterion to tackle the problem of anomaly detection with functional data using the ideas and concepts developed in Section 2. For a real-valued $d$-dimensional random vector $Z$ that admits a continuous density function $f_Z$, define $H_\alpha(A_Z) = \frac{1}{1-\alpha}\log\left(\int_A f_Z^\alpha(\mathbf{z})d\mathbf{z}\right)$ to be the entropy of the Borel-set $A$ with respect to the measure $F_Z$. Then, the $\nu$-Minimal-Entropy Set (MES) is formally defined as:

$$\text{MES}_\nu(Z) := \{\arg\min_{A\subset\mathbb{R}^d} H_\alpha(A_Z) \text{ s.t. } P(A) \geqslant 1-\nu\}.$$

The $\text{MES}_\nu$ is equivalent [14,15] to a $\nu$-High Density Set (HDS) [16] formally defined as $\text{HDS}_\nu(Z) = \{\mathbf{z}\in\mathbb{R}^d | f_Z(\mathbf{z}) > c_\nu\}$, where $c_\nu$ is the largest constant such that $P(\text{HDS}_\nu(Z)) \geqslant 1-\nu$, for $0 < \nu < 1$. Therefore, the complement of MES is a suitable set to define outlier data in the sample, considering $\tilde{x}(t) \notin \text{MES}_\nu$ as an atypical realization of $X$. Next, we give two approaches to estimate MES.

### 3.1. Parametric Approach

Given a random sample of $n$ discrete random paths $\{x_1(t_i),\ldots,x_n(t_i)\}$ for $i = 1,\ldots,m$, we transform this sample into $d$-dimensional vectors $S_n = (\mathbf{z}_1,\ldots,\mathbf{z}_n)$ using the representation and truncation method proposed in this work, numerically implemented in Lines 2–8 in Algorithm 1. Assume further that $f_Z(\mathbf{z},\boldsymbol{\theta})$ is a suitable probability model for the random sample $\mathbf{z}_1,\ldots,\mathbf{z}_n$, then we estimate by Robust Maximum Likelihood (RML) the parameters $\boldsymbol{\theta}$. For instance, in this paper, we

consider $f_Z(\mathbf{z}, \boldsymbol{\theta})$ to be the normal density, and then, RML estimated parameters are $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}, \widehat{\Sigma})$, the robust mean vector and covariance matrix, respectively. For details on robust estimation, we refer to [17]. After the estimation of the distribution parameters, the computation of $H_\alpha$ follows by plugging the estimated density $f_Z(\mathbf{z}, \widehat{\boldsymbol{\theta}})$ into Equation (1). Moreover, for the normal model, the estimated set $\text{MES}_\nu$ is defined trough the following expression:

$$\text{MES}_\nu(S_n) = \{\mathbf{z} \in \mathbb{R}^d | (\mathbf{z} - \widehat{\boldsymbol{\mu}})^T \widehat{\Sigma}^{-1} (\mathbf{z} - \widehat{\boldsymbol{\mu}}) \leqslant \chi^2_d(\nu)\},$$

where $\chi^2_d(\nu)$ is the $1-\nu$ quantile of a Chi-square distribution with $d$-degrees of freedom. Then, if the coefficient $\mathbf{z}_i$, representing $\tilde{x}_i(t)$, lies outside this ellipsoid, we say that the functional datum is atypical. When the proportion of outlier $\nu$ in the sample is known a priori, the $\chi^2_d(\nu)$-quantile can be replaced by the corresponding sample $1-\nu$ Mahalanobis distance quantile, as is the case in Section 4.1.

### 3.2. Non-Parametric Approach

The following are definitions to introduce further non-parametric estimation methods. For the random vector $Z \in \mathbb{R}^d$ distributed according to $F_Z$, let $B_Z(\mathbf{z}, r_\delta) \subset \mathbb{R}^d$ be the $\mathbf{z}$-centered ball with radius $r_\delta$ that fulfills the condition $\delta = \int_{B_Z(\mathbf{z}, r_\delta)} f_Z(\mathbf{z}) d\mathbf{z}$, then the $\delta$-neighbors of the point $\mathbf{z}$ comprise the open set $\Delta_\mathbf{z} = \mathbb{R}^d \bigcap B(\mathbf{z}, r_\delta)$.

**Definition 3** ($\delta$-local $\alpha$-entropy)**.** *Let $\mathbf{z} \in \mathbb{R}^d$, for $\alpha > 0$ and $\alpha \neq 1$; the $\delta$-local $\alpha$-entropy of the r.v. $Z$ is:*

$$h_\alpha(\Delta_\mathbf{z}) = \frac{1}{1-\alpha} \log \left( \int_{\Delta_\mathbf{z}} f_Z^\alpha(\mathbf{z}) d\mathbf{z} \right) \text{ for all } \mathbf{z} \in \mathbb{R}^d.$$

Under mild regularity conditions on $f_Z$, the local entropy measure is a suitable metric to characterize the degree of abnormality of every point $\mathbf{z}$ in the support of $F_Z$. Several natural estimators of local entropy measures can be considered, for instance the (average) distance from the point $\mathbf{z}$ to its $k$-th-nearest neighbor. We estimate MES combining the estimated $\delta$-Local $\alpha$-entropy. As in the parametric case, let $\{x_1(t_i), \ldots, x_n(t_i)\}$ for $i = 1, \ldots, m$ be a random sample of $n$ discrete random paths; we transform this sample into $d$-dimensional vectors $S_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ following Lines 2–8 in Algorithm 1. Next, we estimate the local entropy for these data using the estimator $\widehat{h}_\alpha(\Delta_{\mathbf{z}_i}) = \exp(\bar{d}_k(\mathbf{z}_i, S_n))$, where $\bar{d}_k(\mathbf{z}_i, S_n)$ is the average distance from $\mathbf{z}_i$ to its $k$-th-nearest neighbor [18], and then estimate $\text{MES}_\nu$ solving the following optimization problem:

$$\max_{\rho, \epsilon_1, \ldots, \epsilon_n} (1-\nu)\rho - \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \quad \text{s.t. } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) \geqslant \rho - \epsilon_i,\ \epsilon_i \geqslant 0 \text{ for } i = 1, \ldots, n. \tag{8}$$

The solution to this problem, $\rho^*$, leads to the following decision function:

$$D(\mathbf{z}) = \text{sign}(\rho^* - \hat{h}_\alpha(\Delta_\mathbf{z})),$$

where $D(\mathbf{z}) = +1$ if $\mathbf{z}$ corresponds to the $(1-\nu)$ proportion of curves projected near the origin, that is the set of curves that belongs to a low entropy (high density) set. The following theorem shows that as the number of available curves increases, the estimation method asymptotically detects the proportion $1-\nu$ of curves belonging to the $\text{MES}_\nu$.

**Theorem 1.** *At the solution of the optimization problem stated in Equation 8, the following equality holds:*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{z}_i) = 1 - \nu,$$

*where $I(\mathbf{z}) = 1$ if $\hat{h}_\alpha(\Delta_\mathbf{z}) \leqslant \rho^*$ and $I(\mathbf{z}) = 0$ otherwise.*

## 4. Experimental Section

The aim of this section is to illustrate the performance of the proposed methodology to detect abnormal observations in a sample of functional data. In what follows, for the representation of functional data, we consider the Gaussian kernel function $K(t_l, t_k) = e^{-\sigma\|t_l - t_k\|^2}$. The kernel parameter $\sigma$ and the regularization coefficient $\gamma$ in Algorithm 1 were defined through cross-validation.

### *4.1. Simulation Analysis*

In a Monte Carlo study, we investigate the performance of the proposed method over three data configurations (Scenarios A, B and C). Specifically, we consider the following generating processes: a fraction $1 - \nu$ of $n = 400$ curves are realizations of the following stochastic model:

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

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_4)$ is a normally-distributed multivariate random variable with mean $\boldsymbol{\mu}_{\xi} = (4, 2, 4, 1)$ and diagonal co-variance matrix $\Sigma_{\xi} = \text{diag}(5, 2, 2, 1)$, and $\varepsilon_l(t)$ are independent autocorrelated random error functions.

The remaining proportion of data $n\nu$ with $\nu \in \{1\%, 5\%, 10\%\}$ comprises outliers that contaminate the sample according to the following typical scenarios (see [19]):

(A) Magnitude outliers: $Y_l(t) = \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \varepsilon_l(t)$, for $l = 1, \ldots, \nu n$, and $t \in [0,1]$, where $\boldsymbol{\zeta}$ is a normally-distributed multivariate r.v. with parameters $\boldsymbol{\mu}_{\zeta} = 2.5\boldsymbol{\mu}_{\xi}$ and $\Sigma_{\zeta} = (2.5)^2\Sigma_{\xi}$.

(B) Shape outliers: $Y_l(t) = \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \varepsilon_l(t)$, for $l = 1, \ldots, \nu n$, and $t \in [0,1]$, where $\boldsymbol{\zeta}$ is a normally-distributed multivariate r.v. with parameters $\boldsymbol{\mu}_{\zeta} = (4, -2, 1, 3)$ and $\Sigma_{\zeta} = \Sigma_{\xi}$.

(C) A combination considering $\nu n/2$ outliers from Scenario A and $\nu n/2$ outliers from Scenario B.

To illustrate the generating process, in Figure 2, we show one instance of the simulated paths in Scenario C with $\nu = 10\%$. We test our Parametric entropy (PA) and Non-Parametric entropy (NPA) method against several well-known depth measures for functional anomaly detection, namely: the Modified Band Depth (MBD), the H-Mode Depth (HMD), the Random Tukey Depth (RTD) and the Functional Spatial Depth (FSD) (see [10–13]), respectively, already implemented in the R-package `fda-usc` [20]. For this experiment, the values of the parameter $\nu$ are assumed known in each scenario. With respect to parameters $\sigma$ and $\gamma$ in Algorithm 1, in this simulation exercise, we chose them with a 10-fold cross-validation procedure using a single set of data, which correspond to the first instance of the simulations. The reference values (which remain fixed throughout the simulation exercise) are $\sigma = 10$ and $\gamma = 0.1^5$.

Let P and N be the amount of outlier and normal data in the sample, respectively, and let TP = True Positive and TN = True Negative be the respective quantities detected by different methods; in Table 1, we report the following average metrics TPR = TP/P (True Positive Rate or sensitivity), TNR = TN/N (True Negative Rate or specificity) and the area under the ROC curve (aROC) of each method obtained through the $M = 1000$ replications in the Monte Carlo study.

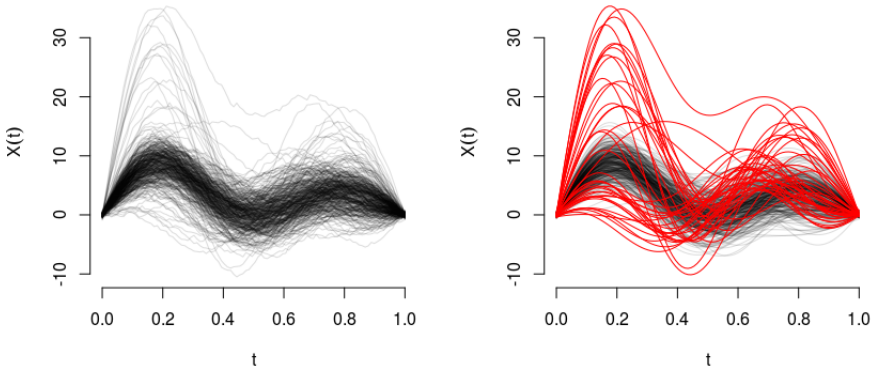


**Figure 2.** (Left) Raw data, 400 curves corresponding to Scenario C with $\nu = 10\%$. (Right) Functional data, in black ("—"), the sample of regular paths $X(t)$, and abnormal curves $Y(t)$ in red ("—").

**Table 1.** Simulation analysis: Scenarios and contamination percentages $\nu$ in columns. In rows, different methods and average sensitivities, specificities and the areas under the ROC curves (aROC) (this last on a scale of $10^2$). The corresponding standard-error is reported in parenthesis.

| Method | Metric | Scenario A | | | Scenario B | | | Scenario C | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 10% | 5% | 1% | 10% | 5% | 1% | 10% | 5% | 1% |
| MBD | TPR | 74.867 (4.699) | 71.010 (7.712) | 55.300 (20.852) | 48.275 (5.914) | 39.395 (9.013) | 13.475 (16.180) | 67.787 (5.351) | 58.365 (7.772) | 36.300 (18.341) |
| | TNR | 97.207 (0.522) | 98.474 (0.406) | 99.548 (0.210) | 94.252 (0.657) | 96.810 (0.474) | 99.126 (0.163) | 96.420 (0.594) | 97.808 (0.409) | 99.356 (0.185) |
| | aROC | 96.662 (1.245) | 97.375 (1.517) | 97.735 (3.059) | 89.393 (2.033) | 91.693 (2.388) | 93.244 (4.425) | 95.272 (1.399) | 95.444 (1.831) | 95.354 (4.370) |
| HMD | TPR | 92.665 (3.295) | 91.545 (5.173) | 88.675 (14.793) | 66.532 (6.084) | 62.780 (8.809) | 47.475 (21.206) | 79.992 (4.562) | 76.765 (7.039) | 66.025 (18.004) |
| | TNR | 99.185 (0.366) | 99.555 (0.272) | 99.885 (0.149) | 96.281 (0.676) | 98.041 (0.463) | 99.469 (0.214) | 97.776 (0.506) | 98.777 (0.370) | 99.656 (0.181) |
| | aROC | 99.200 (0.851) | 99.256 (1.105) | 99.346 (2.391) | 94.980 (1.583) | 96.153 (1.812) | 96.969 (3.473) | 97.676 (1.089) | 97.924 (1.401) | 97.842 (3.542) |
| RTD | TPR | 83.555 (4.743) | 83.045 (0.694) | 76.400 (18.931) | 50.972 (9.409) | 43.940 (1.279) | 22.700 (2.1334) | 71.975 (7.178) | 65.225 (9.716) | 49.700 (1.834) |
| | TNR | 98.174 (0.526) | 99.104 (0.365) | 99.762 (0.191) | 94.544 (1.045) | 97.049 (0.674) | 99.218 (0.215) | 96.889 (0.798) | 98.165 (0.511) | 99.491 (0.184) |
| | aROC | 98.187 (1.094) | 98.605 (1.347) | 98.962 (2.538) | 90.426 (2.817) | 92.510 (2.967) | 94.154 (4.574) | 96.156 (1.580) | 96.345 (1.977) | 96.242 (4.085) |
| FSD | TPR | 81.472 (3.978) | 83.215 (5.947) | 81.925 (16.671) | 50.275 (5.238) | 46.550 (8.018) | 27.400 (19.547) | 74.775 (4.601) | 69.485 (6.859) | 53.775 (16.707) |
| | TNR | 97.941 (0.442) | 99.116 (0.313) | 99.817 (0.168) | 94.475 (0.582) | 97.186 (0.421) | 99.267 (0.197) | 97.197 (0.511) | 98.396 (0.361) | 99.533 (0.168) |
| | aROC | 97.934 (1.030) | 98.738 (1.232) | 99.163 (2.490) | 90.059 (1.794) | 93.279 (2.061) | 95.485 (3.723) | 96.777 (1.158) | 97.148 (1.477) | 97.125 (3.682) |
| Entropy-PA | TPR | **94.150** (3.078) | **93.215** (4.817) | **91.725** (12.591) | **80.740** (6.250) | **77.390** (8.550) | 66.925 (20.330) | **87.550** (4.632) | 84.935 (6.604) | 77.650 (17.015) |
| | TNR | **99.350** (0.342) | **99.649** (0.253) | **99.916** (0.127) | **97.860** (0.694) | **98.810** (0.450) | 99.664 (0.205) | **98.616** (0.514) | 99.207 (0.347) | 99.774 (0.171) |
| | aROC | **99.351** (0.788) | **99.353** (1.078) | **99.374** (2.474) | **97.549** (1.364) | 97.987 (1.495) | 98.301 (2.785) | 98.677 (0.944) | 98.752 (1.208) | 98.641 (3.081) |
| Entropy-NPA | TPR | 92.725 (3.325) | 91.505 (5.228) | 89.050 (14.630) | 74.215 (6.237) | 77.145 (7.904) | **71.250** (19.970) | 87.225 (4.217) | **85.805** (6.198) | **79.775** (16.788) |
| | TNR | 99.191 (0.369) | 99.552 (0.275) | 99.889 (0.147) | 97.135 (0.693) | 98.792 (0.416) | **99.709** (0.201) | 98.586 (0.468) | **99.252** (0.326) | **99.795** (0.169) |
| | aROC | 99.243 (0.815) | 99.266 (1.097) | 99.293 (2.528) | 97.240 (1.130) | **98.253** (1.250) | **98.685** (2.550) | **98.782** (0.856) | **98.880** (1.145) | **98.861** (2.880) |

As can be seen, the PA and NPA entropy methods proposed in this article outperform other recently-proposed depth measures in the three scenarios considered in the experiments when

$\nu = \{0.10, 0.05\}$. In the remaining case (when $\nu = 0.01$), PA and NPA outperform the other methods; however, the standard errors are slightly high to confirm a significant difference between the methods.

When we compare among the proposed methods, the parametric approach seems to be slightly (but consistently) more effective than the non-parametric approach in Scenario A. For Scenarios B and C, both methods provide similar results. It is important to remark that the PA method is especially adequate for Gaussian data, while the NPA method does not assume any distributional hypothesis on the data. In this sense, the simulation results show the robustness of the non-parametric approach even when competing with parametric methods designed for specific distributions.

### *4.2. Outliers in the Context of Mortality-Rate Curve Analysis*

We consider the French mortality rates database, available in the R-package `Demography` [21], to study age-specific male death rates in a logarithmic scale. In Figure 3 (left), each curve corresponds to one year from 1901–2006 (106 paths in total) and accounts for the number of deaths per 1000 of the mean population in the age group (from 0–101 years) in question. As expected, 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.

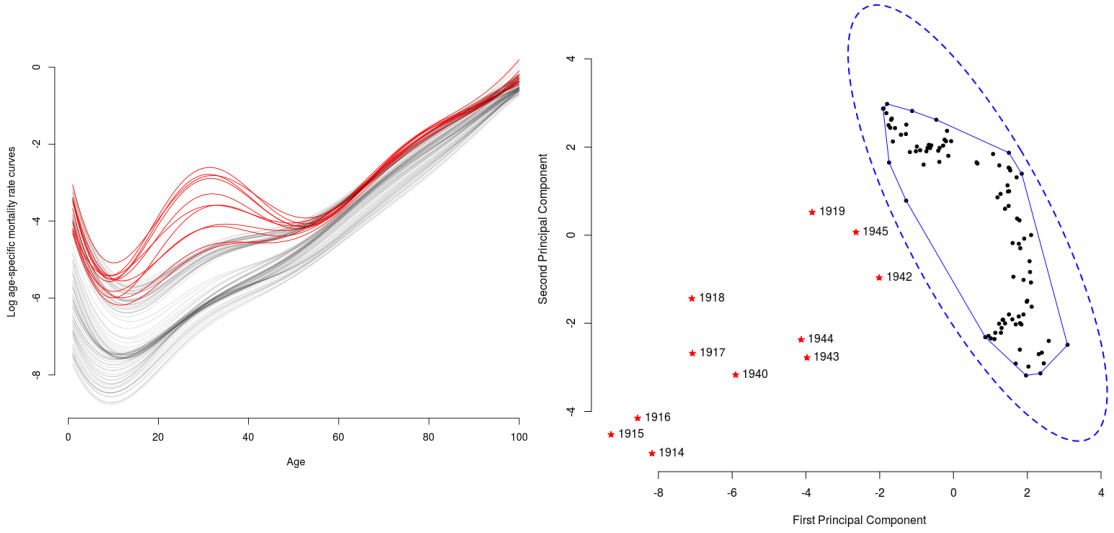


**Figure 3.** French mortality data: On the left, the regular curves in black ("—") and outliers detected in red ("—") for $\nu = 10\%$. On the right, the first two principal components of the kernel eigenfunctions; the area inside the doted blue ellipsoid (- -) corresponds PA estimation of $\text{MES}_{\nu=90\%}$ and the region inside the convex hull in blue (—) to the NPA estimation. The regular curves, represented with black dots (•), lie inside the $\text{MES}_{\nu=90\%}$ and detected outliers with a red asterisk (⁎) outside of $\text{MES}_{\nu=90\%}$.

For some years, the evolution pattern of mortality presents an atypical behavior, mostly coinciding with the first and second World Wars, jointly with the influenza pandemic episode that took place in 1919.

In this experiment, we do not know a priori the proportion of atypical curves. Therefore, after having conducted inference over a wide range of values for $\nu$, as a way to assess the sensitivity and reliability of the inference when determining the number of abnormal curves, we decided to fix $\nu = 10\%$. For further details on the way to choose the parameter $\nu$ (and an extended sensitivity analysis on the values of $\nu$), please refer to § 3.2 in the Supplementary Material. In Figure 3 (left), we highlight in red the anomalous detected curves with both the entropy-PA and NPA methods corresponding to the years 1914–1919 and 1940, 1942–1945, which match with men (between 20 and 40 years old) participating in World War I and II. In Figure 3 (right), we use the first two principal components of the kernel eigenfunctions to project the representation coefficients (in this experiment, in $\mathbb{R}^{14}$) in two

dimensions. As can be seen, the points laying outside the $\text{MES}_{\nu=90\%}$, represented with doted-blue ellipses when estimating it with PA (- -) and the convex hull with a continuous blue line (—) when estimating it with NPA, correspond to the the atypical curves in the sample.

## 5. Discussion

In this article, we propose a definition of entropy for stochastic processes. We provide a reproducing kernel Hilbert space model to estimate entropy from a random sample of realizations of a stochastic process, namely functional data, and introduce two approaches to estimate minimum entropy sets for functional anomaly detection.

In the experimental section, the Monte Carlo simulation illustrates the adequacy of the proposed method in the context of magnitude and shape outliers, outperforming other state of the art methods for functional anomaly detection. In the study of French mortality rates, the parametric and non-parametric approaches for minimum entropy sets estimation show their adequacy to capture anomalous curves, principally associated with the First and Second World Wars and the Influenza episode in 1919.

Regardless of the results presented in the paper, how widely the method can be used in practice, especially with noisier data, is an open question. In this sense, as future work, we will consider testing the performance of the proposed method in other scenarios with different noise assumptions in the observations. Another natural extension for future work entails the study of the asymptotic properties of the $\text{MES}_\nu$ estimators. The extension of the proposed method from the stochastic process to random fields, useful for several statistical and information science areas, seems straightforward, but a wide range of simulations and numerical experiments must be done in order to stress the performance of entropy methods in comparison to other techniques when dealing with abnormal fields. Another natural avenue for future work entails the study of the connections between entropy for stochastic process, as formally defined here, and the maximum entropy principle when estimating the governing parameters of Gaussian processes.

**Supplementary Materials:** The following are available online at www.mdpi.com/1099-4300/20/1/33/s1.

**Acknowledgments:** We thank the referees and the editor for constructive comments and insightful recommendations. This work has been supported by CONICET Argentina Project 20020150200110BA, the Spanish Ministry of Economy and Competitiveness Projects ECO2015–66593-P, GROMA(MTM2015-63710-P), PPI (RTC-2015-3580-7) and UNIKO(RTC-2015-3521-7) and the "methaodos.org" research group at URJC.

**Author Contributions:** All authors have contributed equally to the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| RML | Robust Maximum Likelihood. |
| MES and HDS | Minimum Entropy and High Density Sets, respectively. |
| PA and NPA | Parametric and Non-Parametric approaches. |
| MBD, HMD, RTD, FSD | Modified Band, H-Mode, Random Tukey and Functional Spatial Depths. |

## Appendix A

**Proof Theorem 1.** Consider the following optimization problem:

$$\min_{\beta_1,\ldots,\beta_n} \sum_{i=1}^{n} \beta_i \widehat{h}_\alpha(\Delta_{\mathbf{z}_i}) \quad \text{s.t.} \quad \sum_{i=1}^{n} \beta_i = n(1-\nu) \text{ and } 0 \leqslant \beta_i \leqslant 1 \text{ for } i = 1,\ldots,n. \tag{A1}$$

For the sake of simplicity, consider first the case where $n(1-\nu) \in \mathbb{N}$. Let $q^*$ be the $1-\nu$ quantile of the $S_n$ sample. Then, it can be shown that $\beta_i^* = 1$ if $\hat{h}_\alpha(\Delta_{\mathbf{z}_i}) \leqslant q^*$ and $\beta_i^* = 0$ if $\hat{h}_\alpha(\Delta_{\mathbf{z}_i}) > q^*$ is a solution for the problem stated in Equation (A1). As a consequence:

$$\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{z}_i) = \frac{1}{n}\sum_{i=1}^{n} \beta_i^*.$$

From the constraint in Equation (A1), it holds that $\sum_{i=1}^{n} \beta_i^* = n(1-\nu)$, and then:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \beta_i^* = \lim_{n\to\infty} \frac{1}{n} n(1-\nu) = 1-\nu$$

For the case $n(1-\nu) \notin \mathbb{N}$, it holds that $\begin{cases} \beta_i = 1, & \text{if } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) < q^* \\ \beta_i = n(1-\nu) - [n(1-\nu)], & \text{if } \hat{h}_\alpha(\Delta_{\mathbf{z}}) = q^* \\ \beta_i = 0, & \text{if } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) > q^* \end{cases}$

where $[z]$ stands for the largest integer no greater than $x$. Therefore, the number of $\beta_i^*$'s equating to one is $[n(1-\nu)]$ and:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{z}_i) = \lim_{n\to\infty} \frac{1}{n}([n(1-\nu)] \times 1 + 1) = \lim_{n\to\infty} \frac{[n(1-\nu)]}{n} = 1-\nu.$$

Finally, we show that $\rho^* = q^*$. The dual problem of (A1) is:

$$\max_{b,\epsilon_1,\ldots,\epsilon_n} n(1-\nu)b - \sum_{i=1}^{n} \epsilon_i \quad \text{s.t. } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) \geqslant b - \epsilon_i,\ \epsilon_i \geqslant 0 \text{ for } i = 1,\ldots,n. \tag{A2}$$

By the fundamental theorem of duality, the objective functions of the problems stated in Equations (A1) and (A2) take the same value at their solutions, and as a consequence, $b^* = q^*$ (see [22]). Since Problem (A2) differs from Problem (8) just in the scaling of the objective function, it holds that $\rho^* = b^*$, which concludes the proof. □

## References


1. Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
2. Bosq, D. *Linear Processes in Function Spaces: Theory and Applications*; Springer Science & Business Media: New York, NY, USA, 2012.
3. Ramsay, J.O. *Functional Data Analysis*; Wiley: New York, NY, USA, 2006.
4. Ferraty, F.; Vieu, P. *Nonparametric Functional Data Analysis: Theory and Practice*; Springer: New York, NY, USA, 2006.
5. Berlinet, A.; Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*; Springer: New York, NY, USA, 2011.
6. Kimeldorf, G.; Wahba, G. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **1971**, *33*, 82–94.
7. Cucker, F.; Smale, S. On the mathematical foundations of learning. *Bull. Am. Math. Soc.* **2002**, *39*, 1–49.
8. Muñoz, A.; González, J. Representing functional data using support vector machines. *Pattern Recognit. Lett.* **2010**, *31*, 511–516.
9. Zhu, H.; Williams, C.; Rohwer, R.; Morciniec, M. *Gaussian Regression and Optimal Finite Dimensional Linear Models*; Aston University: Birmingham, UK, 1997.
10. López-Pintado, S.; Romo, J. On the concept of depth for functional data. *J. Am. Stat. Assoc.* **2009**, *104*, 718–734.
11. Cuevas, A.; Febrero, M.; Fraiman, R. Robust estimation and classification for functional data via projection-based depth notions. *Comput. Stat.* **2007**, *22*, 481–496.

12. Sguera, C.; Galeano, P.; Lillo, R. Spatial depth-based classification for functional data. *Test* **2014**, *23*, 725–750.
13. Cuesta-Albertos, J.A.; Nieto-Reyes, A. The random Tukey depth. *Comput. Stat. Data Anal.* **2008**, *52*, 4979–4988.
14. Hero, A. Geometric entropy minimization (GEM) for anomaly detection and localization. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007; pp. 585–592.
15. Xie, T.; Narabadi, N.; Hero, A.O. Robust training on approximated minimal-entropy set. *arXiv* **2016**, arXiv:1610.06806.
16. Hyndman, R.J. Computing and graphing highest density regions. *Am. Stat.* **1996**, *50*, 120–126.
17. Maronna, R.; Martin, R.; Yohai, V. *Robust Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
18. Beirlant, J.; Dudewicz, E.; Györfi, L.; Van der Meulen, E. Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **1997**, *6*, 17–39.
19. Cano, J.; Moguerza, J.M.; Psarakis, S.; Yannacopoulos, A.N. Using statistical shape theory for the monitoring of nonlinear profiles. Applied Stochastic Models in Business and Industry. *Appl. Stoch. Models Bus. Ind.* **2015**, *31*, 160–177.
20. Febrero-Bande, M.; De la Fuente, M.O. Statistical computing in functional data analysis: The R package fda.usc. *J. Stat. Softw.* **2012**, *51*, 1–28.
21. Hyndman, R.J. *Demography Package*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
22. Muñoz, A.; Moguerza, J.M. Estimation of high-density regions using one-class neighbor machines. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 476–480.