UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 13-15
Statistics and Econometrics Series 14
May 2013

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

A New Distance for Data Sets (and Probability Measures) in a RKHS Context

Gabriel Martos and Alberto Muñoz[1]

Abstract

In this paper we define distance functions for data sets (and distributions) in a RKHS context. To this aim we introduce kernels for data sets that provide a metrization of the set of points sets (the power set). An interesting point in the proposed kernel distance is that it takes into account the underlying (data) generating probability distributions. In particular, we propose kernel distances that rely on the estimation of density level sets of the underlying distribution, and can be extended from data sets to probability measures. The performance of the proposed distances is tested on a variety of simulated distributions plus a couple of real pattern recognition problems.

Keywords: Probability measures, kernel, level sets, distances for data sets.

This work was partially supported by projects **DGUCM 2008/00058/002**, **MEC 2007/04438/001** and **MIC 2012/00084/00**.

[1]Department of Statistics, University Carlos III of Madrid, Spain. email: {gabrielalejandro.martos, alberto.munoz}@uc3m.es

A New Distance for Data Sets (and Probability Measures) in a RKHS Context

Gabriel Martos and Alberto Muñoz.

Department of Statistics.

Universidad Carlos III de Madrid.

Abstract

In this paper we define distance functions for data sets (and distributions) in a RKHS context. To this aim we introduce kernels for data sets that provide a metrization of the set of points sets (the power set). An interesting point in the proposed kernel distance is that it takes into account the underlying (data) generating probability distributions. In particular, we propose kernel distances that rely on the estimation of density level sets of the underlying distribution, and can be extended from data sets to probability measures. The performance of the proposed distances is tested on a variety of simulated distributions plus a couple of real pattern recognition problems.

Keywords: Probability measures, kernel, level sets, distances for data sets.

1 Introduction and Related Work

The study of distances between data sets lies at the very core of many methods of cluster analysis, where we have to choose a distance for merging ongoing clusters [47]. Metrics for data sets can be useful in shape recognition and image classification [36, 39, 20], in genetics [1], time series [32] or geometric inference [9], where the Hausdorff distance is a common choice.

However, the Hausdorff distance is designed for dealing with geometric objects, and there are other settings where it is convenient to take into account the underlying statistical distribution to calculate distances between data sets; a classical example is Mahalanobis distance for data sets arising from normal multivariate distributions. Our purpose in this paper is to design a distance between data sets that takes into account the underlying data distribution.

To this aim we will focus on the study of distances between probability measures (PM), also known as distributions. Classical examples of application of distances between PMs in Statistics are homogeneity tests, independence tests and goodness of fit test problems, where the goal is to decide if some available data samples come from the same population or not. These problems can be solved by choosing an appropriate distance between PM. Examples are the Pearsons's goodness of fit test based on the use of the χ^2 distance, and the Kolmogorov-Smirnoff statistics, that uses the L_1 distance between the empirical and theoretical distribution functions. Other examples of distances between PM can also be founded in Clustering [5], Image Analysis [13], Time Series Analysis [25], Econometrics [24, 44] and Text Mining [21], just to name a few.

Next we summarize the most important families of distances between PM. For an exhaustive review of distances between probability distributions and theoretical results, see for instance [12, 48, 26], and references therein.

One of the largest family of dissimilarities between probability distributions is the f-divergences class [11]. Consider two probability densities, say $\mathbb{P}$ and $\mathbb{Q}$, defined on a measurable space $(X, \mathcal{F}, \mu)$, where X is a sample space, $\mathcal{F}$ a σ-algebra of measurable subsets of X and $\mu : \mathcal{F} \to \mathbb{R}^+$ the ambient σ-additive measure. For a convex function f and assuming that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, then the f-divergence from $\mathbb{P}$ to $\mathbb{Q}$ is defined by:

$$d_f(\mathbb{P}, \mathbb{Q}) = \int_X f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q}. \tag{1}$$

Some well known particular cases: for $f(t) = \frac{|t-1|}{2}$ we obtain the *Total Variation* metric; $f(t) =$

$(t-1)^2$ yields the χ^2-distance; $f(t) = (\sqrt{t}-1)^2$ yields the *Hellinger* distance.

The second important family of dissimilarities between probability distributions is made up of Bregman Divergences: Consider a continuously-differentiable real-valued and strictly convex function φ and define:

$$d_\varphi(\mathbb{P}, \mathbb{Q}) = \int_X (\varphi(f_\mathbb{P}) - \varphi(f_\mathbb{Q}) - (f_\mathbb{P} - f_\mathbb{Q})\varphi'(f_\mathbb{Q}))\, d\mu(x), \tag{2}$$

where $f_\mathbb{P}$ and $f_\mathbb{Q}$ represent the density functions for $\mathbb{P}$ and $\mathbb{Q}$ respectively and $\varphi'(f_\mathbb{Q})$ is the derivative of φ evaluated at $f_\mathbb{Q}$ (see [15, 10] for further details). Some examples of Bregman divergences: $\varphi(t) = t^2$ yields the Euclidean distance between $f_\mathbb{P}$ and $f_\mathbb{Q}$ (in L_2); $\varphi(t) = t\log(t)$ yields the *Kullback Leibler* (KL) Divergence; and for $\varphi(t) = -\log(t)$ we obtain the *Itakura-Saito* distance. In general d_f and d_φ are not metrics because the lack of symmetry and because they do not necessarily satisfy the triangle inequality.

A third interesting family of PM distances are integral probability metrics (IPM) [48, 26]. Consider a class of real-valued bounded measurable functions on X, say $\mathcal{H}$, and define the IPM between $\mathbb{P}$ and $\mathbb{Q}$ as

$$d_\mathcal{H}(\mathbb{P}, \mathbb{Q}) = \sup_{f\in\mathcal{H}} \left| \int f d\mathbb{P} - \int f d\mathbb{Q} \right|. \tag{3}$$

If we choose $\mathcal{H} = \{h : \|h\|_\infty \leq 1\}$ then $d_\mathcal{H}$ is the Total Variation distance; when $\mathcal{H} = \{\mathbb{1}_{[(-\infty,x)]} : x \in \mathbb{R}^d\}$, $d_\mathcal{H}$ is the Kolmogorov distance; if $\mathcal{H} = \{e^{\sqrt{-1}\langle\omega,.\rangle} : \omega \in \mathbb{R}^d\}$ the metric computes the maximum difference between characteristics functions. In [40] the authors propose to choose $\mathcal{H}$ as a Reproducing Kernel Hilbert Space and study conditions on $\mathcal{H}$ to obtain proper metrics $d_\mathcal{H}$.

However, there is a serious problem to implement the above described distance functions: in real life we do not know the density functions corresponding to the samples under consideration. For instance suppose we want to estimate the KL divergence (a particular case of (1) taking $f(t) = -\log t$) between two continuous distributions $\mathbb{P}$ and $\mathbb{Q}$ from two given samples. In order

to do this we must choose a number of regions, N, and then estimate the density functions for $\mathbb{P}$ and $\mathbb{Q}$ in the N regions to yield the following estimation:

$$\widehat{KL}(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{N} \hat{p}_i \log \frac{\hat{p}_i}{\hat{q}_i}, \tag{4}$$

where $\{\hat{p}_i, \hat{q}_i\}_{i=1}^{N}$ denotes the estimated density in each region of $\mathbb{P}$ and $\mathbb{Q}$, respectively. It is well known the difficulty in estimating density functions, especially in high dimensional settings.

Non parametric estimators often play a role in estimating such distances. In practical situations there is usually available a not huge data sample, and the use of purely non parametric estimators often results in poor performance [18]. It is also well known that the non-parametric estimations suffers from the "curse of dimensionality": the estimation of general distribution functions becomes intractable as dimension arises. Another important drawback in non-parametric density estimation is the high computation time and huge storage required. This motivates the need of seeking metrics for probability distributions that do not explicitly rely on the estimation of the corresponding probability/distribution functions.

An appealing point of view, initiated by Fisher and Rao [7, 2, 4] and continued with recent development of Functional Data Analysis and Information Geometry Methods [35, 3, 34, 40], is to consider probability distributions as points belonging to some manifold, and then take advantage of the manifold structure to derive appropriate metrics for distributions. This point of view is used, for instance, in the field of Image Analysis [33, 13].

In this work we elaborate on the idea of considering a kernel function for data points with reference to a distribution function, that will be extended to a kernel (and to a distance) for data sets. The article is organized as follows: In Section 2 we introduce kernel functions for data sets with uniform distributions. Section 3 introduces a new metric for general data sets based on the estimation of density level sets. Section 4 shows the performance of the proposed metric on both, simulated and real data sets. Section 5 concludes and shows the next steps for short term future work.

2 A kernel for data sets with reference to a distribution

Consider a measure space $(X, \mathcal{F}, \mu)$, where X is the sample space (a compact set of a real vector space in this work), $\mathcal{F}$ a σ-algebra of measurable subsets of X and $\mu : \mathcal{F} \to \mathbb{R}^+$ the ambient σ-finite measure. A **probability measure** (PM) $\mathbb{P}$ is a σ-additive finite measure absolutely continuous w.r.t. μ that satisfies the three Kolmogorov axioms. By Radon-Nikodym theorem, there exists a measurable function $f_{\mathbb{P}} : X \to \mathbb{R}^+$ (the density function) such that $P(A) = \int_A f_{\mathbb{P}} d\mu$, and $f_{\mathbb{P}} = \frac{dP}{d\mu}$ is the Radon-Nikodym derivative.

From now on we focus on data sets generated from (unknown) PM and we will only mention the corresponding distributional distance measures in Section 3.

It is possible to define distances between data sets using the RKHS representation. The authors in [34, 40] define the induced kernel distance for the sets of points $A = \{x_i\}_{i=1}^n \in X$ and $B = \{y_j\}_{j=1}^m \in X$, by:

$$D_K^2(A, B) = \underbrace{\sum_{x \in A} \sum_{x' \in A} K(x, x') + \sum_{y \in B} \sum_{y' \in B} K(y, y')}_{\text{self similarity}} \underbrace{- 2 \sum_{x \in A} \sum_{y \in B} K(x, y).}_{\text{cross similarity}}, \tag{5}$$

where the kernel measures the *similarity* between the sets A and B. We will work on the construction of a kernel family of functions for data sets (later extended to PM) to embed the data sets into the RKHS structure. This allows us to define a distance measure for data sets, using the natural metric defined in Equation 5. The metrization obtained via the kernel embedding allows us to represent the data sets by points in a finite dimensional vector space, procedure that situates the problem at hand in the context of Functional Data Analysis.

Usually the available data are given as a finite sample. We will consider two *iid* samples $A = s_n(\mathbb{P}) = \{x_i\}_{i=1}^n \in \mathcal{P}(X)$, where $\mathcal{P}(X)$ denotes the power set of X, and $B = s_m(\mathbb{Q}) = \{y_j\}_{j=1}^m \in \mathcal{P}(X)$, generated from the density functions $f_{\mathbb{P}}$ and $f_{\mathbb{Q}}$, respectively and defined on the same

measure space. Define $r_A = \min d(x_l, x_s)$, where $x_l, x_s \in A$. Then r_A gives the minimum resolution for data set A: If a point $z \in X$ is located at a distance smaller than r_A from a point $x \in A$ then, taken $\mathbb{P}$ as reference measure, it is impossible to differentiate z from x. That is, it is not possible to reject the null hypothesis that z is generated from $\mathbb{P}$, given that z is closer to x than any other point from the same distribution. This suggests the following definition:

Definition 1. ***Indistinguishability with respect to a distribution.*** *Let $x \in A$, where A denotes a set of points generated from the probability measure $\mathbb{P}$, and $y \in X$. We say that y is* ***indistinguishable*** *from x whit respect to the measure $\mathbb{P}$ in the set A when $d(x, y) \leq r_A = \min d(x_l, x_s)$, where $x_l, x_s \in A$. We will denote this relationship as: $y \overset{A(\mathbb{P})}{=} x$.*

Given the sets $A = s_n(\mathbb{P})$ and $B = s_m(\mathbb{Q})$, we want to build kernel functions $K : X \times X \to [0, 1]$, such that $K(x, y) = 1$ when $y \overset{A(\mathbb{P})}{=} x$ or $x \overset{B(\mathbb{Q})}{=} y$, and $K(x, y) = 0$ if $y \overset{A(\mathbb{P})}{\neq} x$ and $x \overset{B(\mathbb{Q})}{\neq} y$. For this purpose we can consider smooth indicator functions, for example:

Definition 2. ***Smooth indicator functions.*** *Let $r > 0$ and $\gamma > 0$, define a family of smooth indicator functions with center in x as:*

$$f_{x,r,\gamma}(y) = \begin{cases} e^{-\frac{1}{(\|x-y\|^\gamma - r^\gamma)^2} + \frac{1}{r^{2\gamma}}} & \text{if } \|x - y\| \leq r \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

We represent in Fig. 1 the indicator function in dimensions 1 and 2, with parameters $\gamma = 2$ and $r = 1$. The smooth function $f_{x,r,\gamma}(y)$ act as a bump function with center in the coordinate point given by x: $f_{x,r,\gamma}(y) \approx 1$ for $y \in B_r(x)$, and $f_{x,r,\gamma}(y)$ decays to zero out of $B_r(x)$, according to a rate that depends upon the shape parameter γ.

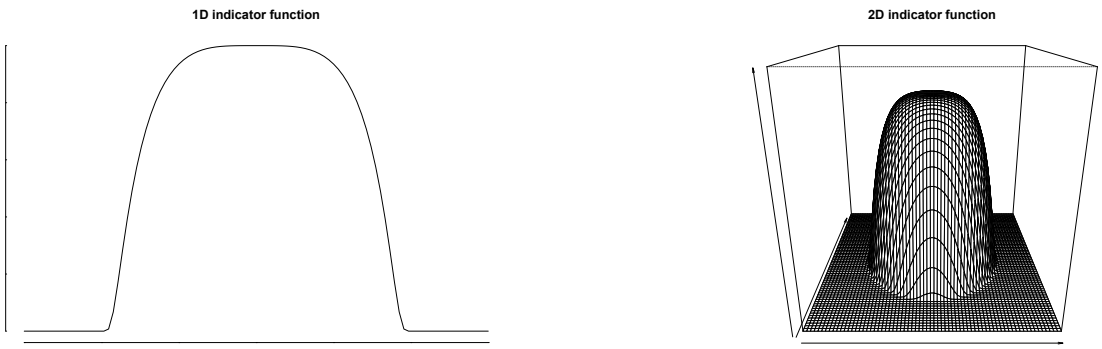

Figure 1: Smooth indicator functions. (a) 1D case. (b) 2D case.

It is clear from the Definition 2 that: $\lim_{r \to 0} f_{x,r,\gamma}(y) = \langle f_{x,r,\gamma}(y), \delta_x() \rangle = \delta_x(y) = \delta(y - x)$, where $\delta()$ is the Dirac-Delta generalized function; for further details on Generalized Functions refers to [42]. This result is important in order to give an asymptotic interpretation to the proposed metric. We are now ready to define an indicator kernel function that summarizes the distinguishability relationship between two points.

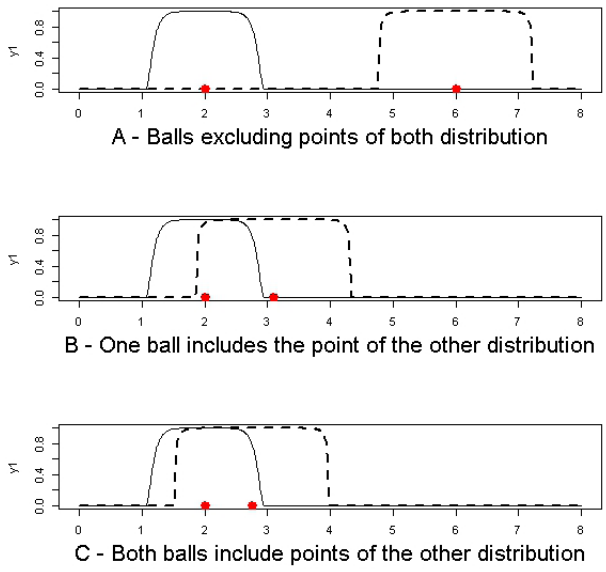

Figure 2: Illustration of the $\overset{A(\mathbb{P})}{=}$ and $\overset{B(\mathbb{Q})}{=}$ relationship using smooth indicator functions.

Definition 3. ***Distributional indicator kernel.*** *Given* $A = s_n(\mathbb{P})$ *and* $B = s_m(\mathbb{Q})$*, define* $K_{A,B} : X \times X \to [0,1]$ *by:*

$$K_{A,B}(x,y) = f_{x,r_A,\gamma}(y) + f_{y,r_B,\gamma}(x) - f_{x,r_A,\gamma}(y) f_{y,r_B,\gamma}(x), \tag{7}$$

where $r_A = \min d(x_l, x_s)$, with $x_l, x_s \in A$, $r_B = \min d(y_l, y_s)$, with $y_l, y_s \in B$ and γ it is a shape parameter. Now, if $d(x,y) > r_A$ and $d(x,y) > r_B$ (see Fig. 2A) then $K_{A,B}(x,y) = 0$: $x \in A \backslash B$ w.r.t. $\mathbb{Q}$ and $y \in B \backslash A$ w.r.t. $\mathbb{P}$. If $d(x,y) > r_A$ but $d(x,y) < r_B$, then $y \in B \backslash A$ w.r.t. $\mathbb{P}$, but $x \overset{B(\mathbb{Q})}{=} y$ at radius r_B and $K_{A,B}(x,y) = 1$ (Fig. 2B). If $d(x,y) < r_A$ but $d(x,y) > r_B$, then $x \in A \backslash B$ w.r.t. $\mathbb{Q}$, but $y \overset{A(\mathbb{P})}{=} x$ at radius r_A and $K_{A,B}(x,y) = 1$. Finally, if $d(x,y) < r_A$ and $d(x,y) < r_B$, then $K_{A,B}(x,y) = 1$ and $y \overset{A(\mathbb{P})}{=} x$ at radius r_A and $x \overset{B(\mathbb{Q})}{=} y$ at radius r_B (Fig. 2C).

Definition 4. ***Kernel for data sets.*** *Given* $A = s_n(\mathbb{P})$ *and* $B = s_m(\mathbb{Q})$*, we consider kernels* $K : \mathcal{P}(X) \times \mathcal{P}(X) \to [0,1]$*, where* $\mathcal{P}(X)$ *denotes the power set of* X*, and for* C *and* D *in* $\mathcal{P}(X)$*, define:*

$$K(C,D) = \sum_{x \in C} \sum_{y \in D} K_{A,B}(x,y). \tag{8}$$

When $C = A$ and $D = B$, we can interpret $K(A,B)$ as a measure for $A \cap B$ by counting, using as equality operators $\overset{A(\mathbb{P})}{=}$ and $\overset{B(\mathbb{Q})}{=}$, the points 'in common': $\mu_{K_{A,B}}(A \cap B) = K(A,B)$. The expression in Equation 8 is directly related with the works in [34, 40]. We next give an explicit kernel function to measure the *cross-similarity* between the sets A and B, as in Equation 5, based on the idea of "indistinguishability" (see Definition 1). Given the identity $A \cup B = \overbrace{(A - B) \cup (B - A)}^{A \Delta B} \cup (A \cap B)$, we will define $\mu_{K_{A,B}}(A \cup B) = N$, where $N = n + m = \#(A \cup B)$, is the counting measure of the set $A \cup B$. Therefore $\mu_{K_{A,B}}(A \Delta B) = N - \mu_{K_{A,B}}(A \cap B)$, and we can take this expression (dividing by N) as a definition for the distance between the sets A and B. Another way to derive

an equivalent distance, is using the expression of Equation 5:

$$\begin{aligned} D_K^2(A,B) &= K(A,A) + K(B,B) - 2K(A,B), \\ &= n + m - 2K(A,B), \\ &= N - 2\sum_{x\in A}\sum_{y\in B} K_{A,B}(x,y). \end{aligned} \tag{9}$$

Note that this kernel distance is defined in terms of it square. We will present a slightly different distance in Definition 5.

It is straight to check that when the size of the sets A and B increases, then the respective radius tends to: $r_A \overset{n\to\infty}{\to} 0$ and $r_B \overset{m\to\infty}{\to} 0$. By Definition 2, the smooth indicator functions tends to the Dirac-Delta Generalized function, therefore $K(A,B) \overset{n,m\to\infty}{\to} \mu(A\cap B)$. In this sense, the Kernel for data sets defined in Equation 8 is an unbiased estimator of the measure that correspond to the intersection between the sets A and B: $\mu(A\cap B)$.

In the general case, $K(C,D)$ can be interpreted as a measure for $C\cap D$ by counting, using as equality operators $\overset{A(\mathbb{P})}{=}$ and $\overset{B(\mathbb{Q})}{=}$, the points 'in common': $\mu_{K_{A,B}}(C\cap D) = K(C,D)$. Therefore the respective distance between C and D obtained with the use of $K(C,D)$, is conditioned to a "resolution" level determined by the sets A and B (this is r_A and r_B).

Definition 5. ***Distance between data sets.*** *Given $A = s_n(\mathbb{P})$ and $B = s_m(\mathbb{Q})$, we define the kernels distance for C and D in $\mathcal{P}(X)$:*

$$d_K(C,D) = 1 - \frac{K(C,D)}{N}, \tag{10}$$

where $N = n_C + n_D = \#(C\cup D)$ and represent the measure of the set $C\cup D$. It is straight to check that $d_K(C,D)$ is a semimetric (using the equality operators $y \overset{A(\mathbb{P})}{=} x$ or $y \overset{B(\mathbb{Q})}{=} x$ where it corresponds).

When $C = A$ and $D = B$ and size of both sets increases, then: $\mu_{K_{A,B}}(A \cap B) \overset{n,m\to\infty}{\to} \mu(A \cap B)$ and $\mu_{K_{A,B}}(A \cup B) \overset{n,m\to\infty}{\to} \mu(A \cup B)$, therefore $\lim_{n,m\to\infty} d_K(A, B) = 1 - \frac{\mu(A\cap B)}{\mu(A\cup B)}$. We can interpret the limit of the proposed distance as a Jaccard distance for data sets. This last expression is clearly related with the distance defined in Equations 5 and 9, because as $K(A, A) \overset{n\to\infty}{\to} \mu(A)$ and $K(B, B) \overset{m\to\infty}{\to} \mu(B)$, then $\lim_{n,m\to\infty} D_K^2(A, B) = \mu(A) + \mu(B) - 2\mu(A \cap B)$; therefore:

$$\lim_{n,m\to\infty} d_K(A, B) = \lim_{n,m\to\infty} \frac{D_K^2(A, B)}{\mu_{K_{A,B}}(A \cup B)}$$

We want to exemplify how the proposed metric works with a synthetic example. We generate two *iid* samples $s_{500}(\mathbb{P}) = A$ and $s_{500}(\mathbb{Q}) = B$, drawn from the bi-dimensional uniform density function inside a ball, with center in zero and radius $r = 1$ ($f_{\mathbb{P}} = U_2(\mu = (0, 0), r = 1)$), and the bi-dimensional Normal distribution function, with parameters $\mu = (0, 0)$ and $\mathbf{\Sigma} = \mathbf{I}_2$ ($f_{\mathbb{Q}} = N_2(\mu = (0, 0), \mathbf{\Sigma} = \mathbf{I}_2)$). We generate new sets: A' and B', by displacing all the points in the sets A and B a constant distance in the same direction. In Figure 3 we represent the sets A, A', B and B'.

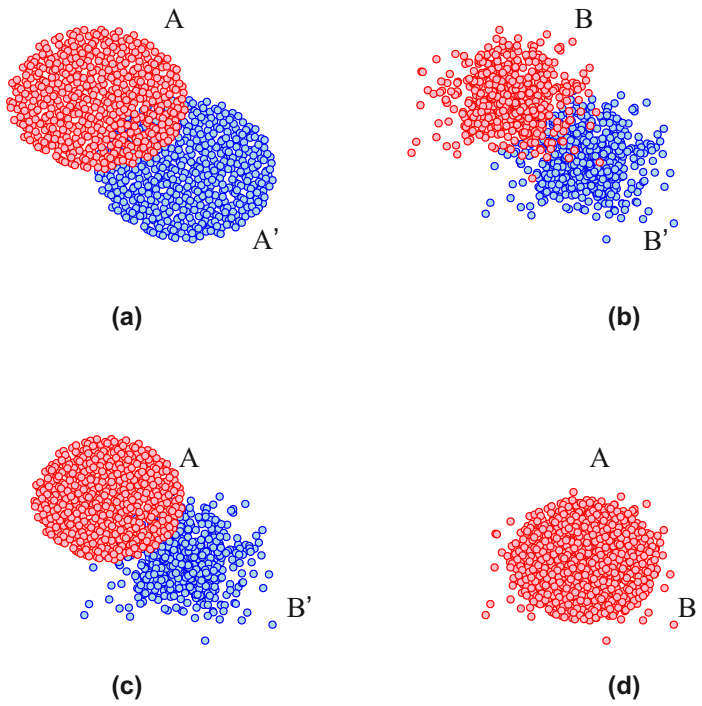

Figure 3: (a) A and A', (b) B and B', (c) A and B', (d) A and B

Therefore using Definition 5, the distance between sets B and B' should be larger compared with the distance between the sets A and A', because as the intersection between the last two sets seems to be bigger, the sets A and A' are more "similar" in relation with the sets B and B'. To verify the similarity relations between the sets, we compute the distance matrix in Table 1:

Table 1: Matrix of distances between data sets: A, A', B and B'.

	A	A'	B	B'
A	0	0.792	0.104	0.821
A'		0	0.823	0.104
B			0	0.864
B'				0

According to Table 1:

$$d_K(B, B') \geq d_K(A, B') \approx d_K(A', B) \geq d_K(A, A') \geq d_K(A, B) \approx d_K(A', B'),$$

that agrees with the visual inspection of Figure 3. For instance: $d_K(B, B') = .864 \geq d_K(A, A') = .792$.

3 A metric for data sets based on the estimation of level sets

In a general distribution, distances between sample points vary depending on the generating PM. Hence, using constant radii to determine the "distinguishability" relationship between points is only adequate if we are working with the uniform PM. In this section we propose a solution to this problem by splitting each data set in density level sets, and then considering difference sets between consecutive density levels, for which density is approximately constant.

Definition 6. ***α-level sets:*** *Given a PM $\mathbb{P}$ with density function $f_{\mathbb{P}}$, α-level sets or minimum volume sets, are defined by $S_\alpha(f_{\mathbb{P}}) = \{x \in X | \, f_{\mathbb{P}}(x) \geq \alpha\}$, such that $P(S_\alpha(f_{\mathbb{P}})) = 1 - \nu$, where $f_{\mathbb{P}}$ is the density function and $0 < \nu < 1$. If we consider an ordered sequence $\alpha_1 < \ldots < \alpha_k$, $\alpha_i \in (0, 1)$, then $S_{\alpha_{i+1}}(f_{\mathbb{P}}) \subseteq S_{\alpha_i}(f_{\mathbb{P}})$.*

Let us define $A_i(\mathbb{P}) = S_{\alpha_i}(f_{\mathbb{P}}) - S_{\alpha_{i+1}}(f_{\mathbb{P}})$, $i \in \{1, \ldots, k-1\}$. We can choose $\alpha_1 \simeq 0$ and $\alpha_k \geq \max_{x \in X} f_{\mathbb{P}}(x)$ (which exists, given that X is compact and $f_{\mathbb{P}}$ continuous); then $\bigcup_i A_i(\mathbb{P}) \simeq Supp(\mathbb{P}) = \{x \in X | \, f_{\mathbb{P}}(x) \neq 0\}$ (equality takes place when $k \to \infty$, $\alpha_1 \to 0$ and $\alpha_k \to 1$). Note that given the definition of the A_i, if $A_i(\mathbb{P}) = B_i(\mathbb{Q})$ for every i when $k \to \infty$, then $\mathbb{P} = \mathbb{Q}$.

Given the definition of the A_i-level set, both $\mathbb{P}$ and $\mathbb{Q}$ are approximately constant on A_i and B_i level sets, respectively. Therefore the use of a constant radii is again adequate when we compare the distance between the sets A_i and B_i.

3.1 Estimation of level sets

To estimate level sets from a data sample we present the following definitions and theorems, adapted from [29, 30, 27, 28].

Definition 7. ***Neighbourhood Measures***. *Consider a random variable X with density function $f(x)$ defined on $\mathbb{R}^d$. Let S_n denote the set of random independent identically distributed (iid) samples of size n (drawn from f). The elements of S_n take the form $s_n = (x_1, \cdots, x_n)$, where $x_i \in \mathbb{R}^d$. Let $M : \mathbb{R}^d \times S_n \longrightarrow \mathbb{R}$ be a real-valued function defined for all $n \in \mathbb{N}$. (a) If $f(x) < f(y)$ implies $\lim_{n\to\infty} P(M(x, s_n) > M(y, s_n)) = 1$, then M is a* ***sparsity measure****. (b) If $f(x) < f(y)$ implies $\lim_{n\to\infty} P(M(x, s_n) < M(y, s_n)) = 1$, then M is a* ***concentration measure****.*

Example. Consider the distance from a point x to its k^{th}-nearest neighbour in s_n, $x^{(k)}$: $M(x, s_n) = d_k(x, s_n) = d(x, x^{(k)})$: it is a sparsity measure.

The Support Neighbour Machine [29] solves the following optimization problem:

$$\begin{aligned} \max_{\rho,\xi} \quad & \nu n\rho - \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & g(x_i) \geq \rho - \xi_i \,, \\ & \xi_i \geq 0, \qquad\qquad i = 1, \ldots, n \,, \end{aligned} \tag{11}$$

where $g(x) = M(x, s_n)$ is a sparsity measure. We present now a theorem taken from [29, 30]:

Theorem 1. *The set $R_n = \{x : h_n(x) = sign(\rho_n^* - g_n(x)) \geq 0\}$ converges to a region of the form $S_\alpha(f) = \{x | f(x) \geq \alpha\}$, such that $P(S_\alpha(f)) = \nu$. Therefore, the Support Neighbour Machine estimates a density contour cluster $S_\alpha(f)$ (around the mode).*

Hence, we take $\hat{A}_i(\mathbb{P}) = \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}}) - \hat{S}_{\alpha_i}(f_{\mathbb{P}})$, where $\hat{S}_{\alpha_i}(f_{\mathbb{P}})$ is estimated by R_n defined above.

Definition 8. ***Weighted level-set distance.*** *Consider data sets $A = s_n(\mathbb{P})$ and $B = s_m(\mathbb{Q})$, generated from PMs $\mathbb{P}$ and $\mathbb{Q}$, respectively. Choose a partition $\alpha_1 < \alpha_2 < \ldots < \alpha_k$, $\alpha_i \in (0, 1)$*

and denote by $\hat{A}_i(\mathbb{P}) = \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}}) - \hat{S}_{\alpha_i}(f_{\mathbb{P}})$ *the estimation of* $A_i = S_{\alpha_{i+1}}(f_{\mathbb{P}}) - S_{\alpha_i}(f_{\mathbb{P}})$ *based on set* A*; and define similarly,* $\hat{B}_i(\mathbb{Q})$*. Then we define the weighted* α*-level set distances between the sets* A *and* B *by*

$$d(A, B) = \sum_{i=1}^{k-1} w_i d_K(A_i, B_i), \tag{12}$$

where $w_1, \ldots, w_{k-1} \in \mathbb{R}^+$. Thus

$$d_K(A, B) = \sum_{i=1}^{k-1} w_i \left(1 - \frac{K(A_i, B_i)}{\#(A_i \cup B_i)}\right) = \sum_{i=1}^{k-1} w_i \left(1 - \frac{\mu_{K_{A,B}}(A_i \cap B_i)}{\mu_{K_{A,B}}(A_i \cup B_i)}\right),$$

where $\mu_{K_{A,B}}$ is the ambient measure. Equation (12) can be interpreted as a weighted sum of Jaccard distances between the $A_i(\mathbb{P})$ and $B_i(\mathbb{Q})$ sets. In the subsection 3.2, we will propose several schemes for setting the weights $\{w_i\}_{i=1}^{k-1}$.

Now we can define a distance between probability distributions $\mathbb{P}$ and $\mathbb{Q}$:

Definition 9. ***Distance for probability distributions.*** *Given two PMs* $\mathbb{P}$ *and* $\mathbb{Q}$*, and samples* $A = s_n(\mathbb{P})$*,* $B = s_m(\mathbb{Q})$*, we define:*

$$d(\mathbb{P}, \mathbb{Q}) = \lim_{(n,m,k)\to\infty} \sum_{i=1}^{k-1} w_i d_K(A_i, B_i), \tag{13}$$

where d_K is given in Definition 5 and $A_i(\mathbb{P})$ and $B_i(\mathbb{Q})$ for $i = 1, .., k-1$, are the partitions of the sets A and B. In practice, we will use $d_K(A, B)$ to estimate the distance between $\mathbb{P}$ and $\mathbb{Q}$.

3.2 Choice of weights for α-level set distances

Now is the time to fix some weighting schemes for the distances defined by eq. (12). Denote by $s_{\mathbb{P}}$ and $s_{\mathbb{Q}}$ the data samples corresponding to set of points/PMs $A(\mathbb{P})$ and $B(\mathbb{Q})$ respectively, and denote by $s_{\hat{A}_i(\mathbb{P})}$ and $s_{\hat{B}_i(\mathbb{Q})}$ the data samples that estimate $A_i(\mathbb{P})$ and $B_i(\mathbb{Q})$, respectively. We estimate the geometric structure of these sets using the following coverings: $\hat{A}_i(\mathbb{P}) = \cup_{x \in s_{\hat{A}_i(\mathbb{P})}} B(x, r_{\hat{A}_i(\mathbb{P})})$, and $\hat{B}_i(\mathbb{Q}) = \cup_{y \in s_{\hat{B}_i(\mathbb{Q})}} B(y, r_{\hat{B}_i(\mathbb{Q})})$, as you can see in Figure 4.

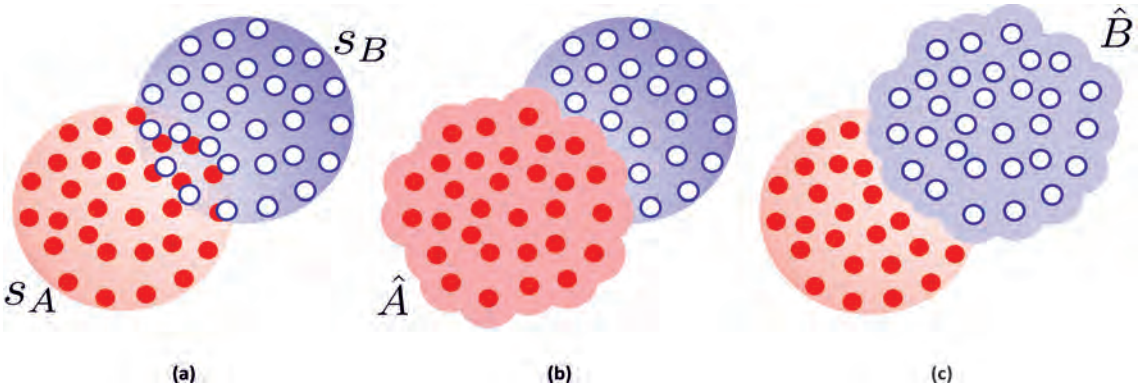

Figure 4: (a) Data samples s_A (dense) and s_B (empty), in this case uniformly distributed. (b) Covering $\hat{A}$. (c) Covering $\hat{B}$.

Let k denote the number of levels in partition $\boldsymbol{\alpha} = \{\alpha_{(i)}\}_1^k$. Denote by $n_{\hat{A}_i(\mathbb{P})}$ the number of data points in $s_{\hat{A}_i(\mathbb{P})}$, $n_{\hat{B}_i(\mathbb{Q})}$ the number of data points in $s_{\hat{B}_i(\mathbb{Q})}$, $r_{\hat{A}_i(\mathbb{P})}$ the (fixed) radius for the covering $\hat{A}_i(\mathbb{P})$ and $r_{\hat{B}_i(\mathbb{Q})}$ the (fixed) radius for the covering $\hat{B}_i(\mathbb{Q})$. We define the following three weighting schemas (see more details in [27, 28]):

Weighting Scheme 1 : Choose w_i in (12) by:

$$w_i = \frac{1}{k} \sum_{x \in s_{\hat{A}_i(\mathbb{P})}}^{n_{\hat{A}_i(\mathbb{P})}} \sum_{y \in s_{\hat{B}_i(\mathbb{Q})}}^{n_{\hat{B}_i(\mathbb{Q})}} \left(1 - I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{B}_i(\mathbb{Q})}}(x, y)\right) \frac{\| x - y \|_2}{(s_{\hat{B}_i(\mathbb{Q})} - \hat{A}_i(\mathbb{P})) \cup (s_{\hat{A}_i(\mathbb{P})} - \hat{B}_i(\mathbb{Q}))}. \tag{14}$$

Weighting Scheme 2 : Choose w_i in (12) by:

$$w_i = \frac{1}{k} \max_{x \in s_{\hat{A}_i(\mathbb{P})}, y \in s_{\hat{B}_i(\mathbb{Q})}} \left\{ (1 - I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{B}_i(\mathbb{Q})}}(x, y)) \parallel x - y \parallel_2 \right\}. \tag{15}$$

Weighting Scheme 3 : Choose w_i in (12) by:

$$w_i = \frac{1}{k} \hat{H} \left(s_{\hat{B}_i(\mathbb{Q})} - \hat{A}_i(\mathbb{P}), s_{\hat{A}_i(\mathbb{P})} - \hat{B}_i(\mathbb{Q}) \right), \tag{16}$$

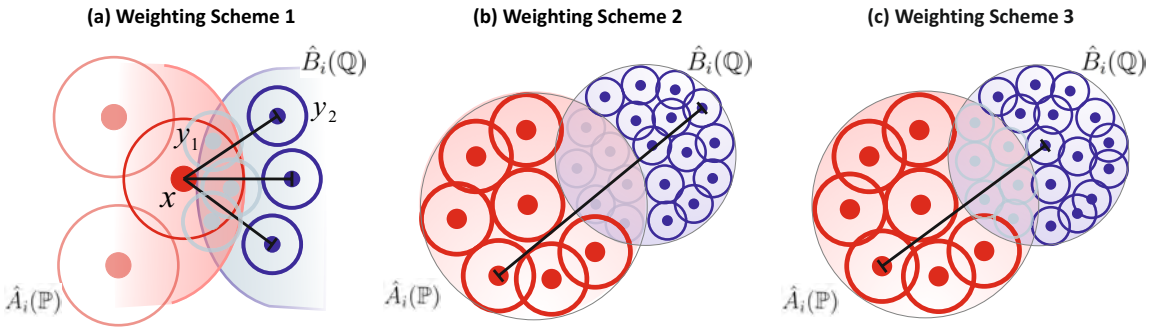

Figure 5: Calculation of weights in the distance defined by equation (12).

where $I_{r_A,r_B}(x,y) = 1$ when y belongs to the covering $\hat{A}$, x belongs to the covering $\hat{B}$ or both events happen. $\hat{H}(\hat{X},\hat{Y})$ denotes the Hausdorff distance (finite size version) between finite sets $\hat{X}$ and $\hat{Y}$ (which estimates the 'theoretical' Hausdorff distance between space regions X and Y). In this case $X = A_i(\mathbb{P}) - B_i(\mathbb{Q})$ and $Y = B_i(\mathbb{Q}) - A_i(\mathbb{P})$.

The intuition behind the three weighting schemes is illustrated in Figure 5. In weighting scheme 1 the weight w_i is a weighted average of distances between a point of $s_{\hat{A}_i(\mathbb{P})}$ and a point of $s_{\hat{B}_i(\mathbb{Q})}$ where $\|x-y\|_2$ is taken into account only when $I_{r_{\hat{A}_i(\mathbb{P})},r_{\hat{B}_i(\mathbb{Q})}}(x,y) = 0$. To illustrate this, consider $x \in s_{\hat{A}_i(\mathbb{P})}$ and $y_1, y_2 \in s_{\hat{B}_i(\mathbb{Q})}$ (Figure 5 (a)). The quantity $\|x-y_1\|_2$ does not contribute to calculation of the weight w_i because y_1 belongs to the (red) covering ball centered at x. That is, y_1 belongs to the cover estimation of $A_i(\mathbb{P})$ and therefore should not be taken into account for the calculation of the distance. On the other hand, $\|x-y_2\|_2$ contributes to the calculation of the weight w_i because y_2 does not belong to the (red) covering ball centered at x. In weighting scheme 2 w_i is proportional to the maximum distance between a point belonging to $\hat{A}_i(\mathbb{P})$ and a point belonging to $\hat{B}_i(\mathbb{Q})$, given that the covering balls centered at such points do not overlap. Figure 5 (c) illustrates the Hausdorff distance between the sets $s_{\hat{B}_i(\mathbb{Q})} - \hat{A}_i(\mathbb{P})$ and $s_{\hat{A}_i(\mathbb{P})} - \hat{B}_i(\mathbb{Q})$.

4 Experimental Work

Being the proposed distances intrinsically nonparametric, there are no simple parameters on which we can concentrate our attention to do exhaustive benchmarking. The strategy will be to compare the proposed distances to other classical PM distances for some well known (and parametrized) distributions and for real data problems. Here we consider distances belonging to the main types of PMs metrics: Kullback-Leibler (KL) divergence [6, 31] (f-divergence and also Bregman divergence), t-test (T) measure (Hotelling test in the multivariate case) and Energy distance [43, 38]. For further details on the sample versions of the above distance functions and their computational subtleties see [37, 8, 45, 31, 41, 17, 43] and references therein.

4.1 Artificial data

4.1.1 Discrimination between normal distributed sets of points

In this experiment we quantify the ability of the considered set/PM distances to discriminate between multivariate normal distributed sets of points. To this end, we generate a data sample of size $100d$ from a $N(\mathbf{0}, \mathbf{I}_d)$ where d stands for dimension and then we generate 1000 iid data samples of size $100d$ from the same $N(\mathbf{0}, \mathbf{I}_d)$ distribution. Next we calculate the distances between each of these 1000 iid data samples and the first data sample to obtain the 95% distance percentile.

Now define $\boldsymbol{\delta} = \delta\mathbf{1} = \delta(1, \ldots, 1) \in \mathbb{R}^d$ and increase δ by small amounts (starting from 0). For each $\boldsymbol{\delta}$ we generate a data sample of size $100d$ from a $N(\mathbf{0} + \boldsymbol{\delta}, \mathbf{I}_d)$ distribution. If the distance under consideration for the (displaced distribution) data sample to the original data sample is larger than the 95% percentile we conclude that the present distance is able to discriminate between both populations and this is the value δ^* referenced in Table 2. To make the process as independent as possible from randomness we repeat this process 100 times and fix δ^* to the present δ value if the distance is above the percentile in 90% of the cases. Thus we are calculating the minimal value δ^* required for each metric in order to discriminate between populations with a 95% confidence level

(type I error = 5%) and a 90% sensitivity level (type II error = 10%). In Table 2 we report the minimum distance ($\delta^*\sqrt{d}$) between distributions centers required to discriminate for each metric in several alternative dimensions, where small values implies better results. In the particular case of the T-distance for normal distributions we can use the Hotelling test to compute a p-value to fix the δ^* value.

Table 2: $\delta^*\sqrt{d}$ for a 5% type I and 10% type II errors.

Metric	d:	1	2	3	4	5	10	15	20	50	100
KL		.870	.636	.433	.430	.402	.474	.542	.536	.495	.470
T		.490	.297	.286	.256	.246	.231	.201	.212	.193	.110
Energy		.460	.283	.284	.250	.257	.234	.213	.223	.198	.141
WLS(1)		.490	.354	.277	.220	.224	.221	.174	.178	.134	.106
WLS(2)		.450	.283	.268	.240	.229	.231	.232	.223	.212	.134
WLS(3)		.490	.424	.329	.300	.291	.237	.240	.225	.219	.141

The data chosen for this experiment are ideal for the use of the T statistics that, in fact, outperforms KL (results in Table 2). However, Energy distance works even better than T distance in dimensions 1 to 4 and WLS(1) performs similarly (slightly better) to T (except for dimension 2) in dimensions upon 3.

In a second experiment we consider again normal populations but different variance-covariance matrices. Define as an expansion factor $\sigma \in \mathbb{R}$ and increase σ by small amounts (starting from 0) in order to determine the smallest σ^* required for each metric in order to discriminate between the $100d$ sampled data points generated for the two distributions: $N(\mathbf{0}, \mathbf{I}_d)$ and $N(\mathbf{0}, (1+\sigma)\mathbf{I}_d)$. If the distance under consideration for the (displaced distribution) data sample to the original data sample is larger than the 95% percentile we conclude that the present distance is able to discriminate between both populations and this is the value $(1+\sigma^*)$ reported in Table 3. To make the process as independent as possible from randomness we repeat this process 100 times and fix σ^* to the present σ value if the distance is above the 90% percentile of the cases, as it was done in the previous experiment.

Table 3: $(1 + \sigma^*)$ for a 5% type I and 10% type II errors.

Metric	dim:	1	2	3	4	5	10	15	20	50	100
KL		3.000	1.700	1.250	1.180	1.175	1.075	1.055	1.045	1.030	1.014
T		–	–	–	–	–	–	–	–	–	–
Energy		1.900	1.600	1.450	1.320	1.300	1.160	1.150	1.110	1.090	1.030
WLS(1)		1.700	1.350	1.150	1.120	1.080	1.050	1.033	1.025	1.015	1.009
WLS(2)		1.800	1.450	1.220	1.200	1.180	1.080	1.072	1.070	1.045	1.020
WLS(3)		1.900	1.450	1.300	1.280	1.380	1.118	1.165	1.140	1.090	1.032

We can see here again that the proposed metric WLS(1) is better than the competitors in all dimensions considered, begin the WLS(2) the second best in performance. There are no entries in Table 3 for the T distance because it was not able to distinguish between the considered populations in none of the considered dimensions.

4.2 Real case-studies

4.2.1 Text Mining

For the first real data example, we consider a collection of 1774 documents (corresponding to 13 topics) extracted from three bibliographic data bases (LISA, INSPEC and Sociological Abstracts). We present a brief summary list of topics considered in each data base:

```
LISA
---------------------------
business archives in de     137
lotka's law  90
biology in de 280
automatic abstracting 69
-----------------------------
INSPEC:
-------------------------------
Self organizing maps 83
```

```
dimensionality reduction 75

power semiconductor devices 170

optical cables 214

feature selection 236

---------------------------------

SOCIOLOGICAL ABSTRACTS:

---------------------------------

Intelligence tests 149

Retirement communities 74

Sociology of literature and discourse 106

Rural areas and rural poverty 91
```

Each document is converted into a vector into the Latent Semantic Space using the Singular Value Decomposition, and the documents corresponding to one topic are considered as a sample from the underlying distribution that generates the topic. Next we calculate the 13×13 distance matrix and perform MDS, not on the individual documents, but on the document sets. The result is shown in Figures 6 and 7, where we can see that close groups correspond to close (in a semantic sense) topics, that indicates the distance is working properly in a nonparametric setting in high dimension.

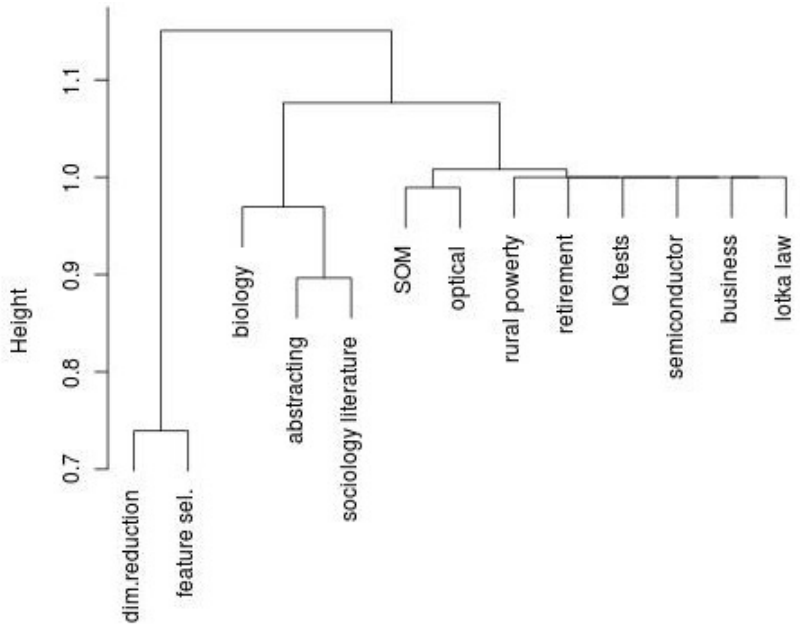

Figure 6: Dendrogram for the 13×13 document data set distance.

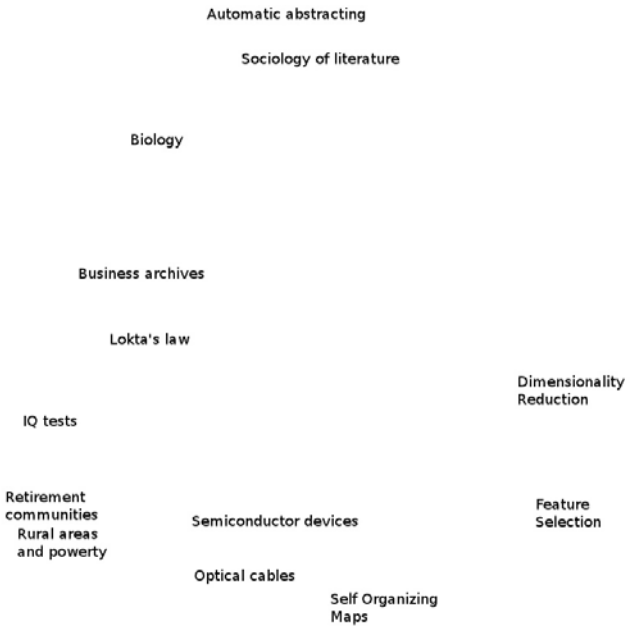

Figure 7: Multidimensional Scaling of the 13 groups of documents.

4.2.2 Shape Classification

As an application of the preceding theory to the field of pattern recognition we consider a cases of shape classification problem. For this we consider the Tree Leaf Database [19]. We represent each leaf by a cloud of points in $\mathbb{R}^2$, as an example of the treatment given to a leaf consider the Figure 8.

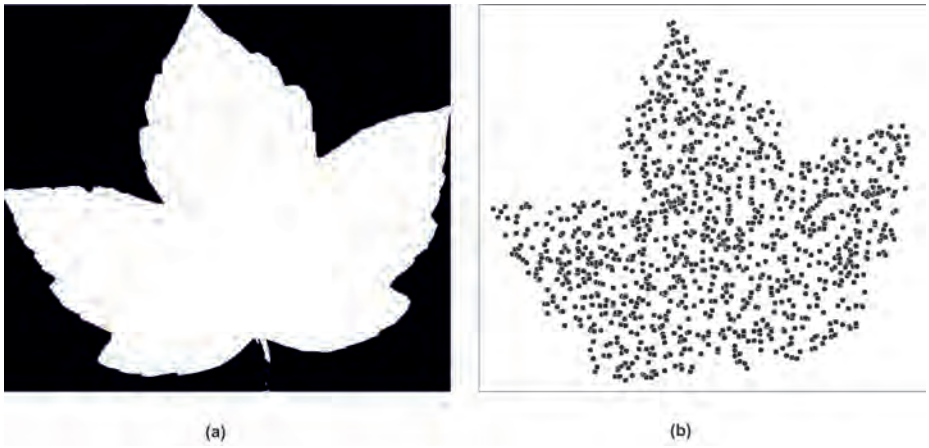

Figure 8: Real image and sampled image of a leaf in the Tree Leaf Database.

For each image i of size $N_i \times M_i$ we generate a uniform sample of size $N_i M_i$ and retain only those points which fall into the white region (image body) whose intensity gray level are larger than a fixed threshold (.99). This yield a representation of the leaf with around one thousand and two thousand points depending on the image. After rescaling and centering, we computed the 10×10 distance matrix (using the WLS(2) distance and the Energy distance in this case) and the MDS plot in Figure 9. It is clear that the WLS distance is able to better account for differences in shapes.

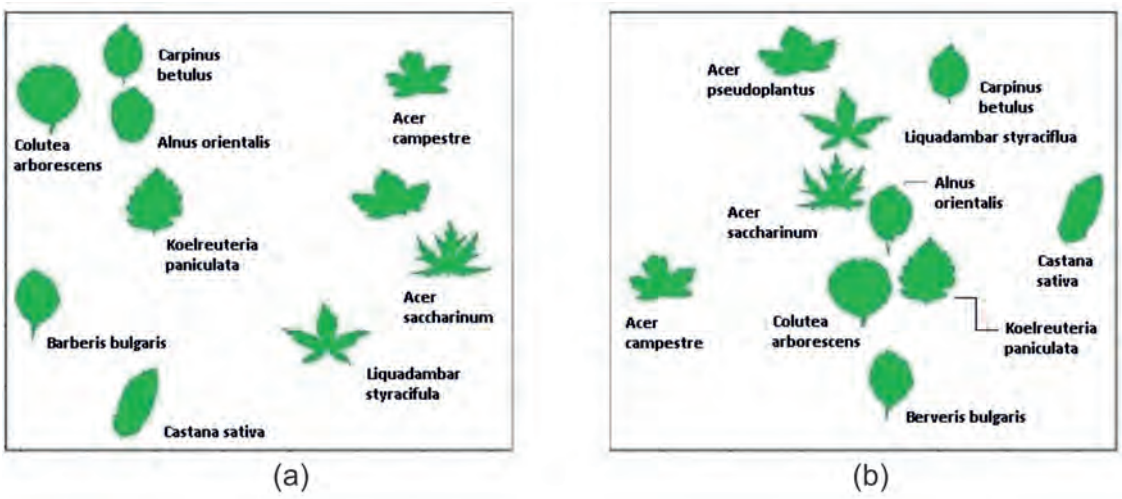

Figure 9: Multi Dimensional Scaling representation for leaf database based on WLS(2) (a) and Energy distance (b).

5 Conclusions

In this paper we afford the task of defining distance functions for data sets (and distributions) in a functional data analysis context. To this aim we introduce kernels that provide a metrization of $\mathcal{P}(X)$, which allows the use of the natural metric induced by the kernel function. An interesting point is that kernels used in this work take into account the underlying (data) generating probability distributions. The estimation of these set/PM kernel distances, does not require the use of either parametric assumptions or explicit probability function parameter estimations, which makes a clear advantage over most well established PM distances, such as the families of Bregman divergences, f-divergences and Integral Probability Metrics.

A battery of real and simulate examples have been used to study the performance of the new distance. In particular in the case of normally distributed data, the generated data sets are ideal for the use of T-statistics, but the proposed distances show superior discrimination power. Regarding the practical applications, the new PM distances have been proven to be very competitive in shape

recognition and text mining problems.

Future Work: Given a positive definite function $K : \mathcal{P}(X) \times \mathcal{P}(X) \to [0, 1]$, as it is defined in Equation 5, by Mercer's theorem there exists an Euclidean space $\mathcal{H}$ and a lifting map $\Phi : \mathcal{P}(X) \to \mathcal{H}$ such that $K(A, B) = \langle \Phi(A), \Phi(B) \rangle$ with $A,\ B \in \mathcal{P}(X)$ [23, 34, 40]. The study of the lifting map $\Phi : \mathcal{P}(X) \to \mathcal{H}$ is the object of our immediate research, in order to understand the geometry induced by the proposed metric and the asymptotic properties of the developed distances.

Acknowledgments This work was partially supported by projects **DGUCM 2008/00058/002**, **MEC 2007/04438/001** and **MIC 2012/00084/00**.

References

[1] C. Ahlbrandt, G. Benson G and W.Casey. *Minimal entropy probability paths between genome families.* J Math Biol. 2004 May;48(5):563-90.

[2] S.-I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen and C. R. Rao. *Differential Geometry in Statistical Inference.* Lecture Notes-Monograph Series, Vol 10, 1987.

[3] S. Amari, H. Nagaoka *Methods of Information Geometry.* American Mathematical Society. 2007.

[4] C. Atkinson and A. F. S. Mitchell. *Rao's Distance Measure.* The Indian Journal of Statistics, Series A. Vol 43, pp 345-365, 1981.

[5] A. Banerjee, S. Merugu, I. Dhillon, J. Ghosh. *Clustering whit Bregman Divergences.* Journal of Machine Learning Research, pp 1705:1749, 2005.

[6] S. Boltz, E. Debreuve and M. Barlaud. *High-dimensional statistical measure for region-of-interest tracking.* Transactions in Image Processing, vol. 18, no. 6, pp 1266:1283, 2009.

[7] J. Burbea and C. R. Rao. *Entropy differential metric, distance and divergence measures in probability spaces: A unified approach.* Journal of Multivariate Analysis, Vol 12, pp. 575-596, 1982.

[8] S.H Cha. *Comprehensive survey on distance/similarity measures between probability density functions.* International Journal of Mathematical Models and Methods in Applied Sciences, vol. 1(4), pp.300-307, 2007.

[9] F. Chazal, D. Cohen-Steiner, and Q. Mrigot. *Geometric Inference for Probability Measures.* Journal on Foundations of Computational Mathematics, 11(6):733-751, 2011.

[10] A. Cichocki and S. Amari. *Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities.* Entropy, 12, 1532-1568, 2010.

[11] I. Csiszár and P. Shields. *Information Theory and Statistics: A Tutorial.* Foundations and Trends in Communications and Information Theory, 2004.

[12] M.M. Deza and E. Deza. *Enciclopedia of Distances.* Springer, 2009.

[13] I.L. Dryden, A. Koloydenko and D. Zhou. *Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging.* The Annals of Applied Statistics, vol. 3, pp. 1102-1123, 2009.

[14] I.L. Dryden, A. Koloydenko and D. Zhou. *The Earth Mover's Distance as a Metric for Image Retrieval.* International Journal of Computer Vision, Vol. 40, pp. 99-121, 2000.

[15] B.A. Frigyik, S. Srivastava and M. R. Gupta. *Functional Bregman Divergences and Bayesian Estimation of Distributions.* IEEE Transactions on Information Theory 54 (11): 51305139, 2008.

[16] A. L. Gibbs and F. E. Su. *On Choosing and Bounding Probability Metrics.* Journal of International Statistical Review, 2002.

[17] M. N. Goria, N. N. Leonenko, V. V. Mergel and P. L. Novi Inverardi. *A new class of random vector entropy estimators and its applications in testing statistical hypotheses.* Journal of Nonparametric Statistics, vol. 13, No. 3, pp. 277-297, 2005.

[18] A. Gretton, K. Borgwardt, M. Rasch, B. Schlkopf, A. Smola. *A kernel method for the two sample problem.* Advances in Neural Information Processing Systems, pp. 513-520, 2007.

[19] Institute of Information Theory and Automation ASCR. *LEAF - Tree Leaf Database.* Prague, Czech Republic, http://zoi.utia.cas.cz/tree_leaves.

[20] S. Joshi, R. Kommaraju, J. Phillips and S. Venkatasubramaian. *Comparing distributions and shapes using the kernel distance.* arXiv:1001.0591v2, 2011.

[21] G. Lebanon. *Metric Learning for Text Documents.* IEEE Trans on Pattern Analysis and Machine Intelligence, 28:4, 497-508,2006.

[22] S. Mallat. *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 11:7, pp. 674-693.

[23] G. Martos and A. Muñoz. *FDA, RKHS and Information Geometry Methods for the Analysis of Time Series and Probability Distributions.* Thesis Proposal, Universidad Carlos III de Madrid, Department of Statistics, 2010.

[24] P. Marriot and M. Salmon. *Aplication of Differential Geometry to Econometrics.* Cambridge University Press, 2000.

[25] Y. Moon, B. Rajagopalan and U. Lall. *Estimation of mutual information using kernel density estimators.* Physical Review E, vol. 52, 3, pp. 2318-2321.

[26] A. Müller. *Integral Probability Metrics and Their Generating Classes of Functions.* Advances in Applied Probability, Vol. 29, No. 2, pp. 429-443. 1997.

[27] A. Muñoz, G. Martos, J. Arriero and J. Gonzalez. *A new distance for probability measures based on the estimation of level sets.* Artificial Neural Networks and Machine Learning (ICANN). Springer Berlin Heidelberg, p. 271-278, 2012.

[28] A. Muñoz, G. Martos and J. Gonzalez. *Nonparametric Distances for Ensembles and Statistical Distributions with Applications.* Journal of Computational and Graphical Statistics (submitted), 2012.

[29] A. Muñoz and J.M. Moguerza. *Estimation of High-Density Regions using One-Class Neighbor Machines.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 28:3, pp. 476-480.

[30] A. Muñoz and J.M. Moguerza. *A Naive Solution to the One-Class Problem and Its Extension to Kernel Methods.* LNCS 3773, pp. 193204, 2005.

[31] X. Nguyen, M. J. Wainwright and M. I. Jordan. *Nonparametric Estimatimation of the Likelihood and Divergence Functionals.* IEEE International Symposium on Information Theory, 2007.

[32] E. Otey and S. Parthasarathy. *A dissimilarity measure for comparing subsets of data: application to multivariate time series.* Fifth IEEE International Conference on Data Mining, 2005, pp. 101112.

[33] X. Pennec. *Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements.* Journal of Mathematical Imaging and Vision, vol. 25, pp. 127-154, 2006.

[34] J. Phillips and S. Venkatasubramanian. *A gentle introduction to the kernel distance.* arXiv preprint, arXiv:1103.1625, 2011.

[35] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis.* New York: Springer. 2005.

[36] Y. Rubner, C. Tomasi and L.J. Guibas. *A Metric for Distributions with Applications to Image Databases.* Sixth IEEE Conf. on Computer Vision, pp.59-66, 1998.

[37] D. Scott. *Multivariate Density Estimation: Theory Practice and Visualization.* Wiley, 1992.

[38] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu. *Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing.* arXiv, 2012.

[39] L. Shamir. *Automatic morphological classification of galaxy images.* Mon Not R Astron Soc. 2009 November 1; 399(3): 13671372.

[40] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Scholkopf. *Hilbert Space Embeddings and Metrics on Probability Measures.* Journal of Machine Learning Research, pp. 1297-1322, 2010.

[41] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Scholkopf and G. R. G. Lanckriet. *Non-parametric estimation of integral probability metrics.* International Symposium on Information Theory, 2010.

[42] R.S. Strichartz. *A Guide to Distribution Theory and Fourier Transforms.* World Scientific, 1994.

[43] G.J. Székely, M.L. Rizzo. *Testing for Equal Distributions in High Dimension.* InterStat, 2004.

[44] A. Ullah. *Entropy, divergence and distance measures with econometric applications.* Journal of Statistical Planning and Inference, 49, 137-162, 1996.

[45] Q. Wang, S. R. Kulkarni and S. Verdú. *Divergence Estimation of Continuoes Distribution Based on Data-dependent Partitions.* IEEE Trans. Information Theory, vol. 51,no. 9, pp. 3064-3074, 2005.

[46] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell. *Distance Metric Learning, with Application to Clustering with Side-information.* Advances in Neural Information Processing Systems, pp 505:512, 2002.

[47] D. Zhou, J. Li and H. Zha. *A New Mallows Distance Based Metric For Comparing Clusterings.* ICML '05 Procs. of the 22nd international conference on Machine learning.

[48] V. M. Zolotarev. *Probability metrics.* Teor. Veroyatnost. i Primenen,28:2, 264287, 1983.