THE ALPHA-BETA-SYMMETRIC DIVERGENCE AND THEIR POSITIVE DEFINITE KERNELS

MACTAR NDAW, MACOUMBA NDOUR
and PAPA NGOM

LMA-Laboratoire de Mathématiques Appliquées
Université Cheikh Anta Diop
BP 5005 Dakar-Fann
Sénégal
e-mail: mactarndaw1@gmail.com
        macoumbandour@hotmail.fr
        papa.ngom@ucad.edu.sn

Abstract

In the field of statistical modelling, the distance or divergence measure is a criterion widely known and widely used tool for theoretical and applied statistical inference and data processing problems. In this paper, we deal with the well-known Alpha-Beta-divergences (which we shall refer to as the AB-divergences), which are a family of cost functions parametrized by two hyperparameters and their tight connections with the notions of Hilbertian metrics and positive definite (pd) kernels on probability measures. An attempt is made to describe this dissimilarity measure, which can be symmetrized using its two tuning parameters, alpha and beta. We compute the degree of symmetry of the AB-divergence on the basis of Hilbertian metrics. We investigate the desirable properties that the proposed approach needs to build a positive definite kernel $\mathcal{K}(x, y)$ corresponding to this symmetric AB-divergence.

We establish the effectiveness of our approach with experiments conducted on Support Vector Machine (SVM) and the applicability of this method is described in an algorithm from this symmetric divergence in image classification.

We perform experiments using the conditionally defined positive $\mathcal{K}$ and the kernel transformed $\mathcal{K}_t$ and show that these kernels have the same proportion of errors for the Euclidian divergence and the Hellinger divergence. We also observe large reductions in error for the Itakura-Saito divergence with the $\mathcal{K}_t$ kernel in classifications than classical kernel methods.

1. Introduction

Over the last few years, the need for specific design of kernels for a given data structure has been recognized by the kernel community. Recently, a Hilbert space embedding for probability measures has been proposed, with applications including dimensionality reduction, independence testing, and machine learning. Therefore, the use of specialized metrics and divergences measures in the successful design of dimensionality reduction techniques has been progressively acquiring much recognition. There are numerous real scenarios and applications for which the parameters of interest belong to non-flat manifolds, and where the Euclidian geometry results are unsuitable to evaluate the similarities. Indeed, this is usual case in the comparison of probability density functions. So, for better results the kernel should be adjusted as accurately as possible to the subjacent structure of the input space. Kernel on probability measures are very handy for dealing with graph problems, trees, manifolds, acoustic and signal processing and they became very popular because of their many applications. In probability theory, the distance between probability measures is used in studying pattern analysis, see John Shawe Taylor and Nello Cristianini [20]. Another application is in giving a bounded probability space $\mathcal{X}$ and using the kernel to compare arbitrary sets in that space, by putting, e.g., the uniform measure on each set. This is extremely useful to compare data of variable length, sequence data in bioinformatics, for example, kernel methods for predicting protein-protein in interactions (Asa Ben-Hur and Willian Stafford Noble [3]).

The 20th century witnessed tremendous efforts to exploit new distance/similarity measures for a variety of applications. There are a substantial number of distance/similarity measures encountered in many different fields, such as biology, chemistry, computer science, ecology, information theory, geology, mathematics, physics, statistics, etc. Distance or similarity measures are essential to solve many pattern recognition problems, such as classification, clustering, and retrieval problems. Various distance/similarity measures that are applicable to compare two probability density functions. The advantages of discriminative learning algorithms and kernel machines are combined with generative modelling using a novel kernel between distributions. In the probability product kernel, data points in the input space are mapped to distributions over the sample space and a general inner product is then evaluated as the integral of the product of pairs of distributions. Recently, developments in machine learning, including the emergence of support vector machines, have rekindled interest in kernel methods (Vapnik [29]; Hastie et al. [13]) and take full advantage of well known probabilistic models. These kernel methods have been widely employed to solve machine learning problems such as classification and clustering. Although there are many existing graph kernel methods for comparing patterns represented by undirected graphs, the corresponding methods for directed structures are less developed. In particular, for domains such as speech and images kernel functions have been suggested as good ways to combine an underlying generative model in the feature space and discriminant classifiers such as SVM.

Of particular concern to mathematicians is that several divergence measures are asymmetric.

However, in support vector machine classifier, asymmetric kernel functions are not used so far, although they are frequently used in other kernel classifies.

In this paper, we suggest an alternative procedure by exploiting the symmetric AB-divergence measures, and present an information theoretic kernel method for assessing the similarity between a pair of directed graphs.

In particular, we show that our kernel method provides an efficient tool in statistical learning theory, and SVM have demonstrated highly competitive performance in numerous real-world applications, such as medical diagnosis, bioinformatics, face recognition, image processing, and text mining, which has established SVM as one of the most popular, state-of-the-art tools for knowledge discovery and data mining. Similar to artificial neural networks, SVM possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy.

The remainder of this paper is organized as follows. In Section 2, we set out the basic notations, the definitions and assumptions. We will show the close relationship between Hilbert metrics and pd kernels so that in general, statements for one category can be easily transferred to another. In Section 3, we define a Hilbertian metrics the Alpha-Beta-Symmetric divergence (ABS-divergence) and the property of this divergence are studied and we are given a corresponding positive definite kernels. In Section 4, the results of the simulations are presented. Therefore, we evaluated the performance of the proposed metrics and kernels in tree classifications. And we proposed and apply an algorithm in experimental dataset to analysis the robustness of the divergence proposed. In the last section we presented the conclusion.

2. Basic Notation and Some Results

For a class of Hilbertian metrics, that are metrics which can be isometrically embedded into a Hilbert space. We will also use the following function class to define this subclass of metrics.

2.1. Hilbertian metrics, positive definite kernels

The positive definite kernel $\mathcal{K}(x, y)$ corresponds to an inner product $\langle \phi_x, \phi_y \rangle_{\mathcal{H}}$ in some feature space $\mathcal{H}$. The class of conditionally positive definite (cpd) kernel is less well known. Nevertheless this class is of great interest since Schölkopf show in (Moreno et al. [26]) that all translation invariant kernel methods can also use the larger class of cpd kernels. Therefore we give a short summary of this type of kernels and their connection to Hilbertian metrics.

Definition 2.1. A real valued function $\mathcal{K}$ on $\mathcal{X} \times \mathcal{X}$ is positive definite (pd) (resp., conditionally positive definite (cpd)) if and only if $\mathcal{K}$ is symmetric and $\sum_{i,j}^{n} c_i c_j \mathcal{K}(x_i, x_j) \geq 0$, for all $n \in \mathbb{N}$, $x_i \in \mathcal{X}, i = 1, \ldots, n$, and for all $c_i \in \mathbb{R}, i = 1, \ldots, n$, (resp., for all $c_i \in \mathbb{R}, i = 1, \ldots, n$ with $\sum_{i}^{n} c_i = 0$).

The following theorem describes the class of Hilbertian metrics:

Theorem 2.2 (Schoenberg [15]).

A metric space $(\mathcal{X}, d)$ can be embedded isometrically into a Hilbert space if and only if $- d^2(x, y)$ is conditionally positive definite (cpd).

Lemma 1 (van den Berg et al. [19]).

Let $\mathcal{K}$ be a kernel defined as $\mathcal{K}(x, y) = \widehat{k}(x, y) - \widehat{k}(x, x_0) - \widehat{k}(x_0, y) + \widehat{k}(x_0, x_0)$, where $x_0 \in \mathcal{X}$. Then $\mathcal{K}$ is pd if and only if $\widehat{k}$ is cpd.

Similar to pd kernel one can also characterize cpd kernel. Presently, one can write all cpd kernel in the form : $\mathcal{K}(x, y) = \frac{-1}{2} \| \phi_x - \phi_y \|_{\mathcal{H}}^2 + f(x) + f(y)$. The cpd kernel corresponding to Hilbertian (semi)-metrics are characterized by $f(x) = 0$ for all $x \in \mathcal{X}$, whereas if $\mathcal{K}$ is pd it follows

that $f(x) = \frac{1}{2}\mathcal{K}(x, x) \geq 0$. We also would like to point out that for SVM the class of Hilbertian (semi)-metrics is more important than the class of pd kernels. Namely, one can show (see Hein and Bousquet [24]), which the solution and optimization problem of the SVM only depends on the Hilbertian (semi)-metric, which is implicitly defined by each pd kernel. Moreover a whole family of pd kernels induces the same metric. Hilbertian metrics since, using Lemma 1, one can always define a corresponding pd kernel. Nevertheless for the convenience of the reader we will often explicity state the corresponding pd kernels.

2.2. Hilbertian metrics on probabilitity measures

For simplicity, we are dealing with the case of discrete probability measures on $D = \{1, 2, \ldots, N\}$, where $1 \leq N \leq \infty$. Given a Hilbertian metrics d on $\mathbb{R}_+$, it is easy to see that the metric on $d_{\mathcal{M}^1_+}$ given by

$D^2_{\mathcal{M}^1_+}(P, Q) = \sum_{i=1}^{N} d^2_{\mathbb{R}_+}(p_i, q_i)$ is a Hilbertian metric on $\mathcal{M}^1_+(D)$. The

following proposition extends the simple discrete case to the general case of a Hilbertian metrics on a probability space $\mathcal{X}$. In order to simplify the notation, we define $p(x)$ to be the Radon-Nikodym derivative $(dP / d\mu)(x)$ of P with respect to the dominating measure μ.

Proposition 1. *Let P and Q be two probability measures on $\mathcal{X}$, μ be an arbitrary dominating measure of P and Q and $d_{\mathbb{R}_+}$ (where $d_{\mathbb{R}_+}$ is the positive part of the real line with 0 included) be a $1/2$-homogeneous Hilbertian metrics on $\mathbb{R}_+$. Then $D^2_{\mathcal{M}^1_+(\mathcal{X})}$ defined as*

$$D^2_{\mathcal{M}^1_+}(P, Q) := \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(p(x), q(x))d_\mu(x) \tag{1}$$

is a Hilbertian metrics on $\mathcal{M}^1_+(\mathcal{X})$. $D^2_{\mathcal{M}^1_+(\mathcal{X})}$ is independent of the dominating measure μ.

Proof. First we show by using the $1/2$-homogeneity of $d_{\mathbb{R}^1_+}$ is independent of the dominating measure μ. We have

$$\int_{\mathcal{X}} d^2_{\mathcal{M}^1_+} \left(\frac{dP}{d\mu}, \frac{dQ}{d\mu} \right) d\mu = \int_{\mathcal{X}} d^2_{\mathcal{M}^1_+} \left(\frac{dP}{d\nu}, \frac{d\nu}{d\mu}, \frac{dQ}{d\nu}, \frac{d\nu}{d\mu} \right) \frac{d\mu}{d\nu} \, d\nu$$

$$= \int_{\mathcal{X}} d^2_{\mathcal{M}^1_+} \left(\frac{dP}{d\nu}, \frac{dQ}{d\nu} \right) d\nu,$$

where we use that $d^2_{\mathbb{R}_+}$ is 1-homogeneous. It is easy that $-D^2_{\mathcal{M}^1_+}(\mathcal{X})$ is conditionally positive definite, simplicity take for every $n \in \mathbb{N}$, $P_1, \dots, P_n$ the dominating measure $\dfrac{\sum_{i=1}^n P_i}{n}$ and use that $-D^2_{\mathbb{R}_+}$ is conditionally positive definite.

Remark 1. It is in principe very easy to build Hilbertian metrics on $\mathcal{M}^1_+(\mathcal{X})$ using arbitrary Hilbertian metrics on $\mathbb{R}_+$ and plugging it into the Equation 1.

But the key property of the method we propose is the independence of the metric d on $\mathcal{M}^1_+(\mathcal{X})$ of the dominating measure. That is we have generated a metric which is invariant with respect to general coordinate transformations on $\mathcal{X}$, therefore we call it a *covariant metric*.

2.3. λ-homogeneous Hilbertian metrics and positive definite kernels on $\mathbb{R}_+$

In this paper, we consider the class of Hilbertian metrics on probability measure, therefore the Hilbertian metrics on $\mathbb{R}_+$ is the main element of our approach. This is the event we require that the Hilbertian metrics on $\mathbb{R}_+$ is λ-homogeneous. The class of λ-homogeneous Hilbertian metrics on $\mathbb{R}_+$ was characterized by Fuglede:

Definition 2.3 (Topsøe [12] and Fuglede [5])**.**

A Hilbertian metrics is λ-homogeneous if and only if $d^2(cp, cq) = c^\lambda d^2(p, q)$ for all $c \in \mathbb{R}_+$.

Theorem 2.4 (Fuglede [5])**.**

A symmetric function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $d(x, y) = 0 \Leftrightarrow x = y$ is a γ-homogeneous, continuous Hilbertian metrics d on $\mathbb{R}_+$ if and only if there existe a (necessarily unique) non-zero bounded measure $\rho \geq 0$ on $\mathbb{R}_+$ such that d^2 can be written as

$$d^2(x, y) = \int_{\mathbb{R}_+} |x^{(\gamma+i\lambda)} - y^{\gamma+i\lambda}|^2 d\rho(\lambda). \tag{2}$$

Using Lemma 1, we define the corresponding class of pd kernel on $\mathbb{R}_+$ by choosing $x_0 = 0$. We will see later that this corresponds to choosing the zero-measure as origin of the RKHS (reproducing kernel Hilbert space).

Corollary 1. *A symmetric function $k : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $k(x, y) = 0 \Leftrightarrow x = 0$ is a 2γ-homogeneous continuous pd kernel $\mathcal{K}$ on $\mathbb{R}_+$ if and only if there exists a (necessarily unique) non-zero bounded symmetric measure $\kappa \geq 0$ on $\mathbb{R}$ such that $\mathcal{K}$ is given as*

$$\mathcal{K}(x, y) = \int_{\mathbb{R}} x^{(\gamma+i\lambda)} y^{(\gamma-i\lambda)} d\kappa(\lambda). \tag{3}$$

Proof. If we have the form given in (2), then it is obviously 2γ-homogeneous and since $\mathcal{K}(x, y) = x^{2\gamma}\kappa(\mathbb{R})$ we have $\mathcal{K}(x, x) = 0 \Leftrightarrow x = 0$.

The other direction follows by first noting that $\mathcal{K}(0, 0) = \langle \phi_0, \phi_0 \rangle = 0$ and then by applying Theorem 2.4, where κ is the symmetrized version of ρ around the origin, together with Lemma 1 and $\mathcal{K}(x, y) = \langle \phi_x, \phi_y \rangle = \frac{1}{2}(-d^2(x, y) + d^2(x, 0) + d^2(y, 0))$. $\qquad\qquad\square$

Andrezej Cichocki et al. ([2]) proposed an interesting two-parameter family of metrics, the AB-divergence defined:

Definition 2.5. The function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined as:

$$d_{AB}^{(\alpha,\beta)}(x, y) = \begin{cases} -\dfrac{1}{\alpha\beta}\left(x^\alpha y^\beta - \dfrac{\alpha}{(\alpha+\beta)}\,x^{\alpha+\beta} - \dfrac{\beta}{(\alpha+\beta)}\,y^{\alpha+\beta}\right) \text{ for } \alpha,\,\beta,\,\alpha+\beta \neq 0, \\[2ex] \dfrac{1}{\alpha^2}\left(x^\alpha \log(\dfrac{x^\alpha}{y^\alpha}) - x^\alpha + y^\alpha\right) \quad \text{for} \quad \alpha \neq 0,\, \beta = 0, \\[2ex] \dfrac{1}{\alpha^2}\left(\log(\dfrac{y^\alpha}{x^\alpha}) + (\dfrac{y^\alpha}{x^\alpha})^{-1} - 1\right) \quad \text{for} \quad \alpha = -\beta \neq 0, \\[2ex] \dfrac{1}{\beta^2}\left(y^\beta \log(\dfrac{y^\beta}{x^\alpha}) - y^\beta + x^\alpha\right) \qquad \text{for} \quad \alpha = 0;\, \beta \neq 0, \\[2ex] \dfrac{1}{2}(\log x - \log y)^2 \quad \text{for} \quad \alpha,\, \beta = 0, \end{cases}$$

(4)

where $d_{AB}^{(\alpha,\beta)}(x, y)$ is a divergence on $\mathbb{R}_+$.

2.4. A brief recall of support vector machine (SVM)

SVM were developed by Cortes and Vapnik [9] for binary classification. Their approach may be roughly sketched as follows:

● Class separation: basically, we are looking for the optimal separating hyperplane between the two classes by maximizing the margin between the classes closest points, the points lying on the boundaries are called support vectors, and the middle of the margin is our optimal separating hyperplane.

● Nonlinearity: when we cannot find a linear separator, data points are projected into an (usually) higher-dimensional space where the data points effectively become linearly separable (this projection is relised via kernel techniques).

● Problem solution: the whole task can be formulated as a quadratic optimization problem which can be solved by known techniques etc.

An implicit mapping Φ was used by SVM of the input data into a high-dimensional feature space defined by a kernel function, the inner product $\langle \Phi(x), \Phi(y) \rangle$ between the images of two data points x, y was returning by the function in the feature space. The kernel function can be represented as

$$\mathcal{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle, \tag{5}$$

where $\Phi : X \to H$ is the projection function, this function project x and y into the feature space H.

Relationship between the kernel method and SVM: Schölkopf showed that the class of cpd kernel can be used in SVM due to the translation invariant of the maximal margin problem in the RKHS, and the kernel can be used in SVM to the classification, the regression etc. if we found a good kernel function. The advantage of kernel method and SVM is that we can found and used a kernel for a problem particular that could be applied directly to data without the need for a feature extraction process. This was used in (Hein and Bousquet [24]) to show that the properties of the SVM only depend on the Hilbertian metrics. That is all cpd kernel are generated by a Hilbertian metric $d(x, y)$ through $\mathcal{K}(x, y)$ $= -d^2(x, y) + g(x) + g(y)$, where $g : \mathcal{X} \to \mathbb{R}$ and the solution of the SVM only depends on the Hilbertian metric $d(x, y)$.

3. Main Results

Generally the AB-metrics is not symmetric. We extend and improve this family a two-parameter symmetric. The metrics we propose is very interesting since it is a symmetric and smoothed variant from AB-metrics. This allows us to recover all property in Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ from the family of two-parameter.

Theorem 3.1. *The function* $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ *defined as*:

$$d_{ABS}^{\alpha,\beta}(x, y) = \begin{cases} \dfrac{1}{\alpha\beta}(x^\alpha - y^\alpha)(x^\beta - y^\beta) \ for \ \alpha, \beta \neq 0, \\[2mm] \dfrac{1}{\alpha^2}(x^\alpha - y^\alpha)\log(\dfrac{x^\alpha}{y^\alpha}) \ for \ \alpha \neq 0, \beta = 0, \\[2mm] \dfrac{1}{\alpha^2}((x^\alpha - y^\alpha)\log(\dfrac{x^\alpha}{y^\alpha}) + (\dfrac{x^\alpha}{y^\alpha})^{-1} + (\dfrac{y^\alpha}{x^\alpha})^{-1} - 2) \ for \ \alpha = -\beta \neq 0, \\[2mm] \dfrac{1}{\beta^2}(x^\beta - y^\beta)\log(\dfrac{x^\beta}{y^\beta}) \ for \ \alpha = 0, \beta \neq 0, \\[2mm] \dfrac{1}{2}(\log x - \log y) \ for \ \alpha = \beta = 0, \end{cases}$$

(6)

is a γ*-homogeneous Hilbertian metrics on* $\mathbb{R}_+$*, note that* $d_{ABS}^{(\alpha,\beta)}$ *is symmetric.*

Proof. The proof for the symmetry is trivial because this function is symmetric by construction and $d_{ABS}^{(\alpha,\beta)}(cx, cy) = c^{\alpha+\beta}d_{ABS}^{(\alpha,\beta)}(x, y)$, where $\gamma = \alpha + \beta$, then $d_{ABS}^{(\alpha,\beta)}$ is γ-homogeneous. Second for simplicity note that $\mathcal{K}(x, y) = -d^2(x, y)$, where d^2 is a Hilbertian metrics. The all conditions for theorem of Schoenberg satisfied we have $-d^2$ cpd.   □

We can now apply the principle to construct Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$, of building Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ and use the family of γ-homogeneous Hilbertian metrics $d_{ABS}^{(\alpha,\beta)}$ on $\mathbb{R}_+$.

Definition 3.2. We proposed two ways to build the symmetry divergence:

Type-1:

$$D_{ABS}^{(\alpha,\beta)} = \frac{1}{2}[D_{AB}^{(\alpha,\beta)}(P, Q) + D_{AB}^{(\alpha,\beta)}(Q, P)].$$

(7)

Type-2:

$$D_{ABS}^{(\alpha,\beta)}(P, Q) = \frac{1}{2}[D_{AB}^{(\alpha,\beta)}(P, \frac{P+Q}{2}) + D_{AB}^{(\alpha,\beta)}(Q, \frac{P+Q}{2})]. \qquad (8)$$

For the construction ABS-divergence, we used the definition above and we apply the Proposition 1 from the Section 2. We obtain the symmetric ABS-divergence (Type-1) defined as:

Definition 3.3. Let P and Q be two probability measures on $\mathcal{X}$ (probability space) and μ be an dominating measure of P, Q and $d_{ABS}^{(\alpha,\beta)}$ be a γ-homogeneous Hilbertian metrics on $\mathbb{R}_+$, then $D_{ABS}^{(\alpha,\beta)}$ defined as:

$$D_{ABS}^{(\alpha,\beta)}(P, Q) = \begin{cases} \dfrac{1}{\alpha\beta} \displaystyle\int_{\mathcal{X}} (p^\alpha - q^\alpha)(p^\beta - q^\beta) d\mu(x) \text{ for } \alpha \neq 0, \beta \neq 0, \\[2mm] \dfrac{1}{\alpha^2} \displaystyle\int_{\mathcal{X}} (p^\alpha - q^\alpha)\log(\dfrac{p^\alpha}{q^\alpha}) d\mu(x) \text{ for } \alpha \neq 0, \beta = 0, \\[2mm] \dfrac{1}{\alpha} \displaystyle\int_{\mathcal{X}} ((p^\alpha - q^\alpha)\log(\dfrac{p^\alpha}{q^\alpha}) + \dfrac{q^\alpha}{p^\alpha} + \dfrac{p^\alpha}{q^\alpha} - 2) d\mu(x) \text{ for } \alpha = -\beta \neq 0, \\[2mm] \dfrac{1}{\beta^2} \displaystyle\int_{\mathcal{X}} (p^\beta - q^\beta)\log(\dfrac{p^\beta}{q^\beta}) d\mu(x) \text{ for } \alpha = 0, \beta \neq 0, \\[2mm] \dfrac{1}{2} \displaystyle\int_{\mathcal{X}} (\log x - \log y) d\mu(x) \text{ for } \alpha = \beta = 0, \end{cases}$$

$$(9)$$

is a γ-homogeneous Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$.

The ABS-divergence has the following basic properties:

Properties 1.

(1) Convexity: $D_{ABS}^{(\alpha,\beta)}(P, Q)$ is convex with respect to both P and Q.

(2) Strict Positvity: $D_{ABS}^{(\alpha,\beta)}(P, Q) \geq 0$ and $D_{ABS}^{(\alpha,\beta)}(P, Q) = 0$ if and only if $P = Q$.

(3) Continuity: The ABS-divergences is continuous function of real variant (α, β) in the whole range including singularities.

(4) Symmetric: $D_{ABS}^{(\alpha, \beta)}(P, Q) = D_{ABS}^{(\alpha, \beta)}(Q, P)$.

(5) γ-homogeneous: $D_{ABS}^{(\alpha, \beta)}(cP, cQ) = c^{\alpha+\beta} D_{ABS}^{(\alpha, \beta)}(P, Q)$.

We used the instrument of building Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ and use the family of $(\alpha + \beta)$-homogeneous Hilbertian metrics $d_{ABS}^{(\alpha, \beta)}$ on $\mathbb{R}_+$. This yield as special case the following measures on $\mathcal{M}_+^1(\mathcal{X})$.

$$D_{ABS}^{\alpha, \beta}(P, Q) = \int_{\mathcal{X}} \varphi_{\alpha, \beta}(p(x)q(x)) d\mu(x).$$

Table (a). $D_{ABS}^{\alpha, \beta}(P, Q)$ divergence

Divergence $D_{ABS}^{\alpha, \beta}(P, Q)$	Function $\varphi_{\alpha, \beta}(p(x)q(x))$	Name
$D_{ABS}^{(1,1)}(P, Q)$	$(p(x) - q(x))^2$	Euclidian
$D_{ABS}^{(1/2,1)}(P, Q)$	$2(\sqrt{p(x)} - \sqrt{q(x)})(p(x) - q(x))$	V_1-Hellinger
$D_{ABS}^{(1/2,-1)}(P, Q)$	$2(\sqrt{p(x)} - \sqrt{q(x)}) \dfrac{(p(x) - q(x))}{p(x)q(x)}$	V_2-Hellinger
$D_{ABS}^{(1/2,1/2)}(P, Q)$	$4(\sqrt{p(x)} - \sqrt{q(x)})^2$	Hellinger
$D_{ABS}^{(1,0)}(P, Q)$	$(p(x) - q(x)) \log(\dfrac{p(x)}{q(x)})$	Jeffrey

$D_{ABS}^{(1,1)}$ corresponds to the square of Euclidian metric, $D_{ABS}^{(1/2,1/2)}$ corresponds to the Hellinger metric which is well known in the statistics community, $D_{ABS}^{(1,0)}$ correspond to the Jeffreys metric, $D_{ABS}^{(1/2,1)}$ V_1-Hellinger and $D_{ABS}^{(1/2,-1)}$ V_2-Hellinger is a variant of Hellinger metrics.

For completeness we also give the corresponding pd kernels on $\mathcal{M}_+^1(\mathcal{X})$, where we take in Lemma 1 the zero measure as $x_0 \in \mathcal{M}_+^1(\mathcal{X})$. This choice seems strange at first since we are dealing with probability measures. But in fact the whole framework presented in this paper can easily be extended to all finite, positive measure on $\mathcal{X}$. For this, set zero measure is a natural choice of the origin.

$$\mathcal{K}_{(1,1)}(P,\, Q) = \int_{\mathcal{X}} p(x)q(x)d\mu(x),$$

$$\mathcal{K}_{(1/2,1)}(P,\, Q) = \int_{\mathcal{X}} (q(x)\sqrt{p(x)} + p(x)\sqrt{q(x)})d\mu(x),$$

$$\mathcal{K}_{(1/2,1/2)}(P,\, Q) = 4\int_{\mathcal{X}} \sqrt{p(x)q(x)}d\mu(x),$$

$$\mathcal{K}_{(1,0)}(P,\, Q) = \frac{1}{2}\left(\int_{\mathcal{X}} (q(x) - p(x))\log\left(\frac{p(x)}{q(x)}\right) - p(x) - q(x))d\mu(x).$$

Using two-parameters is so difficult in practice that is the reason why we will proposed a one-parameter family to improve the ABS-divergence.

Proposition 2. *The function $d : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ defined as:*

$$d_t^2(x,\, y) = \begin{cases} \frac{1}{2}\left(\dfrac{x^t - y^t}{t}\right)^2 & \text{for } t \neq 0, \\ \frac{1}{2}(\log x - \log y) & \text{for } t = 0, \end{cases} \tag{10}$$

is a $2t$-homogeneous Hilbertian metric on $\mathbb{R}_+$ if $t \neq 0$.

Proof. Note that d_t^2 is symmetric by construction and it's easy to verified the property of Hilbertian metric. Therefore, we show that:

$$d_t^2(cx,\, cy) = \frac{1}{2}\left(\frac{(cx)^t - (cy)^t}{t}\right)^2 = \frac{c^{2t}}{2}\left(\frac{x^t - y^t}{t}\right)^2 = c^{2t}d_t^2(x,\, y). \qquad \square$$

We used the Proposition 1 of building Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ and use the family of $2t$-homogeneous Hilbertian metrics d_t^2 on $\mathbb{R}_+$. Therefore, we obtain as special case the following measures on $\mathcal{M}_+^1(\mathcal{X})$.

$$D_t^2(P, Q) = \int_{\mathcal{X}} \varphi_t(p(x)q(x))d\mu(x).$$

Table (b). Divergence using $2t$-homogeneous Hilbertian metric

Divergence $D_t^2(P, Q)$	Function $\varphi_t(p(x)q(x))$	Name
$D_1^2(P, Q)$	$\frac{1}{2}(p(x) - q(x))^2$	Euclidian
$D_{-1}^2(P, Q)$	$\frac{1}{2}(\frac{1}{p(x)} - \frac{1}{q(x)})^2$	S-Euclidian
$D_{-\frac{1}{2}}^2(P, Q)$	$2(\sqrt{p(x)} - \sqrt{q(x)})^2$	Hellinger
$D_{\frac{1}{2}}^2(P, Q)$	$2(\frac{1}{\sqrt{p(x)}} - \frac{1}{\sqrt{q(x)}})^2$	S-Itakura Saito

D_1^2 correspond to the square of Euclidian metric, D_{-1}^2 is an other version of Euclidian metric, $D_{1/2}^2$ corresponds two Hellinger metric. The Hellinger metric is well-known in the statistics community. $D_{-\frac{1}{2}}$ is a symmetrized Itakura-Saito distance (called also the COSH distance) modified.

For completeness we also give the corresponding pd kernels on $\mathcal{M}_+^1(\mathcal{X})$, where we take in Lemma 1 the zero measure as $x_0 \in \mathcal{M}_+^1(\mathcal{X})$. This choice seems strange at first since we are dealing with probability measures. But in fact the whole framework presented in this paper can easily be extended to all finite, positive measures on $\mathcal{X}$. For this, set zero measure is a natural choice, the origin.

$$\mathcal{K}_t(P,\,Q) = \frac{1}{2}\int_\mathcal{X} \frac{x^t y^t}{t^2}\,d\mu,$$

$$\mathcal{K}_1(P,\,Q) = \frac{1}{2}\int_\mathcal{X} p(x)q(x)d\mu(x),$$

$$\mathcal{K}_{-1}(P,\,Q) = \frac{1}{2}\int_\mathcal{X} \frac{1}{p(x)q(x)}\,d\mu(x),$$

$$\mathcal{K}_{\frac{1}{2}}(P,\,Q) = 2\int_\mathcal{X} \sqrt{p(x)q(x)}\,d\mu(x),$$

$$\mathcal{K}_{-\frac{1}{2}}(P,\,Q) = 2\int_\mathcal{X} \frac{1}{\sqrt{p(x)q(x)}}\,d\mu(x).$$

4. Numerical Studies

To show the interest of our study the examples of applications have been proposed. For the SVM, we made studies on the classification of the genes and on the sex of the cats knowing the weight of the heart and the body. As regards classification of images, we use our divergences to separate them into two classes.

4.1. Application in SVM

The performance of metrics and kernels has been compared in classification using some data sets. All data sets were split into a training (80%) and a test (20%) set. For the problem we use SVM method. For all experiments we use the one-parameter family d_t^2 of Hilbertian metric, the corresponding kernel cpd is $\mathcal{K} = -D_t^2$ with varying penality constants C in the SVM, and we use the transformed kernel (Gaussian transformation):

$$\mathcal{K}_t(P,\,Q) = e^{-D_t^2(P,\,Q)/2\sigma^2}.$$

The test error was evaluated by the best parameters C and σ. The best constant penality C and σ was found by cross-validation.

We evaluated the performance of the proposed metrics and kernels in three classifications tasks. Firstly, we generated a artificial data and we consider the test error for kernels proposed. Secondly as a real world application, let us test the ability of SVM to predict the class of a tumour from gene expression data. We use a publicly available datasets of gene expression data for 128 differents individuals with acute lymphoblastic leukemia (ALL). Here we focus on predicting the type of the disease (*B*-cell or *T*-cell). Therefore we test a SVM classifier for cancer diagnosis from gene expression data, and we test the ability of a SVM to predict the class of the disease from gene expression. Finally, we apply the data from support functions and datasets for Venables and Ripley's MASS. We use the anatomical data from domestic cats, the heart and body weights of samples of male and female cats used for digitalis experiments. The cats were all adult, over 2kg body weight. We presented the classification error according to the sex of the cats.

The tables shows the test errors for the kernels corresponding to $t = \{-1; -1/2; 1/2; 1\}$ from the ABS-divergence resp. and their Gaussian transformation. The first line shows the kernels directly (dir) and the second line the Gaussian transformation (tran).

Table 1. Test error using data artificial

Divergence	Euclidian			Hellinger			Itakura-Saito			S-Euclidian		
	error	C	σ	error	C	σ	error	C	σ	error	C	σ
Data (dir)	0.0001	10	–	0.005	10	–	0.26	100	–	0.13	10	–
Artificial (tran)	0.0001	10	1.5	0.005	10	0.5	0.125	100	1.5	0.17	100	1.5

Table 1 shows that the errors committed using the conditionally defined positive $\mathcal{K}$ and the kernel transformed $\mathcal{K}_t$ are in the same proportion for the Euclidian divergence and the Hellinger divergence with $\sigma = 1.5$ and the constant $C = 10$. While for the Itakura-Saito divergence the errors committed in the classifications with the $\mathcal{K}_t$ kernel are smaller than that of the $\mathcal{K}$ kernel with $\sigma = 1.5$ and $C = 100$. So the

classification with the transformed core is better than the one used directly. For the S-Euclidian divergence by varying the constant C, $C = 10$ for the kernel $\mathcal{K}$ and $C = 100$ for the kernel transforms $\mathcal{K}_t$, we find that the transformed nucleus offers a better classification result. Thus in all cases we notice that the transformed nuclei give the best results. For the sake of using the transformed nuclei for artificial data.

Table 2. Test error using data "ALL" gene

Divergence	Euclidian			Hellinger			Itakura-Saito			S-Euclidian		
	error	C	σ	error	C	σ	error	C	σ	error	C	σ
Data "ALL" (dir)	0.679	10	–	0.414	10	–	0.257	10	–	0.461	10	–
gene (tran)	0.0234	10	0.5	0.156	100	1.5	0.0468	100	1.5	0.164	100	1.5

With the data gene (ALL), we find that the kernel transformed offers the best results for the classification. However with the use of kernels $\mathcal{K}$ it is the kernel constructed with the divergence of Itakura-Saito that gives the best classification results, followed by Hellinger, S-Euclidian and the Euclidian divergence. Whereas if we use kernel transformed, it is those built with Euclidian divergence that give the best results, followed by that of Itakura-Saito, Hellinger and S-Euclidian.

Table 3. Test error using data cats "MASS"

Divergence	Euclidian			Hellinger			Itakura-Saito			S-Euclidian		
	error	C	σ	error	C	σ	error	C	σ	error	C	σ
Data "MASS" (dir)	0.236	1	–	0.326	1	–	0.340	1	–	0.326	10	–
cats (tran)	0.194	1	0.5	0.326	10	0.5	0.368	1	0.5	0.181	1	0.5

With the data cats (MASS), if we use kernel $\mathcal{K}$ we have almost the same results for all measures of divergence. However with the modified kernel the classification obtained with the S-Euclidian divergence and the Euclidian divergence offer the best results. It should also be noted that for the divergence of Itakura-Saito it is the kernel $\mathcal{K}$ that gives the best results. In conclusion, we retain that the choice of the best method depends on data and constants.

4.2. Applications in images classifications

In this work, we apply the above proposed algorithm and also implement with the metrics proposed in the section above, to color image segmentation. The divergence we proposed gives a good classification with different threshold (k). In our experiments we used the data from the image (a) and apply our algorithm with different threshold. From these figures, we can see the results experimented by our algorithm. These results have been obtained with our divergences following some (k) values.
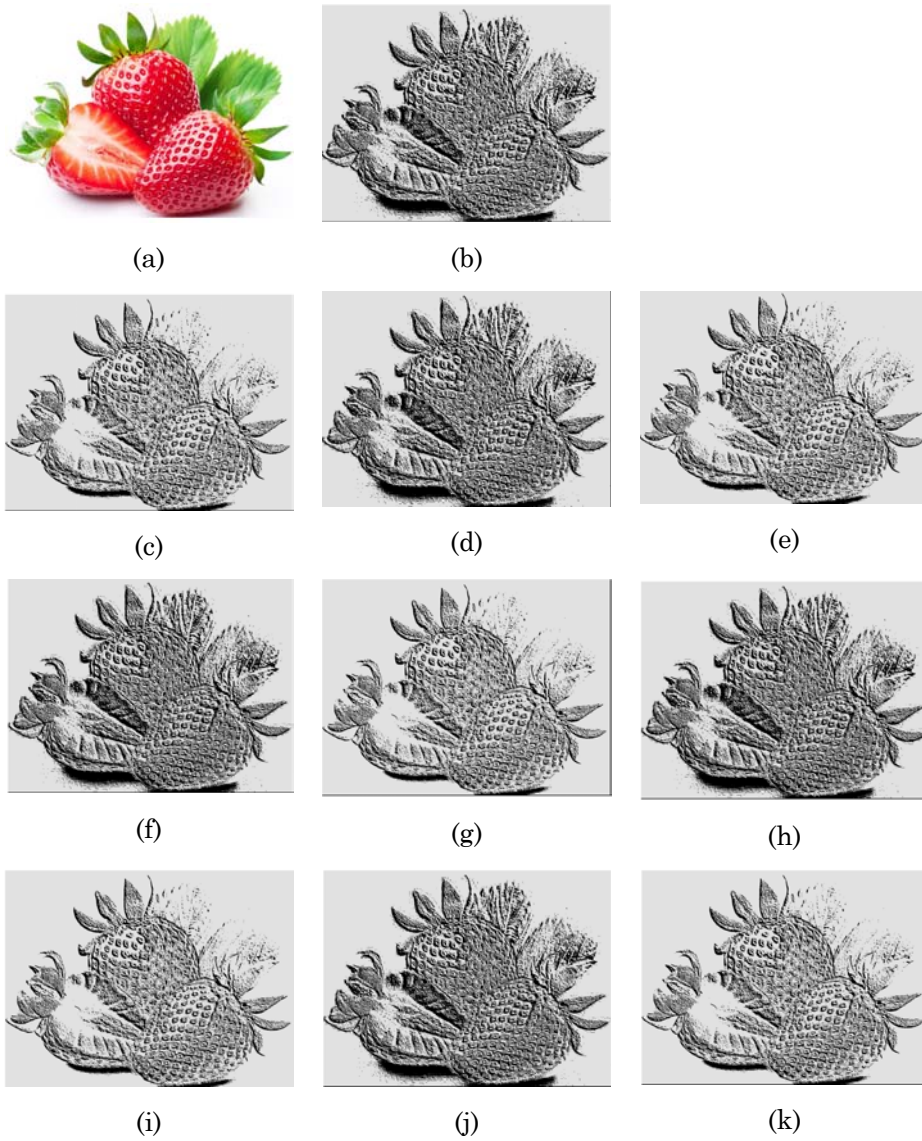
Algorithm		
$X \in M_{l,c}(\mathbb{R})$		matrix origin
$X' \in M_{l,c}(\mathbb{R})$		matrix sortie
X'	$\leftarrow$	matrix null order $l \times c$
for i	$\mathrm{rang}\,(1,\, l-1):$	
P_1	$\leftarrow$	$X_{i,\,j-1}$
P_2	$\leftarrow$	$X_{i-1,\,j}$
If	$\mathrm{norm}\,(P_1,\, P_2) < k$	
P_0	$\leftarrow$	$(225, 225, 225)$
else:		
P_0	$\leftarrow$	$(0, 0, 0)$
X'	$\leftarrow$	P_0

4.2.1. Facial image segmentation
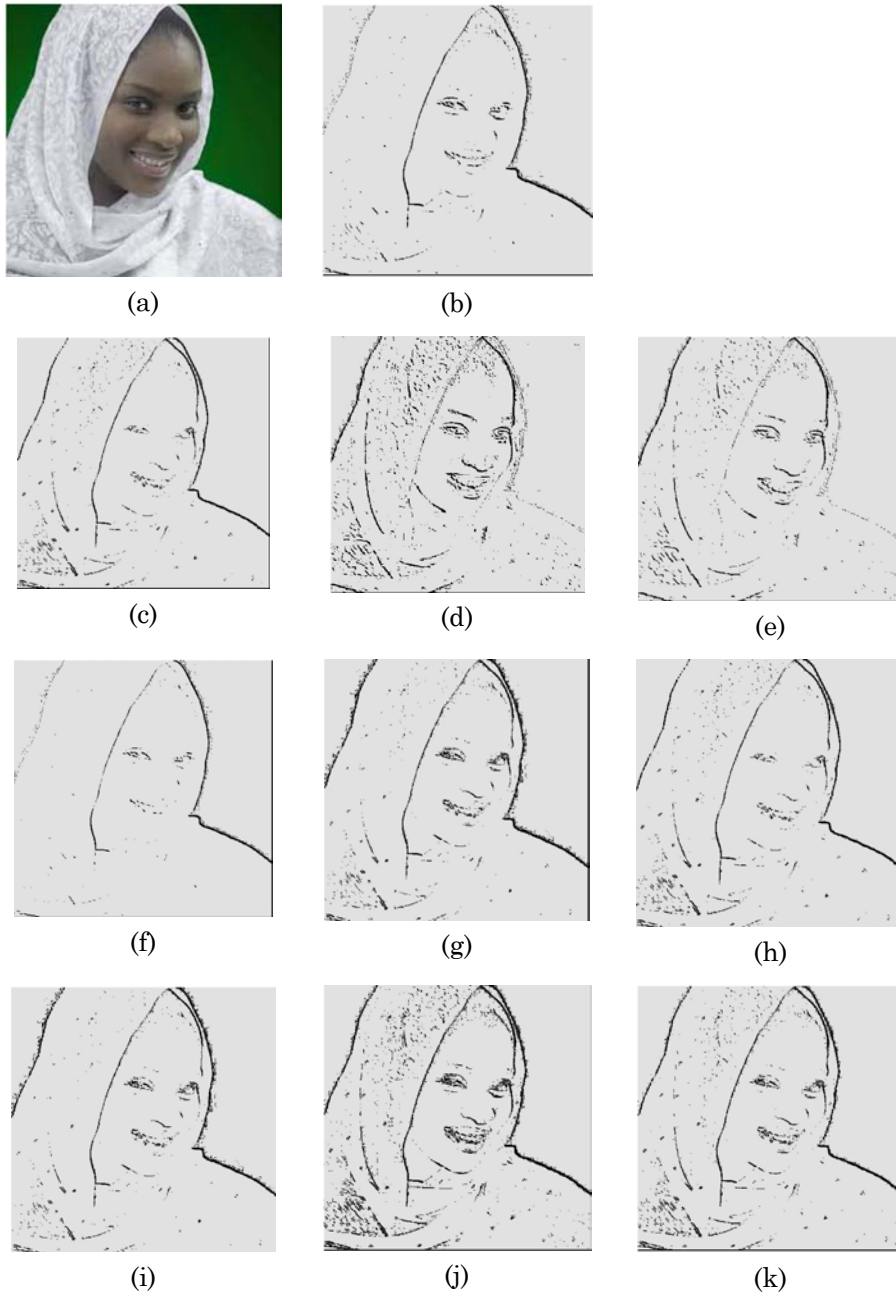
Figure 1. Results facial image segmentation.

In this part of the work we will make a brief presentation of the classification of an image. A presentation of the transformation with the corresponding divergence and the optimal threshold. The color image segmentation results is that: (a) corresponds to the original image, (b) image $D_{ABS}^{(1,0)}$-divergence with $k = 2.4$, (c) image $D_{ABS}^{(1,1)}$-divergence with $k = 1354$, (d) image $D_{ABS}^{(1/2,-1)}$-divergence with $k = 2.5$, (e) image $D_{ABS}^{(1/2,-1)}$-divergence with $k = 3.4$, (f) image $D_{ABS}^{(1,0)}$-divergence with $k = 3.5$, (g) image $D_{ABS}^{(1/2,1/2)}$-divergence with $k = 4.5$, (h) image $D_{ABS}^{(1,1)}$-divergence with $k = 1350$, (i) image $D_{ABS}^{(1/2,1/2)}$-divergence with $k = 5.5$, (j) image $D_{ABS}^{(1/2,1)}$-divergence with $k = 50$, and (k) image $D_{ABS}^{(1/2,1)}$-divergence with $k = 80$.

From these figures (Figure 1), we can observe the experimental results of our algorithm on a facial image. Using on divergences in the proposed algorithm, we can observe the separation into two classes of our image according to the metric used with an adequate threshold. We can see some differences between the segmentation results images (b)-(k). For example, the images (b) and (f) given by the $D_{ABS}^{(1,0)}$ divergence showed a bleary delimitation of the original image (a); for the images (c), (e), (h), (i) we used (resp.,) $D_{ABS}^{(1,1)}$, $D_{ABS}^{(1/2,-1)}$, $D_{ABS}^{(1,1)}$, and $D_{ABS}^{(1/2,1/2)}$ divergences they correctly delineates the contour of the image (a); for the images (d), (g), (j) the $D_{ABS}^{(1/2,-1)}$, $D_{ABS}^{(1/2,1/2)}$, and $D^{(1/2,1)}$ divergences results are relatively homogeneous and do a good work.

4.2.2. Fruit image segmentation

(a)   (b)

(c)   (d)   (e)

(f)   (g)   (h)

(i)   (j)   (k)

Figure 2. Fruit image segmentation.

The images in Figure 2 correspond to a classification using the proposed algorithm. The classification uses our divergences with a certain decision threshold k. Image (a) corresponds to the original images. The image (b) corresponds to the classification using the divergence $D_{ABS}^{(1,1)}$ with a threshold $k = 0.1$, the image (c) is that of this divergence with the threshold $k = 0.5$. The images (d) and (e) correspond to the classification the divergence $D_{ABS}^{(\frac{1}{2},\frac{1}{2})}$ with the thresholds $k = 0.1$ and $k = 0.5$, respectively. The classification made with the divergence $D_{ABS}^{(\frac{1}{2},1)}$ with the threshold $k = 0.1$ and $k = 0.5$, respectively corresponding to the (f) and (g). The images (h) and (i) are obtained with the divergence $D_{ABS}^{(\frac{1}{2},-1)}$ with the thresholds $k = 0.1$ and $k = 0.5$. The images (j) and (k) correspond to the classification with the divergence $D_{ABS}^{(1,0)}$ with $k = 0.1$ and $k = 0.5$, respectively.

From these figures (Figure 2), we can observe the experimental results of our algorithm on a fruit image. Using on divergences in the proposed algorithm, we can observe the separation into two classes of our image according to the metrics used with an adequate threshold. For $k = 0.1$ figures (b), (d), (f), (h) present the better representation image than $k = 0.5$ images (c), (e), (g), (k). The fruits images representations depend on two things k and the divergences. But in this case we have a few differences between our divergences.

5. Conclusion

A general method to build Hilbertian metrics on probability measures from Hilbertian on $\mathbb{R}_+$ was presented. Using results from Cichocki and Amari from the Alpha-Beta-divergence, then we generalized this framework by incorporating the symmetry property. We propose a new

variant of Alpha-Beta-Symmetric divergence metrics (ABS-divergence) and kernels associated. Our main contributions consist of, first is to construct a new family of metrics, ABS-divergence and kernels, and second is to be integrated into SVM and algorithm classification. Our results, which are based on a choice of ABS divergence parameters leading to symmetric kernel $\mathcal{K}_t$, are very efficient compared to classical $\mathcal{K}$-based methods.

Acknowledgements

References

[1] Andrzej Cichocki, Sergio Cruces and Shun-ichi Amari, Log-determinant revisited: Alpha-beta and gamma log-det divergences, Entropy 17(5) (2015), 2988-3034.

DOI: https://doi.org/10.3390/e17052988

[2] Andrzej Cichocki, Sergio Cruces and Shun-ichi Amari, Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization, Entropy 13(1) (2011), 134-170.

DOI: https://doi.org/10.3390/e13010134

[3] Asa Ben-Hur and Willian Stafford Noble, Kernel methods for predicting protein-protein-interactions, Bioinformatics 21(1) (2005), i38-i46.

DOI: https://doi.org/10.1093/bioinformatics/bti1016

[4] B. Frénay and M. Verleysen, Parameter-insensitive kernel in extreme learning for nonlinear support vector regression, Neurocomputing 74(16) (2011), 2526-2531.

DOI: https://doi.org/10.1016/j.neucom.2010.11.037

[5] B. Fuglede, Spirals in Hilbert space: With an application in information theory, Expositiones Mathematicae 23(1) (2005), 23-45.

DOI: https://doi.org/10.1016/j.exmath.2005.01.014

[6] B. Schölkopf and A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[7]   Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf and Gert R. G. Lanckriet, Hilbert space embeddings and metrics on probability measures, Journal of Machine Learning Research 11 (2010), 1517-1561.

[8]   Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[9]   C. Cortes and V. Vapnik, Support-vector networks, Machine Learning 20(3) (1995), 273-297.

DOI: https://doi.org/10.1007/BF00994018

[10] D. B. Thiyam, S. Cruces, J. Olias and A. Cichocki, Optimization of alpha-beta log-det divergences and their application in the spatial filtering of two class motor imagery movements, Entropy 19(3) (2017), Article 89.

DOI: https://doi.org/10.3390/e19030089

[11] D. Olszewski and B. Ster, Asymmetric clustering using the alpha-beta divergence, Pattern Recognition 47(5) (2014), 2031-2041.

DOI: https://doi.org/10.1016/j.patcog.2013.11.019

[12] F. Topsøe, Jenson-Shannon divergence and norm-based measures of discrimination and variation, Preprint (2003).

[13] Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning, Springer Series in Statistics, 2001.

[14] Ho Chung Leon Law, Dougal J. Sutherland, Dino Sejdinovic and Seth Flanman, Bayesian Approaches to Distribution Regression, Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS), Lanzarote, Spain, PMLR: Volume 84 (2018).

[15] I. J. Schoenberg, Metric space and positive definite functions, Transactions of the American Mathematical Society 44(3) (1938), 522-536.

DOI: https://doi.org/10.1090/S0002-9947-1938-1501980-0

[16] F. Amara, M. Fezari and H. Bourouba, An improved GMM-SVM system based on distance metric for voice pathology detection, Applied Mathematics and Information Sciences An International Journal 10(3) (2016), 1061-1070.

DOI: https://doi.org/10.18576/amis/100324

[17] I. W. Sumarjaya, A survey of kernel-type estimators for copula and their applications, Journal of Physics: Conference Series 893 (2017), Article 012027.

DOI: https://doi.org/10.1088/1742-6596/893/1/012027

[18] J. Lafferty and G. Lebanon, Diffusion kernels on statistical manifolds, Journal of Machine Learning Research 6 (2005), 129-163.

[19] C. van den Berg, J. P. R. Christensen and P. Ressel, Harmonic Analysis on Semigroups, Springer, New York 29 (1984), 438.

[20] John Shawe-Taylor and Nello Cristianini, Kernel Methods for Pattern Analysis, 2004.

[21] K. Muandet, K. Fukumizu, B. Sriperumbudur and B. Schölkopf, Kernel mean embedding of distribution: A review and beyond, Foundations and trends in Machine Learning 10(1-2) (2017), 1-141.

DOI: http://dx.doi.org/10.1561/2200000060

[22] M. Hein, T. N. Lal and O. Bousquet, Hilbertian Metrics on Probability Measures and their Application in SMVs, Accepted at DAGM Springer 3175 (2004), 270-277.

[23] M. Hein, O. Bousquet and B. Schölkopf, Maximal margin classification for metric spaces, Journal of Computer and System Sciences 71(3) (2005), 333-359.

DOI: https://doi.org/10.1016/j.jcss.2004.10.013

[24] M. Hein and O. Bousquet, Maximal margin classification for metric spaces, In 16th Annual Conference on Learning Theory (COLT), 2003.

[25] O. Chapelle, P. Haffner and V. Vapnik, Support vector machines for histogram-based image classification, IEEE Transaction on Neural Networks 10(5) (1999), 1055-1064.

DOI: https://doi.org/10.1109/72.788646

[26] P. J. Moreno, P. P. Ho and N. Vasconcelos, A Kullback-Leibler divergence based kernel for SVM classification in multimedia application, NIPS 16 (2003).

[27] T. Jebara and R. Kondor, Bhattacharyya and Expected Likelihood Kernels, In 16th Annual Conference on Learning Theory (COLT) (2003), 57-71.

[28] T. J. Abrahamsen, L. K. Hansen and O. Winther, Kernel Methods for Machine Learning with Life Science Applications, Technical University of Denmark (DTU), PHD, No. 299, 2013.

[29] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

∎