entropy

MDPI

Article

Combining Entropy Measures for Anomaly Detection

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

[1] Department of Statistics, Universidad Carlos III de Madrid, 28903 Getafe, Madrid, Spain
[2] Department of Computer Science and Statistics, University Rey Juan Carlos, 28933 Móstoles, Madrid, Spain; javier.moguerza@urjc.es
[3] Department of Mathematics and Statistics, Universidad Torcuato Di Tella and CONICET, Buenos Aires C1428BCW, Argentina; gmartos@utdt.edu
* Correspondence: albmun@est-econ.uc3m.es (A.M.); nihernan@est-econ.uc3m.es (N.H.); Tel.: +34-91-624-9579 (A.M.)

Received: 31 July 2018; Accepted: 10 September 2018; Published: 12 September 2018

Abstract: The combination of different sources of information is a problem that arises in several situations, for instance, when data are analysed using different similarity measures. Often, each source of information is given as a similarity, distance, or a kernel matrix. In this paper, we propose a new class of methods which consists of producing, for anomaly detection purposes, a single Mercer kernel (that acts as a similarity measure) from a set of local entropy kernels and, at the same time, avoids the task of model selection. This kernel is used to build an embedding of data in a variety that will allow the use of a (modified) one-class Support Vector Machine to detect outliers. We study several information combination schemes and their limiting behaviour when the data sample size increases within an Information Geometry context. In particular, we study the variety of the given positive definite kernel matrices to obtain the desired kernel combination as belonging to that variety. The proposed methodology has been evaluated on several real and artificial problems.

Keywords: entropy kernel; kernel combination; Karcher mean; anomaly detection; functional data

1. Introduction

Usual Data Mining tasks, such as classification, regression and anomaly detection, are heavily dependent on the geometry of the underlying data space. Kernel Methods, such as Support Vector Machines (SVM), provide the control on the data space geometry through the use of a Mercer kernel function [1,2]. Such functions, defined in the next section, induce embeddings of the data in feature spaces where Mercer kernels act as inner products. The choice of the appropriate kernel, including its parameters, is a particular case of model selection problems.

For instance, when working with SVM, a delicate parameterization is needed; otherwise, solutions might be suboptimal. In other words, the choice of a suitable kernel function and its parameters will affect both the geometry of the data embedding and the success of the algorithms [3,4]. A typical way to proceed is by means of cross-validation procedures [5]. However, these parameter calibration strategies, although intuitive and simple from an applied point of view, have some important drawbacks. In particular, their computational burden is of practical relevance when implementing cross–validation strategies in problems that involve calibrating a medium to large amount of parameters. An appealing alternative to model selection when working with SVM is to combine or merge different kernel functions into a single kernel [6,7].

Functional data [8] present the particularity of being intrinsically infinite dimensional. This peculiarity implies that classical procedures for multivariate data must be adapted or redesigned to cope with functional data. The statistical distribution of data is a basic element to afford outlier detection problems. Entropies are natural functions to use in anomaly detection problems given that

any definition of entropy should produce large values for scattered distributions and small values for concentrated distributions. In addition, statistical distributions are a particular case of functional data and in this way entropy comes then into play in this context.

In this paper, we present an alternative proposal to solving anomaly detection problems that avoids the selection of kernel hyperparameters. A novelty of this work is that the methodology is developed to deal with functional data. We will explore several kernel combination techniques, including some methods from Information Geometry that respect the geometry of the manifold that contains the Gram matrices associated with the Mercer kernels involved.

The paper is organized as follows: Section 2 describes the functional data analysis methods used to produce the data representations from kernels, as well as the minimum entropy method used in this paper for anomaly detection. Section 3 develops several methods to obtain kernel combinations for the task of outlier detection. Section 4 illustrates the theory with simulations and examples; and Section 5 concludes the work.

2. Reproducing Kernel Hilbert Spaces for Multivariate and Functional Data

Let X be the "space" where the data live (a compact metric space). A Mercer kernel is a function $K : X \times X \to \mathbb{R}$ symmetric, continuous and such that, for all finite sets $S = \{x_1, \ldots, x_n\} \subset X$, the matrix whose entries are $K(x_i, x_j)_{i,j \in \{1, \cdots, n\}}$ is positive semidefinite. Often, we will use the term "kernel function" when referring to a Mercer kernel. Kernel functions admit expansions of the type $K(x, z) = \sum_i \phi(x)^T \phi(z)$ for some $\phi : X \longrightarrow \mathbb{R}^d$, where d is usually large. In particular, $\phi(X)$ is some manifold embedded in $\mathbb{R}^d$ [9]. For $x \in X$, denote K_x the function $K_x : X \to \mathbb{R}$ given by $K_x(z) = K(x, z)$. There exists a unique Hilbert space H_K of functions on X made up of the span of the set $\{K_x | x \in X\}$, such that for all $f \in H_K$ and $x \in X$, $f(x) = \langle K_x, f \rangle_{H_K}$. The Hilbert space H_K is said to be a Reproducing Kernel Hilbert Space (RKHS) [10]. Next, we describe the use of RKHS for data analysis, differentiating between the multivariate and functional cases.

In the multivariate case, we consider data sets $S = \{x_1, \ldots, x_n\} \subset X$, where X is a compact subset of $\mathbb{R}^D$. Consider the RKHS H_K and the linear integral operator L_K defined by $L_K(f) = \int_X K(\cdot, s) f(s) ds$. Since X is compact and K continuous, L_K has a countable sequence of eigenvalues $\{\lambda_j\}$ and eigenfunctions $\{\phi_j\}$, and K can be expressed as $K(x, y) = \sum_j \lambda_j \phi_j(x) \phi_j(y)$, where the convergence is absolute and uniform (Mercer's theorem).

Consider the Gram matrix $K_S = K(x_i, x_j)$, $i, j \in \{1, \cdots, n\}$. This matrix is real, symmetric and positive definite (by definition of K) and $K_S(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, where $\phi(x_i) = (\sqrt{\lambda_j} \phi_j(x_i))_j$ is the mapping $\phi : X \to \mathbb{R}^d$. Straightforwardly, K is the standard scalar product in $\mathbb{R}^d$. Thus, the use of K induces both a data transformation and a metric on the original data given by:

$$d_K^2(x_i, x_j) = \|\phi(x_i) - \phi(x_j)\|^2 = \phi(x_i)^T \phi(x_i) + \phi(x_j) \phi(x_j) - 2\phi(x_i) \phi(x_i) = K(x_i, x_i) + K(x_j, x_j) - 2K(x_i, x_j) \quad (1)$$

Equation (1) shows that the choice of the kernel K determines the geometry of the data set after the transformation $X \to \phi(X)$.

Now, we consider the functional data case, that is, the case where data are functions or, by generalization, infinite dimensional objects (such as images, for instance). Let $(\Omega, \mathcal{F}, P)$ be a probability space, where $\mathcal{F}$ is the σ-algebra in Ω and P a σ-finite measure. We consider random elements (functions) $X(\omega, t) : \Omega \times T \to \mathbb{R}$ in a metric space (T, τ), where $T \subset \mathbb{R}$ is compact and we assume $X(\omega, \cdot)$ to be continuous functions. We consider kernels $K_X(s, t) = \mathbb{K}(X(\omega, s), X(\omega, t))$ (the classical choice is $K_X(s, t) = \mathbb{E}(X(\omega, s) X(\omega, t))$). Then, there exists a basis $\{e_i\}_{i \geqslant 1}$ of $\mathcal{C}(T)$ such that for all $t \in T$

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

for appropriate coefficients, where the e_i are the eigenfunctions associated with the integral operator of $K_X(s, t)$.

In real data analysis, we do not have theoretical random paths, or functional data described by mathematical equations, but finite samples from such processes. For instance, if we are considering normal distributions as the object of analysis, we will not know the vectors of means and real covariance matrices(μ and Σ), but a sample $X = \{x_i\} \in \mathbb{R}^n$ from which we will estimate the covariance matrix $S = \frac{1}{n} XX^T$. In the case of functions, X will be a compact space or manifold in an Euclidean space, $Y = \mathbb{R}$, and there will be available sample curves f_n identified with data sets $\{(x_i, y_i) \in X \times Y\}_{i=1}^n$. Let $K : X \times X \to \mathbb{R}$ a Mercer Kernel and H_K its associated RKHS. Then, the coefficients in Equation (2) can be approximated by solving the following optimization problem [11]:

$$\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \gamma \|f\|_K^2, \tag{3}$$

where $\gamma > 0$ and $\|f\|_K^2$ represents the norm of the function f in H_K. The solution, that constitutes an example of Equation (2), is given by $f^*(x) = \sum_j \hat{\lambda}_j \phi_j(x)$, where the $\hat{\lambda}_j$ are the weights of the projection of the function corresponding to the sample $\{(x_i, y_i)\}$ onto the function space generated by the eigenvalues of L_K.

Next, we use local entropies for anomaly detection through kernel combinations. For this preliminary work, we explore linear combinations and Karcher means, to validate the intuition that the use of a more natural mean than the arithmetic mean will produce better practical results, as far as positive definite matrices are involved.

2.1. Local Entropy Kernels

In order to link the metric induced by the kernel function and the underlying (empirical) density in the data, we propose *local entropy kernels*. Consider a measurable cover on $(\Omega, \mathcal{F}, P)$—the probability space where the random element of interest X is defined—say $\{\Omega_i\}_{i \geqslant 1}$, where $\bigcup_{i \geqslant 1} \Omega_i = \Omega$ and $\Omega_i \cap \Omega_j = \varnothing$ for any $i \neq j$; we can define the α-Entropy [12] of X as follows:

$$H_\alpha(X) = \frac{1}{1-\alpha} \sum_{i \geqslant 1} P(\Omega_i) \log \left(P(\Omega_i)^{\alpha - 1} \right), \text{ for } \alpha \geqslant 0 \text{ and } \alpha \neq 1. \tag{4}$$

The parameter α defines to which entropy inside the family of α entropies we are referring to. For instance, when $\alpha = 0$, then H_α is the Hartley entropy, when $\alpha \to 1$ then H_α converges to the Shannon entropy and when $\alpha \to \infty$ then H_α converges to the Min-entropy measure. Let $\mathcal{S}_\Omega$ be the collection of finite partitions of Ω, for any subset $A = \bigcup_{i=1}^n A_i \in \mathcal{S}_\Omega$, the entropy of A can be computed as follows:

$$H_\alpha(A) = \frac{1}{1-\alpha} \sum_{i=1}^{n} P(A_i) \log \left(P(A_i)^{\alpha - 1} \right). \tag{5}$$

This paves the way to define the Δ-local entropy [13] corresponding to any subset $\Delta \in \mathcal{F}_\Omega$ as follows

$$h_\alpha(\Delta) = \inf_{\tilde{\Delta} \in \mathcal{S}_\Omega} H_\alpha(\tilde{\Delta}), \text{ such that } \Delta \subset \tilde{\Delta}. \tag{6}$$

Let $(X_1, \ldots, X_n)$ be a random sample drawn *i.i.d.* from P, we would like to compute the local entropies of the corresponding random sets $\Delta_1, \ldots, \Delta_n$, where $\Delta_i = \Omega \bigcap B(X_i^{-1}(\omega), r)$ and $B(X_i^{-1}(\omega), r) \in \Omega$ is the open ball with center in ω and a (data driven) small radius r. In practice, given a sample $S_n = (x_1, \ldots, x_n)$, we compute the local entropy using the estimator $\widehat{h}_\alpha(\Delta_i) = \bar{d}_k(x_i, S_n)/(1-\alpha)$, where $\bar{d}_k(x_i, S_n)$ is the average distance from x_i to its k^{th}-nearest neighbour. Notice that the locality parameter k in $\bar{d}_k(x, S_n)$, which represents the number of neighbours that we take into

account to approximate the local entropy around x, is related to r in $\Delta_x = \Omega \bigcap B(x^{-1}(\omega), r)$. We then consider $\varphi(x) = \widehat{h}_\alpha(\Delta_x)$, with $\alpha = 0$, so to define the local entropy kernel as

$$K(x, y) = \varphi(x)^T \varphi(y). \tag{7}$$

In the next section, we discuss how to avoid model selection problems. To this aim, a set of local entropy kernels is initially estimated from the data. Then, we estimate an average local entropy kernel that takes into account the particular geometry of the space of positive definite matrices. In this way, we obtain a unique low dimensional data representation, from which outliers are detected. This approach does not include neither a model selection step nor a parameter estimation procedure.

3. Kernel Combination for Anomaly Detection

Consider a data sample $S_n = \{x_1, \ldots, x_n\} \subset X$, where the x_i can be multivariate or functional data, and consider a set of m Mercer kernels (or matrices) $K_1^e, \cdots, K_m^e$, that induce m different data embeddings $\phi_j : X \to \mathbb{R}^{d_j}$, where $K_j^e(x, y) = \phi_j(x)^T \phi_j(y)$. As stated in Equation (1), each of the kernels induces a kernel distance d_{K_j} on the original data space X, corresponding to the Euclidean distance on the manifold $Z_j = \phi_j(X)$.

Next, we define a new set of transformations, suitable for anomaly detection, in line with the theory of Section 2.1 by:

$$\varphi_j(x) = d_{K_j}(\phi_j(x), \phi_j(S_n)). \tag{8}$$

The corresponding kernels suitable for outlier detection are

$$K_j(x, y) = \varphi_j(x)^T \varphi_j(y). \tag{9}$$

Now, kernel functions are positive definite type functions, i.e., the empirical kernel matrix K—obtained via the evaluation of the kernel function into the set of n training points—belongs to the cone of symmetric positive semidefinite matrices $\mathcal{P} := \{K \in \mathbb{R}^{n \times n} | K = K^T, K \geq 0\}$. Let $K_1, \ldots, K_m$ be the empirical kernel matrices defined in Equation (9), all of them in $\mathcal{P}$, and let $(w_1, \ldots, w_m)^T$ be a suitable non-negative vector of combination parameters, then define the "fusion" kernel $\mathcal{K}$

$$\mathcal{K}(w_1, \ldots, w_m) := w_1 K_1 +, \ldots, +w_m K_m \geq 0.$$

In the context of SVM classification problems, the goal is to find the parameters $w_1, \ldots, w_m$ that maximize the optimal margin. Instead, in anomaly detection cases, the goal is to estimate the parameters $w_1, \ldots, w_m$ that produce a suitable data representation. This is achieved when the regular data within the sample –represented in the coordinate space provided by the fusion kernel $\mathcal{K}$–have a reduced entropy or equivalently is scarcely scattered and those observations that are atypical in the sample are projected in distant regions from that of the regular data.

Next we consider three particular combination schemes. The first is rather straightforward, the second proposes the mean in the manifold that contains the kernels, and the third is a weighting scheme that assigns the weights according to the use of appropriate choices of entropy functions.

Definition 1 (Multivariate sparsity measures)**.** *Consider m different sparsity measures $\phi_1, \cdots, \phi_m$ and let $K_1, \cdots, K_m$ be the corresponding set of Mercer kernels, where $K_i(x, y) = \phi_i^T(x)\phi_i(y)$. We define a multivariate concentration measure by $\mathbf{\Phi} = (\phi_1, \cdots, \phi_m) : X \to \mathbb{R}^m$.*

The corresponding kernel, evaluated at the sample S, will be

$$K_\Phi(x_i, x_j) = \mathbf{\Phi}(x_i)^T \mathbf{\Phi}(x_j) \quad = \quad (\phi_1(x_i), \cdots, \phi_m(x_i))^T (\phi_1(x_j), \cdots, \phi_m(x_j)) = \sum_{i=1}^m \phi_i^T(x_i)\phi_i(x_j) = \sum_{i=1}^m K_i(x_i, x_j) \tag{10}$$

Thus, the kernel corresponding to a multivariate sparsity measure $\mathbf{\Phi} = (\phi_1, \cdots, \phi_m)$ is the sum of the univariate kernels K_i associated with the ϕ_i. This fact allows us to interpret linear combination of kernels $\sum w_i K_i$ as coming from (weighted) multivariate sparsity measures.

3.1. Entropy Weighting

Definition 2 (K-entropy of a data set)**.** *Consider a Mercer kernel K acting on a space X, a sample data set S_n and the corresponding transformation $\phi : X \to \mathbb{R}^d$ induced by K, where $K(x,y) = \phi(x)^T\phi(y)$. The **K-entropy** of S_n is defined by:*

$$E_K(S_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} |K(x_i,x_j)| = \sum_{i=1}^{n}\sum_{j=1}^{n} |\phi(x_i)^T\phi(x_j)|. \tag{11}$$

In the context of outlier detection, consider $K_1, \cdots, K_m$, obtained from sparsity measures. From Equation (9), if a point x is an outlier, then it will be off the main bulk of data points and, thus, $\varphi_j(x) = d_{K_j}(\phi_j(x), \phi(S_n))$ will be large and the same will be true for $K_j(x, x_i)$ for most $x_i \in S_n$. As a consequence, $E_{K_j}(S_n)$ will tend to be large. On the other hand, and following the same reasoning, if a particular kernel K_j induces a representation not suitable for detecting the outliers, then $E_{K_j}(S_n)$ will be small. Thus, the measure defined in Equation (11) acts as a true entropy for matrices: If data are very concentrated after the transformation induced by K, then the entropy of the data (measured by the) set will be low.

We establish the entropy-weighting scheme by solving the following semidefinite optimization problem:

$$\begin{array}{c} \max\limits_{\alpha_1,\cdots,\lambda_m} \sum_{j=1}^{m} \lambda_j E_{K_j}(S_n) \\ s.t. \sum_{j=1}^{m} \lambda_j K_j \succeq 0, \quad \sum_{j=1}^{m} \lambda_j = 1, \text{ and } 0 \leqslant \lambda_j \leqslant u_j, \end{array} \tag{12}$$

where $u_j \in [(0,1]$ are some positive constants that may be associated with each kernel matrix K_j. We refer to [14] for a detailed description of the basics of semidefinite programming.

Theorem 1. *Consider the previous semidefinite optimization problem. If $K_1, \cdots, K_m \succeq 0$ and $u_i = \dfrac{E_{K_j}(S_n)}{\sum_j E_{K_j}(S_n)}$, then the solution to the optimization problem is given by $\lambda_j^* = \dfrac{E_{K_j}(S_n)}{\sum_j E_{K_j}(S_n)}$.*

Proof. Given that $\lambda_j^* = \dfrac{E_{K_j}(S_n)}{\sum_j E_{K_j}(S_n)} \geqslant 0$ and $K_j \succeq 0$, the constraint $\sum_{j=1}^{m} \lambda_j K_j \succeq 0$ holds. In addition,

$$\sum_{j=1}^{m} \lambda_j^* = \sum_{j=1}^{m} \frac{E_{K_j}(S_n)}{\sum_j E_{K_j}(S_n)} = 1.$$

Since all the λ_j^* reach their upper bound, the theorem holds and the solution is unique. □

Thus, the entropy-weighting scheme will be:

$$K^* = \frac{E_{K_1}(S_n)}{\sum_j E_{K_j}(S_n)} K_1 + \frac{E_{K_2}(S_n)}{\sum_j E_{K_j}(S_n)} K_2 + \cdots + \frac{E_{K_m}(S_n)}{\sum_j E_{K_j}(S_n)} K_m. \tag{13}$$

3.2. Karcher Mean

Next, we introduce the Karcher mean [15–17] of kernel matrices as an alternative approach to the linear combinations of matrices presented in Section 3.1. The Karcher mean preserves the particular Riemannian manifold in which the kernel matrices lie and constitutes a natural definition for the geometric mean of the matrices.

The set of positive definite square matrices $\mathcal{P}$ is a Riemannian manifold, with inner product $\langle A, B\rangle_X = Tr(X^{-1}AX^{-1}B)$ on the tangent space to $\mathcal{P}$ at the point X. The distance between $A, B \in \mathcal{P}$ is given by $d_{\mathcal{P}}(A, B) = \|\log(A^{-1/2}BA^{-1/2})\|_F$, where $\|\cdot\|_F$ is the Frobenius norm, that is $\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$. Given $K_1, \ldots, K_m$ kernel matrices, the Karcher mean, denoted onwards as $\overline{\overline{K}}$, is defined as the minimizer of the function $f(X) = \sum_{i=1}^m d_{\mathcal{P}}(X, K_i)^2$, and it is the unique solution $X \in \mathcal{P}$ of the matrix equation $\sum_{i=1}^m \log(K_i^{-1}X) = 0$.

4. Experimental Section

In this section, we illustrate, with the aid of multiple numerical examples and real data sets, the performance of the proposed methodology when the goal is to detect abnormal observations in a sample. We consider a list of several kernel functions, namely: (i) the Gaussian kernel $K_G(x_i, x_j) = e^{-\sigma\|x_i - x_j\|^2}$ with parameter σ defined in a grid of values ranging in $\sigma \in \{0.1^3, 0.1^2, 0.1, 1, 10, 50, 100, 500, 10^3\}$; (ii) the linear kernel $K_L(x_i, x_j) = \langle x_i, x_j\rangle$ and (iii) the second degree polynomial kernel $K_P(x_i, x_j) = (\langle x_i, x_j\rangle + 1)^2$. As it was explained in Section 1, the combination methods proposed can be considered as an alternative to model selection techniques for outlier detection purposes. Therefore, the results obtained are presented jointly with the single kernel methods. Our combination methods are denoted as: (i) the average kernel ($\bar{K}$), (ii) the kernel constructed using the *Karcher* mean of the single kernel functions ($\overline{\overline{K}}$) and (iii) the minimum entropy linear combination kernel or entropy kernel (E_K).

For comparison purposes, we consider several alternative approaches for anomaly detection in both the multivariate and the functional data frameworks. In the multivariate case, we consider some alternative well-known techniques in the field of machine learning. These methods are: (i) LOF [18] and (ii) HiCS [19]. In the functional case, we test our proposals against three widely used depth measures: the Modified Band Depth (MBD) [20], the Modal Depth (HMD) [21] and the Random Tukey Depth (RTD) [22]. This depth measures induce an order with respect to the functional data set that can be used to determine which observations (curves) are far from the deepest or central point and can be classified as outliers.

Each database presents a set of regular observations and has been contaminated with abnormal or outlying observations. 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 Tables 1 and 2, we report the following average metrics TPR = TP/P (True Positive Rate or sensitivity), TNR = TN/N (True Negative Rate or specificity). For the comparison with other techniques using real data sets, in Tables 4 and 5, we report the area under the ROC curve (AUC) for each experiment.

4.1. Synthetic Data

For the simulated experiment, we consider two synthetic data schemes. The first scheme has been built by generating a synthetic multivariate data set, while, for the second scheme, we have generated a synthetic functional data set.

Synthetic multivariate data: We consider a conditionally normal bivariate distribution model [23] for regular data and outliers were sampled from three different standard Gaussian models. The sample size is $n = 1000$. The data for the experiment, illustrated in Figure 1, was obtained using a Gibbs sampler.

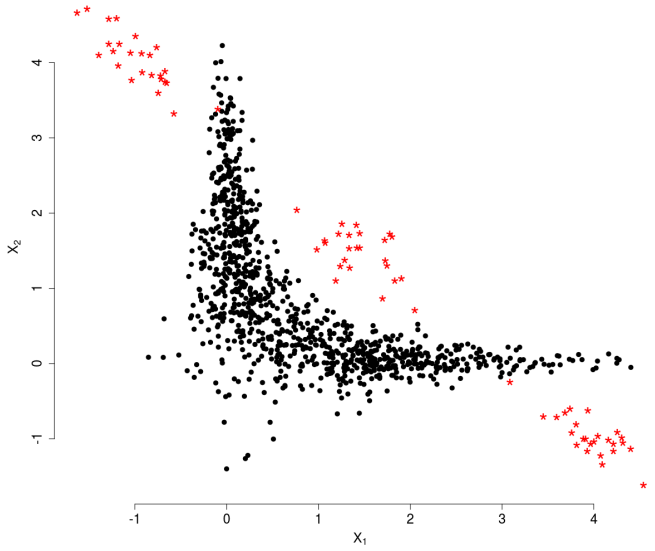

Figure 1. Main data in black (•) and outlying observations in red (*).

Synthetic functional data: We consider random samples of Gaussian processes $\{x_1(t), \ldots, x_n(t)\}$, with sizes 4000 and 2000, where a proportion $\nu = 0.1$, known a priori, present an atypical pattern, and the remaining $n(1-\nu)$ curves are considered the main data. We consider the following generating processes:

$$\begin{aligned} X_l(t) &= \sum_{j=1}^{2} \xi_j \sin(j\pi t) + \varepsilon_l(t), \text{ for } l = 1, \ldots, (1-\nu)n, \\ Y_l(t) &= \sum_{j=1}^{2} \zeta_j \sin(j\pi t) + \varepsilon_l(t), \text{ for } l = 1, \ldots, \nu n, \end{aligned}$$

where $t \in [0,1]$, $\varepsilon(t)$ are independent autocorrelated random error functions and (ξ_1, ξ_2) is a normally-distributed bivariate random variable (NDMRV) with mean $\boldsymbol{\mu}_\xi = (1,2)$ and diagonal co-variance matrix $\boldsymbol{\Sigma}_\xi = \text{diag}(1,1)$. To generate the outliers, we consider (ζ_1, ζ_2) NDMRV with parameters $\boldsymbol{\mu}_\xi = (4,5)$ and $\boldsymbol{\Sigma}_\zeta = \Sigma_\xi$. The data are plotted in Figure 2.

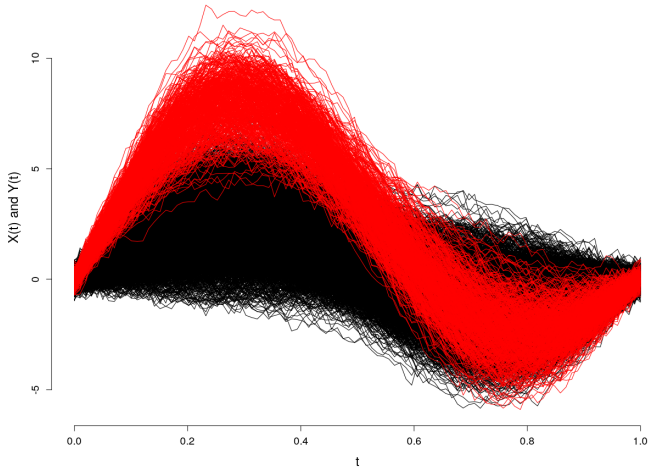

Figure 2. Main data in black (—) and outlying observations in red (—).

Table 1 shows the results of the experiment using synthetic multivariate data. Best results are marked using bold enhanced text. It can be observed that the proposed combination methods, namely the mean, the weighted entropy and the Karcher mean perform as well as the best single kernel in terms of the TNR. With respect to the TPR, the best combination method is the one based on the calculation of the Karcher mean.

Table 1. Percentage of TPR (sensitivity) and TNR (specificity) for synthetic multivariate data.

Experiment	$K_{G_{\sigma=0,1^3}}$	$K_{G_{\sigma=0,1^2}}$	$K_{G_{\sigma=0,1}}$	$K_{G_{\sigma=1}}$	$K_{G_{\sigma=10}}$	$K_{G_{\sigma=50}}$	$K_{G_{\sigma=100}}$	$K_{G_{\sigma=500}}$	$K_{G_{\sigma=10^3}}$	K_L	K_P	$\bar{K}$	$\bar{\bar{K}}$	E_K
TPR	**69.3**	**69.3**	**69.3**	65.3	0.0	0.0	0.0	0.0	0.0	**69.3**	62.6	62.6	**69.3**	62.6
TNR	**97.7**	**97.7**	**97.7**	**97.7**	92.5	92.5	92.5	92.5	92.5	**97.7**	**97.7**	**97.7**	**97.7**	**97.7**

In Table 2, the results of the experiment using synthetic functional data are presented. In this case, two of the three proposals, the mean and the weighted entropy are always able to perform as well as the best single kernel (the polynomial kernel) in terms of both the TNR and TPR. The method based on the calculation of the Karcher mean obtains good results with respect to the TNR measure.

Table 2. Percentage of TPR (sensitivity) and TNR (specificity) for synthetic functional data.

Experiment	Train Set (n = 4000)		Test Set (n = 2000)	
	TPR	**TNR**	**TPR**	**TNR**
$K_{G_{\sigma=0,1^3}}$	90.25	98.91	89.5	98.55
$K_{G_{\sigma=0,1^2}}$	90.00	98.88	90.5	98.44
$K_{G_{\sigma=0,1}}$	2.25	89.13	5.0	84.11
$K_{G_{\sigma=1}}$	0.25	88.91	0.0	81.00
$K_{G_{\sigma=10}}$	32.00	92.44	20.0	91.05
$K_{G_{\sigma=50}}$	23.75	91.52	20.0	97.77
$K_{G_{\sigma=100}}$	24.00	91.55	20.0	99.22
$K_{G_{\sigma=500}}$	14.50	90.50	44.0	57.00
$K_{G_{\sigma=10^3}}$	38.75	93.19	44.0	54.88
K_L	90.25	98.91	89.0	98.55
K_P	**94.75**	**99.41**	**95.5**	**99.27**
$\bar{K}$	**94.75**	**99.41**	**95.5**	**99.27**
$\bar{\bar{K}}$	38.75	93.19	44.0	55.05
E_K	**94.75**	**99.41**	**95.5**	**99.27**

4.2. Real Data

Regarding real data, we also differentiate between multivariate and functional data. To test and compare proposals using multivariate data, we consider six databases from the UCI machine learning repository [24] which are available and properly described in [25]. The testing and comparison of our proposals using functional data are carried out over two functional data sets: (i) Poblenou NOx Emissions (NOx). This data set contains the nitrogen oxide (NO_x) emissions levels measured every hour by a control station in Poblenou in Barcelona (Spain). The data are publicly available in the R-package `'fda.usc'` [26]. In the data set, working day NO_x emissions, considered as regular data, and weekend day NO_x emissions considered as atypical data can be distinguished; (ii) Vertical Density Profiles (VDP). This data set contains 24 curves of Vertical Density Profiles which come from the manufacture of engineered woodboards. Each one consists of 314 measurements taken 0.002 inches (see [27] for further details). In Table 3, we give the details about the sample size, the dimension and the percentage of outlier observations for each of the data sets. The NOx and VDP data sets are illustrated in Figure 3.

Table 3. Summary of the data sets.

Data Set	Sample Size (n)	Dimension (d)	% of Outilers
Glass	214	9	9 (4.2%)
Vertebral	240	6	30 (12.5%)
Breast	683	9	239 (35%)
Breast (Diagnosis)	278	30	21 (5.6%)
Pima (Diabetes)	768	8	268 (35%)
Cardio	1831	21	176 (9,6%)
VDP	24	∞ (sampled at 314 points)	35 (30.3%)
NOx	115	∞ (sampled at 24 points)	3 (12.5%)

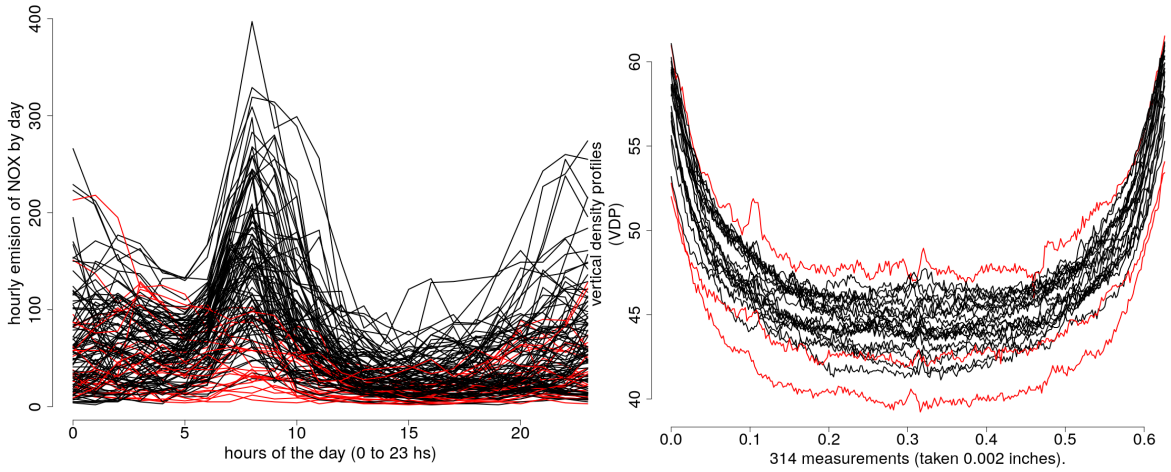

Figure 3. NOX (**left**) and VDP (**right**) functional data sets. The sample of regular curves in black ("—"), and abnormal curves in red ("—").

Table 4 shows the results of the experiment using real multivariate data. It can be observed that the best overall method in average is the weighted entropy proposal. In particular, this method attains the best results for two of the six data bases (Pima and Cardio), and for the rest of the sets its results are close to the best ones. Although the proposed methodologies seem to perform systematically better than other machine learning approaches, it is not clear, in terms of the AUC, whether for some data bases (Glass, Breast Cancer, Breast Cancer Diagnostic and Pima) the difference is statistically significant.

Table 4. Area under the ROC curve (AUC) for multivariate data sets.

Experiment	Glass	Vertebral	Breast Cancer	Breast Cancer (Diag.)	Pima	Cardio
$K_{G_{\sigma=0,1^3}}$	88.13	57.5	60.9	94.8	49.5	90.0
$K_{G_{\sigma=0,1^2}}$	88.51	71.4	60.8	94.7	49.7	89.5
$K_{G_{\sigma=0,1}}$	**91.49**	63.0	60.5	94.8	50.6	69.6
$K_{G_{\sigma=1}}$	82.87	66.9	62.2	94.4	71.9	65.3
$K_{G_{\sigma=10}}$	76.40	77.9	68.6	86.1	74.8	49.2
$K_{G_{\sigma=50}}$	51.49	**89.0**	68.1	85.2	48.6	46.8
$K_{G_{\sigma=100}}$	56.42	79.2	68.1	84.7	43.8	45.4
$K_{G_{\sigma=500}}$	53.90	85.8	68.1	65.2	52.4	64.8
$K_{G_{\sigma=10^3}}$	57.51	79.5	68.1	72.9	62.2	67.0
K_L	87.86	72.8	60.8	94.8	49.1	90.1
K_P	85.85	74.1	59.4	**96.3**	49.4	**94.8**
$\bar{K}$	85.85	66.4	62.9	94.3	48.8	**94.8**
$\overline{\overline{K}}$	88.13	82.2	60.7	94.4	**78.7**	63.4
E_K	88.13	82.2	61.4	94.4	**78.7**	**94.8**
LOF	76.8	59.3	56.4	86.9	70.9	59.6
HiCS	80.0	56.6	59.3	94.2	72.4	63.0

In Table 5, the results of the experiment using real functional data are presented. For the VDP data set, in terms of the AUC measure, the weighted entropy and the mean proposals perform as well as the best single kernels and the MBD. For the NOx data set, the best overall method is the one based on the calculation of the Karcher mean, followed closely by the MBD approach.

Table 5. Area under the ROC curve (AUC) for functional data sets.

Experiment	VDP	NOx
$K_{G_{\sigma=0,1^3}}$	**100.0**	65.7
$K_{G_{\sigma=0,1^2}}$	73.8	42.9
$K_{G_{\sigma=0,1}}$	88.8	48.1
$K_{G_{\sigma=1}}$	53.9	48.1
$K_{G_{\sigma=10}}$	50.7	48.1
$K_{G_{\sigma=50}}$	45.2	48.1
$K_{G_{\sigma=100}}$	52.3	48.1
$K_{G_{\sigma=500}}$	52.3	48.1
$K_{G_{\sigma=10^3}}$	52.3	48.1
K_L	**100.0**	57.7
K_P	**100.0**	52.0
$\bar{K}$	**100.0**	65.1
$\bar{\bar{K}}$	63.4	**70.9**
E_K	**100.0**	65.1
MBD	**100.0**	69.6
HMD	98.4	51.6
RTD	87.3	62.8

4.3. Robustness of the Karcher Mean

In this experiment we explore the robustness of the proposed procedure in the context of detection of atypical functional data. To this aim, we generate $n = 100$ independent sample paths from the following Gaussian stochastic model:

$$X(t) = \xi_1 \sin(t) + \xi_2 \cos(t) \text{ for } t \in [0, \pi], \text{ and } \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N\left[\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.75 & -0.5 \\ -0.5 & 0.75 \end{pmatrix}\right] \tag{14}$$

that is (ξ_1, ξ_2) follows a zero mean bi-variate normal distribution with covariance parameters $\sigma_{11} = \sigma_{22} = 0.75$ and $\sigma_{12} = \sigma_{21} = -0.5$. Using the representation techniques introduced in § 2, we can represent these curves as points in $\mathbb{R}^2$ and, moreover, we can estimate (by Maximum Likelihood) a covariance matrix $\hat{\Sigma}$ using this data representation. We replicate the previous generating process 10 times, obtaining 10 covariance matrices estimates, namely $\hat{\Sigma}_i$ for $i = 1, \ldots, 10$. Next, we construct the *mean estimated covariance matrix* as, $\bar{\hat{\Sigma}} = \sum_{i=1}^{10} \hat{\Sigma}_i/10$, and the *Karcher mean estimated covariance matrix*, $\bar{\bar{\Sigma}}(\hat{\Sigma}_1, \ldots, \hat{\Sigma}_{10})$. The estimations are illustrated in Figure 4-left, where each ellipse (in grey"—") corresponds to the following equation:

$$\frac{(x_1 \cos(\hat{\theta}_i) + x_2 \sin(\hat{\theta}_i))^2}{\sqrt{\hat{\lambda}_{1,i}\chi^2_{2,0.99}}} + \frac{(x_2 \cos(\hat{\theta}_i) - x_1 \sin(\hat{\theta}_i))^2}{\sqrt{\hat{\lambda}_{2,i}\chi^2_{2,0.99}}} = 1, \quad \text{for } i = 1, \ldots, 10,$$

where $\chi^2_{2,0.99}$ is the value of a Chi-square with two degrees of freedom that accumulates 0.99 probability, $\hat{\lambda}_{1,i}$ and $\hat{\lambda}_{2,i}$ are the estimated eigenvalues, corresponding to each estimate $\hat{\Sigma}_i$, and $\hat{\theta}_i$ is the estimated rotation angle with respect to the 'x_1' axis. In addition, in the same Figure, the estimated mean $\bar{\hat{\Sigma}}$ (its corresponding ellipse estimation) is shown in red ("- - -"), and in blue ("- - -") the Karcher mean. To introduce some anomaly in our data, in Figure 4-right, we added one ellipse constructed with an anomalous bivariate distribution with covariance matrix with elements $\sigma_{11} = \sigma_{22} = 7.5$ and

$\sigma_{12} = \sigma_{21} = -10$; this atypical covariance matrix corresponds to a different stochastic Gaussian model from the baseline introduced in Equation (14).

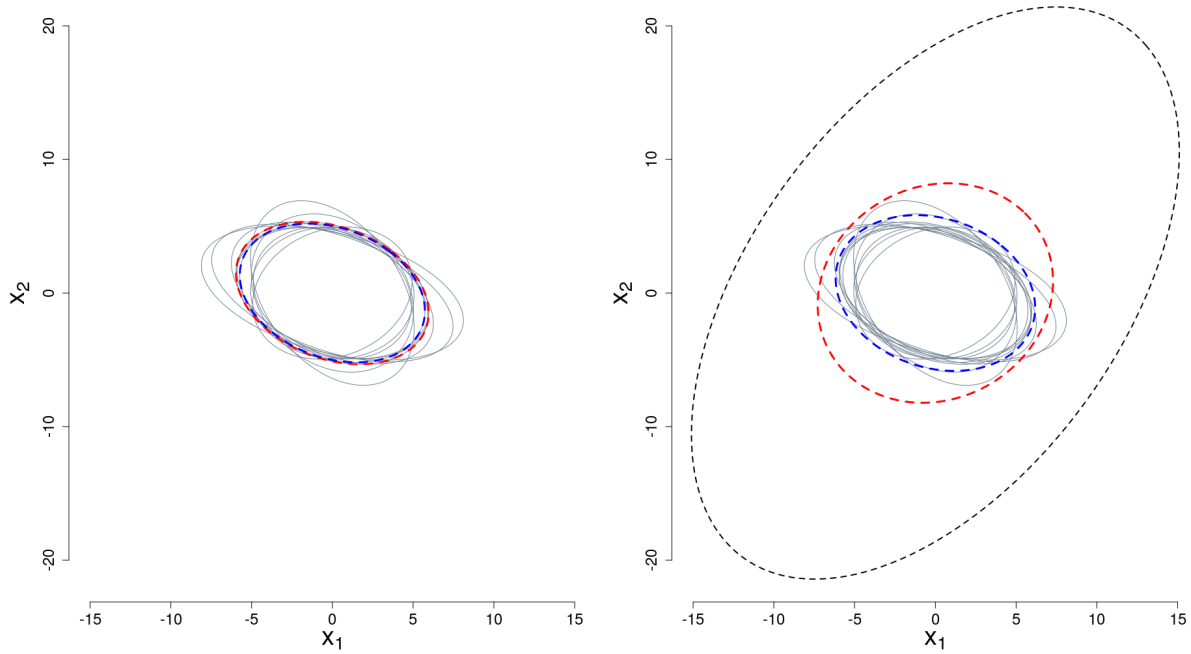

Figure 4. First two coordinates of ten 99th percentile ellipses ("—"). $\widehat{\overline{\Sigma}}$ ellipse in red ("- - -") and $\overline{\overline{\Sigma}}$ ellipse in blue ("- - -"). Left panel: Gaussian scenario; Right panel: Gaussian scenario contaminated with anomalous covariance matrix in black ("- - -").

It can be observed in Figure 4-left that the average covariance matrix and the Karcher mean of the covariance matrix generate similar 99th percentile ellipses. Since the generated covariance matrices $\widehat{\Sigma}_i$ are located in a small region within the cone of semi-definite positive matrices, such a region can be approximated by a linear subspace that contains the average covariance matrix. On the other hand, in Figure 4-right, the curvature of the cone is depicted by the difference in the dispersion of the anomalous covariance matrix, illustrated by the ellipse with a black-dashed line. In this scenario, the Karcher mean of the covariance matrices generates similar 99th percentile ellipse with respect to the regular scenario (left panel), which shows the robustness of the Karcher mean in the presence of outliers. Nevertheless, in the contaminated scenario (right panel), the 99th percentile ellipse generated with the simple average mean of the covariance matrices changes radically with respect to the regular scenario. The robustness in the estimation of the covariance matrix allows us to ensure that the procedure proposed in this paper, based on the estimation of the Karcher mean in the cone of positive definite matrices, will be useful when solving atypical functional data identification problems.

Last but not least, the relevant aspect of this numerical example is that, using the Karcher mean as an estimator of the center of the distribution of semi-definite positive matrices, we are minimizing the Riemannian distance, as it is defined in Section 3, and, as a consequence, the proposed method is able to identify the anomalous covariance matrix with respect to the pattern given by the rest of the distributions.

5. Discussion

In this work, we have explored how to combine different sources of information for anomaly detection within the framework of Entropy measures. We define entropies associated with the transformation induced by Mercer kernels, both for random variables and for data sets. We propose a new class of combination methods that generate a single Mercer kernel (that acts as a similarity measure) for anomaly detection purposes from a set of entropy measures in the context of density

estimation. In particular, three combination schemes have been proposed and analysed, namely: (i) an average of the kernel matrices; (ii) the mean in the manifold that contains the kernels; and (iii) a weighting scheme that assigns the weights as the solution of an optimization problem that seeks to maximize a particular kernel entropy. Such proposals, based on the idea of building the final combined kernel matrix within the same variety where the kernel matrices to be combined live, seem to be the most successful ones on average.

An innovative application of this methodology is the use of the Karcher mean as part of a method to identify anomalous covariance matrices. The success of this proposal is due to the fact that the Karcher mean acts as an estimator of the center of the distribution of semi-definite positive matrices, while minimizing their Riemannian distance, allowing the identification of the outlying matrices with respect to the pattern given by such an estimator.

A relevant aspect for the method applicability in real problems is its complexity and costs in comparison with other alternatives. The proposals whose structure is based on a linear combination of kernel matrices have a very low computational cost based on the computation of products of constants and sums of matrices. The proposal based on the use of the Karcher mean has the typical drawback of any semidefinite programming problem, that is, the computational and memory costs are related to the size of the matrices involved. Current systems are not able to deal with dense large matrices, given that processing time and memory grow quasi-exponentially as the size of the matrices increase. See [28] for a discussion on these aspects and current trends to improve the performance of methods for the solution of semidefinite programming problems. Most applications for general dense matrices in semidefinite programming involve a few hundred data cases. Fortunately, in this particular application (outlier detection), we do not need to work with the full database to success. Due to the presence of statistical regularities, a few thousand data cases will usually be enough to collect all the relevant statistical aspects of the data set at hand.

Further research is to be afforded, especially regarding the possibility of exploring other embeddings of the data. For instance, higher dimensional transformations specific for anomaly detection could be designed. In this regard, care should be taken with the scaling of such transformations, as dimensions with large magnitudes with respect to the others may lead to suboptimal results. In this work, for multivariate data, we have compared the methodologies proposed with some multivariate outlier detection techniques. In the future, systematic experiments comparing with other well known methodologies such as XBGOD [29], LODES [30], iForest [31] or MASS [32] are to be carried out. Regarding these multivariate techniques, another interesting research line is the extension of such methodologies to functional data analysis. In this regard, suitable multivariate representations of functional data similar to those in [2] should be explored.

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

Funding: This research was funded by CONICET Argentina project 20020150200110BA, the Spanish Ministry of Economy and Competitiveness projects ECO2015–66593-P and GROMA (MTM2015-63710-P). The APC was funded by Project GROMA (MTM2015-63710-P).

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

References

1. Moguerza, J.M.; Muñoz, A. Support vector machines with applications. *Stat. Sci.* **2006**, *21*, 322–336. [CrossRef]
2. Muñoz, A.; González, J. Representing functional data using support vector machines. *Pattern Recognit. Lett.* **2010**, *31*, 511–516. [CrossRef]
3. Carl, G.; Sollich, P. Model selection for support vector machine classification. *Neurocomputing* **2003**, *55*, 221–249.
4. Chapelle, O.; Vapnik, V.; Bousquet, O.; Mukherjee, S. Choosing multiple parameters for support vector machines. *Mach. Learn.* **2002**, *46*, 131–159. [CrossRef]

5. Wahba, G. Support Vector machines, Reproducing Kernel Hilbert Spaces, and randomized GACV. In *Advances in Kernel Methods: Support Vector Learning*; Schoelkopf, B., Burges, C., Smola, A., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp. 69–88.
6. Lanckriet, G.R.; Cristianini, N.; Bartlett, P.; Ghaoui, L.E.; Jordan, M.I. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.* **2004**, *5*, 27–72.
7. De Diego, I.M.; Muñoz, A.; Moguerza, J.M. Methods for the combination of kernel matrices within a support vector framework. *Mach. Learn.* **2010**, *78*, 137–172. [CrossRef]
8. Hsing, T.; Eubank, R. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
9. Smola, A.; Gretton, A.; Song, L.; Schölkopf, B. A Hilbert space embedding for distributions. In Proceedings of the International Conference on Algorithmic Learning Theory, Sendai, Japan, 1–4 October 2007; pp. 13–31.
10. Berlinet, A.; Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*; Springer: New York, NY, USA, 2011.
11. Kimeldorf, G.; Wahba, G. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **1971**, *33*, 82–94. [CrossRef]
12. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960.
13. Martos, G.; Hernández, N.; Muñoz, A.; Moguerza, J.M. Entropy measures for stochastic processes with applications in functional anomaly detection. *Entropy* **2018**, *20*, 33. [CrossRef]
14. Vandenberghe, L.; Boyd, S. Semidefinite programming. *SIAM Rev.* **1996**, *38*, 49–95. [CrossRef]
15. Karcher, H. Riemannian center of mass and mollifier smoothing. *Commun. Pure Appl. Math.* **1977**, *30*, 509–541. [CrossRef]
16. Arnaudon, M.; Barbaresco, F.; Yang, L. Medians and means in Riemannian geometry: Existence, uniqueness and computation. *arXiv* **2011**, arXiv:1111.3120.
17. Bini, D.A.; Iannazzo, B. Computing the Karcher mean of symmetric positive definite matrices. *Linear Algebra Appl.* **2013**, *438*, 1700–1710. [CrossRef]
18. Breunig, M.M.; Kriegel, H.-P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, TX, USA, 15–18 May 2000; pp. 93–104.
19. Keller, F.; Müller, E.; Böhm, K. HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. In Proceedings of the 2012 IEEE 28th International Conference on Data Engineering, Washington, DC, USA, 1–5 April 2012; pp. 1037–1048.
20. López-Pintado, S.; Romo, J. On the concept of depth for functional data. *J Am. Stat. Assoc.* **2009**, *104*, 718–734. [CrossRef]
21. Cuevas, A.; Febrero, M.; Fraiman, R. Robust estimation and classification for functional data via projection-based depth notions. *Comput. Stat.* **2007**, *22*, 481–496. [CrossRef]
22. Cuesta-Albertos, J.A.; Nieto-Reyes, A. The random Tukey depth. *Comput. Stat. Data Anal.* **2008**, *52*, 4979–4988. [CrossRef]
23. Gelman, A.; Meng, X.L. A note on bivariate distributions that are conditionally normal. *Am Stat.* **1991**, *45*, 125–126.
24. Blake, C.L.; Merz, C.J. UCI Repository of Machine Learning Databases. Available online: http://archive.ics.uci.edu/ml/index.php (accessed on 10 September 2018).
25. Rayana, S. ODDS Library. Available online: http://odds.cs.stonybrook.edu (accessed on 10 September 2018).
26. 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. [CrossRef]
27. Moguerza, J.M.; Muñoz, A.; Psarakis, S. Monitoring nonlinear profiles using support vector machines. In *Iberoamerican Congress on Pattern Recognition*; Springer: Berlin, Germany, 2007; pp. 574–583.
28. Zheng, Y.; Yan, Y.; Liu, S.; Huang, X.; Xu, W. An Efficient Approach to Solve the Large-Scale Semidefinite Programming Problems. *Math. Probl. Eng.* **2012**, *2012*, 764760. [CrossRef]
29. Zhao, Y.; Hryniewicki, M.K. XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Rio, Brazil, 8–13 July 2018.

30. Sathe, S.; Aggarwal, C. LODES: Local Density Meets Spectral Outlier Detection. In Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, FL, USA, 5–7 May 2016; pp. 171–179.
31. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (2008), Pisa, Italy, 15–19 December 2008; pp. 413–422.
32. Ting, K.M.; Chuan, T.S.; Liu, F.T. *Mass: A New Ranking Measure for Anomaly Detection*; Technical Report TR2009/1; Gippsland School of Information Technology, Monash University: Victoria, Australia, 2009.