A New Distance for Data Sets in a Reproducing Kernel Hilbert Space Context

Alberto Muñoz[1], Gabriel Martos[1], and Javier González[2]

[1] University Carlos III, Department of Statistics, Madrid, Spain
[2] J. Bernoulli Institute for Mathematics and Computer Science, University of Groningen, The Netherlands
{alberto.munoz,gabrielalejandro.martos}@uc3m.es, j.gonzalez.hernandez@rug.nl

Abstract. In this paper we define distance functions for data sets in a reproduncing kernel Hilbert space (RKHS) context. To this aim we introduce kernels for data sets that provide a metrization of the power set. The proposed distances take into account the underlying generating probability distributions. In particular, we propose kernel distances that rely on the estimation of density level sets of the underlying data distributions, and that can be extended from data sets to probability measures. The performance of the proposed distances is tested on several simulated and real data sets.

1 Introduction

The study of distances between data sets lies at the core of many methods of analysis in image processing [9], genetics [1], time series [7], etc. In this paper we define distances between data sets that take 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. These problems can be solved by choosing an appropriate distance between PM *e.g.* the χ^2 or L_1 distance. Other examples of distances between PM can also be founded in Clustering, Image Analysis, Time Series Analysis, etc. For a review of interesting distances between probability distributions and theoretical results, see for instance [4]. In many practical situations the size of the available sample is small, and the use of purely non parametric estimators often results in a poor performance. 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 distribution functions. 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 therefore to a distance) for data sets. This paper is organized as follows: In Section 2 we introduce kernel functions for data sets with uniform distributions. Section 3 introduces a new

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 222–229, 2013.

metric for general data sets based on the estimation of density level sets. Section 4 shows the performance of the proposed metric on simulated and real data sets.

2 A Kernel for Data Sets with Reference to a Distribution

Consider a measure space $(\mathcal{X}, \mathcal{F}, \mu)$, where $\mathcal{X}$ is the sample space (a compact set of a real vector space in this work), $\mathcal{F}$ is a σ-algebra of measurable subsets of $\mathcal{X}$ and $\mu : \mathcal{F} \to \mathbb{R}^+$ is the ambient σ-finite measure, the Lebesgue 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 $f_{\mathbb{P}} : \mathcal{X} \to \mathbb{R}^+$ (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. In Section 3 we will discuss the corresponding distributional distance measures. Consider two *iid* samples $A = s_n(\mathbb{P}) = \{x_i\}_{i=1}^n \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of all subsets of $\mathcal{X}$ including the empty set and itself (the power set of $\mathcal{X}$), and $B = s_m(\mathbb{Q}) = \{y_j\}_{j=1}^m \in \mathcal{P}(\mathcal{X})$, 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 $z \in \mathcal{X}$ 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 $y \in \mathcal{X}$. We say that y is **indistinguishable** from x with 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 consider the following smooth indicator functions.

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

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

Of course, other definitions of $f_{x,r,\gamma}$ are possible. 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 $B_r(x)$, depending on the shape parameter γ (see Fig. 1).

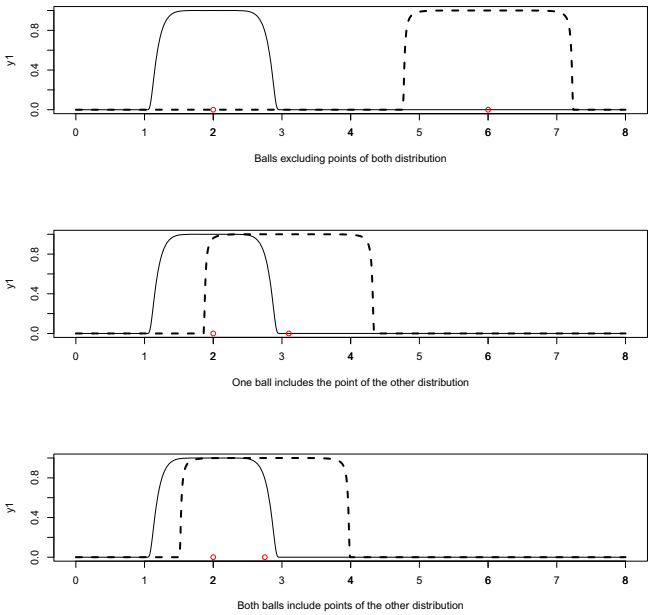

Fig. 1. Illustration of the $\overset{A(\mathbb{P})}{=}$ and $\overset{B(\mathbb{Q})}{=}$ relationship using smooth indicator functions

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

$$K_{A,B}(x,y) = f_{x,r_A,\gamma}(y) + f_{y,r_B,\gamma}(x) - f_{x,r_A,\gamma}(y) f_{y,r_B,\gamma}(x), \tag{2}$$

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 $d(x,y) > r_B$ (Fig. 1, top) 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 $K_{A,B}(x,y) = 1$. 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 $K_{A,B}(x,y) = 1$ (Fig. 1, center). 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. 1, bottom).

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

$$K(C,D) = \sum_{x \in C} \sum_{y \in D} K_{A,B}(x,y). \tag{3}$$

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':

$\mu_{K_{A,B}}(A \cap B) = K(A,B)$. 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.

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 $\mathcal{P}(\mathcal{X})$:*

$$d_K(C,D) = 1 - \frac{K(C,D)}{N}, \tag{4}$$

where $N = n_C + n_D = \#(C \cup D)$ and represent the measure of the set $C \cup D$.

It is straightforward to check that $d_K(C,D)$ is a semi-metric (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 the 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 $\lim_{n,m \to \infty} d_K(A,B) = 1 - \frac{\mu(A \cap B)}{\mu(A \cup B)}$, that is the Jaccard distance for data sets.

3 A Metric for Data Sets Based on Estimation of Level Sets

Using constant radii in Eq. (1) 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.

Consider the α-level sets defined by $S_\alpha(f_\mathbb{P}) = \{x \in X | \, f_\mathbb{P}(x) \geq \alpha\}$, where $P(S_\alpha(f_\mathbb{P})) = 1 - \nu$, where $f_\mathbb{P}$ is the density function and $0 < \nu < 1$. If we consider an ordered sequence $\alpha_1 < \ldots < \alpha_k$, 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 $\alpha_k \geq \min\{\max_{x \in X} f_\mathbb{P}(x), \max_{x \in X} f_\mathbb{Q}(x)\}$; then $\bigcup_i A_i(\mathbb{P}) \simeq Supp(\mathbb{P}) = \{x \in X | \, f_\mathbb{P}(x) \neq 0\}$. Note that given the definition of the A_i, if $A_i(\mathbb{P}) = B_i(\mathbb{Q})$ for every i when $(n,m,k) \to \infty$, 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. To estimate level sets $\hat{S}_{\alpha_i}(f_\mathbb{P})$ from a data sample in this work we use the algorithm introduced in [5]. Next 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 6. ***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*

$\alpha_1 < \alpha_2 < \ldots < \alpha_k$, $\alpha_i \in (0, \min\{\max_{x\in X} f_{\mathbb{P}}(x), \max_{x\in X} f_{\mathbb{Q}}(x)\})$. *Then we define the weighted α-level set distances between the sets A and B by*

$$d(A,B) = \sum_{i=1}^{k-1} w_i d_K(A_i, B_i), \quad w_1, \ldots, w_{k-1} \in \mathbb{R}^+ \tag{5}$$

In the practice to compute the distance in Eq. (5), we have to use: $\hat{A}_i(\mathbb{P}) = \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}}) - \hat{S}_{\alpha_i}(f_{\mathbb{P}})$ the estimation of $A_i = S_{\alpha_{i+1}}(f_{\mathbb{P}}) - S_{\alpha_i}(f_{\mathbb{P}})$ based on set A; and the respective estimation for $\hat{B}_i(\mathbb{Q})$. In this paper we choose the weights by

$$w_i = \frac{1}{k} \sum_{x\in s_{\hat{A}_i(\mathbb{P})}}^{n_{\hat{A}_i(\mathbb{P})}} \sum_{y\in s_{\hat{B}_i(\mathbb{Q})}}^{n_{\hat{B}_i(\mathbb{Q})}} \frac{\left(1 - I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{B}_i(\mathbb{Q})}}(x,y)\right) \parallel x - y \parallel_2}{(s_{\hat{B}_i(\mathbb{Q})} - \hat{A}_i(\mathbb{P})) \cup (s_{\hat{A}_i(\mathbb{P})} - \hat{B}_i(\mathbb{Q}))}, \tag{6}$$

where $s_{\mathbb{P}}$ and $s_{\mathbb{Q}}$ are the data samples corresponding to set of points/PMs $A(\mathbb{P})$ and $B(\mathbb{Q})$ respectively, $s_{\hat{A}_i(\mathbb{P})}$ and $s_{\hat{B}_i(\mathbb{Q})}$ denote the data samples that estimate $A_i(\mathbb{P})$ and $B_i(\mathbb{Q})$, respectively. $\hat{A}_i(\mathbb{P}) = \cup_{x\in s_{\hat{A}_i(\mathbb{P})}} B(x, r_{\hat{A}_i(\mathbb{P})})$, and $\hat{B}_i(\mathbb{Q}) = \cup_{y\in s_{\hat{B}_i(\mathbb{Q})}} B(y, r_{\hat{B}_i(\mathbb{Q})})$ are the covering estimations of the sets $A_i(\mathbb{P})$ and $B_i(\mathbb{Q})$ respectively, and $I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{B}_i(\mathbb{Q})}}(x,y)$ is an indicator function that takes value 1 when y belongs to the covering estimation of the set $A_i(\mathbb{P})$, x belongs to the covering estimation of the set $B_i(\mathbb{Q})$ or both events happen, and value 0 otherwise. Note that 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 $I_{r_{\hat{A}_i(\mathbb{P})}, r_{\hat{B}_i(\mathbb{Q})}}(x,y) = 0$. Other definitions of w_i are possible and give rise to different distance measures.

4 Experimental Work

The proposed distances are intrinsically non-parametric, so no tuning parameters have to be fixed or evaluated via simulation. The strategy to test the Weighted level-set distance (WLS) will be to compare it to other classical PM distances for some well known (and parametrized) distributions and for real data problems. We consider distances belonging to the main types of PMs metrics: Kullback-Leibler (KL) divergence [6] (f-divergence and also Bregman divergence), t-test (T) measure (Hotelling test in the multivariate case) and Energy distance [12]. For further details on the sample versions of the above distance functions and their computational subtleties see [6,11,12].

4.1 Synthetically Generated Data

Case I: Discrimination between Gaussian distributed sets of points with equal covariance structure

We quantify the ability of the considered set/PM distances to discriminate between multivariate normal distributed sets of points. To this end, we generate

Table 1. $\delta^* \sqrt{d}$ 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		.490	.354	.277	.220	.224	.221	.174	.178	.134	.106

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 1. To take into account the randomness in the experiment 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 calculate 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 1 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 case of the T-distance for normal distributions we can use the Hotelling test to compute a p-value to fix the δ^* value.

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

Case II: Discrimination between Gaussian distributed sets of points with different covariance structure

In this experiment we consider again normal populations but different variance-covariance matrices. Define as expansion factor $\sigma \in \mathbb{R}$, which we gradually increase (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 2. To take into account the randomness in the experiment we repeat it 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 2. $(1 + \sigma^*)$ 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		1.700	1.350	1.150	1.120	1.080	1.050	1.033	1.025	1.015	1.009

(a) (b)

Fig. 2. Real image and sampled image of a leaf in the Tree Leaf Database

We can see here again that the proposed metric WLS is better than the competitors in all dimensions considered. There are no entries in Table 2 for the T distance because it was not able to distinguish between the considered populations in the considered dimensions.

4.2 Real Case-Study: Shape Classification

As an application of the preceding theory to the field of pattern recognition we consider a problem of shape classification, using the Tree Leaf Database [3]. 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 Fig. 2. For each image i of size $N_i \times M_i$, we generate a sample of size $N_i \times M_i$ from a uniform distribution 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 distance and the Energy distance in this case) and the Multidimensional Scaling (MDS) plot in Fig. 3. It is clear that the WLS distance is able to better account for differences in shapes.

Future Work: Given a positive definite function $K : \mathcal{P}(X) \times \mathcal{P}(X) \to [0, 1]$, as it is defined in Eq. (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 $A, B \in \mathcal{P}(X)$ [8,10]. 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.

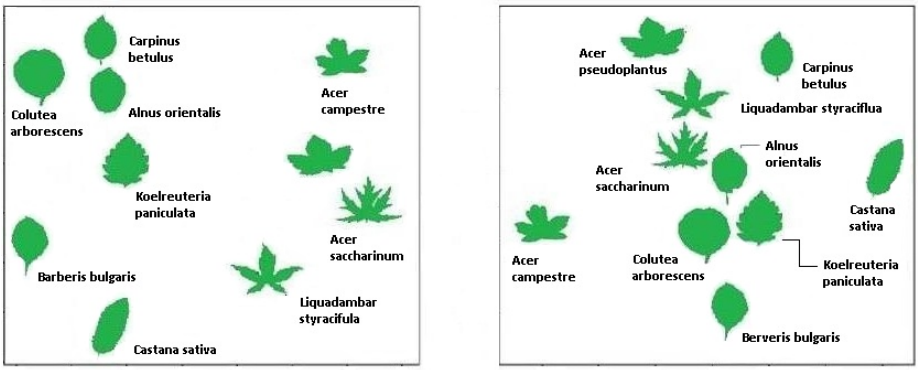

Fig. 3. MDS representation for leaf database based on WLS (a); Energy distance (b)

Acknowledgments. This work was partially supported by projects **MIC 2012/ 00084/00**, **ECO2012-38442**, **DGUCM 2008/00058/002** and **MEC 2007/04438/001**.

References

1. Ahlbrandt, C., Benson, G., Casey, W.: Minimal entropy probability paths between genome families. J. Math. Biol. 48(5), 563–590 (2004)
2. Dryden, I.L., Koloydenko, A., Zhou, D.: The Earth Mover's Distance as a Metric for Image Retrieval. Internat. Journal of Comp. Vision 40, 99–121 (2000)
3. Institute of Information Theory and Automation ASCR. LEAF - Tree Leaf Database. Prague, Czech Republic, `http://zoi.utia.cas.cz/tree_leaves`
4. Müller, A.: Integral Probability Metrics and Their Generating Classes of Functions. Advances in Applied Probability 29(2), 429–443 (1997)
5. Muñoz, A., Moguerza, J.M.: Estimation of High-Density Regions using One-Class Neighbor Machines. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(3), 476–480 (2006)
6. Nguyen, X., Wainwright, M.J., Jordan, M.I.: Nonparametric Estimatimation of the Likelihood and Divergence Functionals. In: IEEE International Symposium on Information Theory (2007)
7. Otey, E., Parthasarathy, S.: A dissimilarity measure for comparing subsets of data: application to multivariate time series. In: Fifth IEEE International Conference on Data Mining, pp. 101–112 (2005)
8. Phillips, J., Venkatasubramanian, S.: A gentle introduction to the kernel distance. arXiv preprint, arXiv:1103.1625 (2011)
9. Rubner, Y., Tomasi, C., Guibas, L.J.: A Metric for Distributions with Applications to Image Databases. In: Sixth IEEE Conf. on Computer Vision, pp. 59–66 (1998)
10. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Scholkopf, B.: Hilbert Space Embeddings and Metrics on Probability Measures. Journal of Machine Learning Research, 1297–1322 (2010)
11. Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Scholkopf, B., Lanckriet, G.R.G.: Non-parametric estimation of integral probability metrics. In: International Symposium on Information Theory (2010)
12. Székely, G.J., Rizzo, M.L.: Testing for Equal Distributions in High Dimension. InterStat (2004)