DISSIMILARITY MEASURES IN FEATURE SPACE

Frédéric Desobry and Manuel Davy

IRCCyN, UMR CNRS 6597, 1 rue de la Noë - BP92101, 44321 Nantes Cedex 3 - France
{Frederic.Desobry,Manuel.Davy}@irccyn.ec-nantes.fr

ABSTRACT

In this paper, we present a study of the statistical behavior of the dissimilarity measure $\mathcal{D}_{\mathcal{S}}$, proposed in [1] and which results from a machine learning-based quantile estimation approach, namely: single-class support vector machine. This dissimilarity measure possesses the interesting property of being asymptotically equivalent to the Fisher ratio when dealing with radial Gaussian probability density functions. More generally, it can be efficiently applied to non-connected quantiles, and to noisy data sets, as outliers are taken into account by the SVM. A generalisation of $\mathcal{D}_{\mathcal{S}}$ is then proposed, which results in the design of a more general class of dissimilarity measures, also defined in feature space and with the same properties.

1. INTRODUCTION

Many Signal Processing applications require the comparison of data sets, via a dissimilarity measure. In [1], we presented an abrupt change detection algorithm based on a dissimilarity measure $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$. Other typical examples can be found in Independent Component Analysis (contrast functions) and in Pattern Recognition. In this paper, we show that the dissimilarity measure introduced in [1] has good properties, and that it can be generalised to a wide class of measures in the so-called *feature space*.

More precisely, let $\boldsymbol{x}_1 = \{x_1^1, \ldots, x_1^{m_1}\}$ (resp. $\boldsymbol{x}_2 = \{x_2^1, \ldots, x_2^{m_2}\}$) be a set of vectors in a space $\mathcal{X}$ i.i.d. sampled according to an unknown probability density function (pdf) p_1 (resp. p_2). We want a dissimilarity measure $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ to be small if the vectors in $\boldsymbol{x}_1$ are located in the same part of the space as the vectors in $\boldsymbol{x}_2$ and of course, $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ must be large in any other case. This enables the following test, typically implemented in decision algorithms:

$$\begin{cases} H_0: & \mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) & \leq & \eta & \text{(Sets are similar)} \\ H_1: & \mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) & > & \eta & \text{(Sets are dissimilar)} \end{cases} \quad (1)$$

where η is a threshold that tunes the sensibility/robustness compromise. In, e.g., detection problems, η tunes the false alarm/miss alarm ratio.

Under the assumption that p_1 and p_2 are unknown, if one wants to build $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ upon statistics based on the sole knowledge of the data sets $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, possible approaches include the comparison of statistics directly computed on the data, of inferred distributions, of empirical density estimates... The approach in [1] was slightly different, as we had defined a dissimilarity measure $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ built on estimated quantiles of p_1 and p_2. Roughly, a quantile $\mathcal{R}^{\mathcal{X}}_{\boldsymbol{x}_1}$ for the pdf p_1 is a region of the space $\mathcal{X}$ that contains most of p_1 probability mass. In [1], quantile estimation was performed via a single class ν-support vector machine (SVM), and $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ was defined as a Fisher-like ratio aimed at comparing the estimated quantiles $\mathcal{R}^{\mathcal{X}}_{\boldsymbol{x}_1}$ and $\mathcal{R}^{\mathcal{X}}_{\boldsymbol{x}_2}$ of pdfs p_1 and p_2. This is recalled in Section 2.

Though $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ is defined as a Fisher-like ratio in feature space, we have no guaranty that it actually behaves like the true Fisher ratio when p_1 and p_2 *are* Gaussian pdfs, i.e. when the Fisher ratio is most relevant. Assume p_1 and p_2 are Gaussian with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and with covariance matrices Σ_1 and Σ_2. The Fisher ratio is [2]

$$\mathcal{D}_F(p_1, p_2) \;=\; (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}} (\Sigma_1 + \Sigma_2)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad (2)$$

In the case where $\Sigma_1 = \Sigma_2 \triangleq \Sigma$, $\mathcal{D}_F(p_1, p_2)$ is a distance between p_1 and p_2, as it equals the Kullback-Leibler divergence, and also the Mahalanobis distance denoted $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_{\Sigma,\mathcal{X}}$ (see, e.g., [3]). In Section 3, we show that the dissimilarity measure defined in [1] actually behaves asymptotically (as m_1,$m_2 \to \infty$) like the Fisher ratio (or equivalently, the Kullback-Leibler divergence) in the case of Gaussian distributions with identical covariance matrices. Moreover, $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ has a relevant behavior even when $\mathcal{D}_F(\cdot,\cdot)$ does not.

The results obtained in Section 3 are a starting point to build a large class of dissimilarity measures also defined in feature space, which all possess the above asymptotic property. These dissimilarity measures are built in Section 4, using so-called *metric preserving* functions. Section 5 is dedicated to the comparison of the Fisher ratio to some dissimilarity measures designed in Section 4. Finally, conclusion and future research directions are proposed in 6.

2. A FIRST DISSIMILARITY MEASURE

For the sake of brevity, we do not recall here single-class ν-SVM. Sufficient elements can be found in, e.g., [4]. We employ the following usual notations: $k(\cdot,\cdot)$ is a kernel inducing a mapping $\phi(\cdot)$ from input space $\mathcal{X}$ to feature space $\mathcal{H}$,

and the α_j^i, $i = 1, \ldots, m_j$, $j = 1, 2$, are the weights yielded by the SVM. The inner product in $\mathcal{H}$ between two mapped training vectors $< \phi(x_i), \phi(x_j) >_{\mathcal{H}}$ equals $k(x_i, x_j)$. In the following, we assume that $k(\cdot, \cdot)$ takes values between 0 and 1, and is such that $\forall x \in \mathcal{X}, k(x, x) = 1$[1].

Assume that two single-class classifiers are trained *independently* on the sets $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, yielding the regions $\mathcal{R}^{\mathcal{X}}_{\boldsymbol{x}_1}$ and $\mathcal{R}^{\mathcal{X}}_{\boldsymbol{x}_2}$ or, equivalently in feature space $\mathcal{H}$, the hyperplanes $\mathcal{W}_1$ and $\mathcal{W}_2$ parametered by $(\mathbf{w}_1, \rho_1)$ and $(\mathbf{w}_2, \rho_2)$. In $\mathcal{H}$, the vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ define a 2-dimensional plane, denoted $\mathcal{P}$, that intersects the hypersphere $\mathcal{S}$ along a circle with center $\mathbf{O}$ and radius 1, as depicted in Fig. 1. Actually, in the pathological case where $\mathbf{w}_1$ and $\mathbf{w}_2$ are collinear, there is an infinity of planes $\mathcal{P}$, and one can select any of them. We recall that $\mathbf{w}_1$ (resp. $\mathbf{w}_2$) equals $\sum_{i=1}^{m_1} \alpha_1^i \phi(x_1^i)$ (resp. $\sum_{i=1}^{m_2} \alpha_2^i \phi(x_2^i)$).

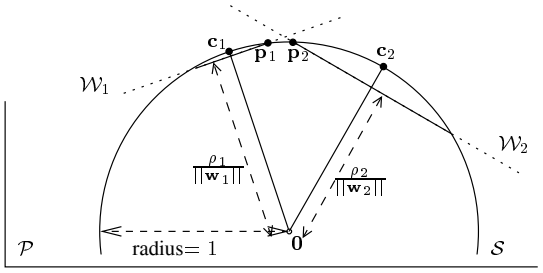

Fig. 1. The SV single-class classifiers yield two hyperplanes $\mathcal{W}_1(\mathbf{w}_1, \rho_1)$ and $\mathcal{W}_2(\mathbf{w}_2, \rho_2)$. The circle represented corresponds to the intersection of the plane $\mathcal{P}$ (uniquely defined by $\mathbf{w}_1$ and $\mathbf{w}_2$) and the hypersphere $\mathcal{S}$. The intersection of the line $\big(0, \mathbf{w}_1\big)$ (resp. $\big(0, \mathbf{w}_2\big)$) with $\mathcal{S}$ yields $\mathbf{c}_1$ (resp. $\mathbf{c}_2$), and the intersection of the hyperplane $\mathcal{W}_1$ (resp. $\mathcal{W}_2$) with $\mathcal{S}$ in the plane $\mathcal{P}$ yields two points, any of which is denoted $\mathbf{p}_1$ (resp. $\mathbf{p}_2$).

In feature space, the hyperplane $\mathcal{W}_1$ (resp. $\mathcal{W}_2$) bounds the segment of $\mathcal{S}$ where all the mapped inliers of $\boldsymbol{x}_1$ (resp. $\boldsymbol{x}_2$) lie. Mapped training vectors located right on the boundary $\mathcal{W}_1$ (resp. $\mathcal{W}_2$) are called *Margin support vectors*, and are denoted $\mathbf{x}_1^{\text{MSV}}$ (resp. $\mathbf{x}_2^{\text{MSV}}$). In [1], we proposed the following intra-regions/inter-regions ratio inspired by the Fisher ratio [2]:

$$\mathcal{D}_{\mathcal{S}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \;=\; \frac{d_{\text{arc}}(\mathbf{c}_1, \mathbf{c}_2)}{d_{\text{arc}}(\mathbf{c}_1, \mathbf{p}_1) + d_{\text{arc}}(\mathbf{c}_2, \mathbf{p}_2)} \qquad (3)$$

where $d_{\text{arc}}(\mathbf{a}, \mathbf{b}) = \widehat{\mathbf{ab}}$ denotes the arc distance between two vectors $\mathbf{a}$ and $\mathbf{b}$ on $\mathcal{S}$. Considering $\boldsymbol{x}_1$ only, we see that the arc distance $d_{\text{arc}}(\mathbf{c}_1, \mathbf{p}_1)$ is a measure of the spread of samples of the mapped training set $\phi(\boldsymbol{x}_1)$ in feature space. The more these samples are spread, the higher the distance $d_{\text{arc}}(\mathbf{c}_1, \mathbf{p}_1)$, and the smaller the margin $\rho_1 / ||\mathbf{w}_1||$. The dissimilarity measure $\mathcal{D}_{\mathcal{S}}(\cdot, \cdot)$ thus has the expected behavior in feature space, namely it is high for well separated sets, and it is small for strongly overlapping sets. The pathological case where $\mathbf{w}_1$ and $\mathbf{w}_2$ are collinear and where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ have zero spread is not considered, as it can easily be dealt with by adding some small $\varepsilon > 0$ to $\mathcal{D}_{\mathcal{S}}(\cdot, \cdot)$ denominator.

As the mapping $\phi(\cdot)$ is generally unknown, the computation of $\mathcal{D}_{\mathcal{S}}(\cdot, \cdot)$, which is defined in feature space, is possible only if we can express it as a function of the kernel $k(\cdot, \cdot)$ applied to vectors of the input space $\mathcal{X}$. In [1], we had shown that the arc distance $d_{\text{arc}}(\mathbf{a}, \mathbf{b})$ can be computed in terms of the kernel since

$$d_{\text{arc}}(\mathbf{a}, \mathbf{b}) = \arccos\big(< \mathbf{a}, \mathbf{b} >_{\mathcal{H}}\big) = \arccos\big(1 - \frac{1}{2}||\mathbf{a} - \mathbf{b}||^2_{\mathcal{H}}\big) \qquad (4)$$

Note that the $\arccos$ function is defined properly because the vectors we consider are all located in the same (positive) orthant of $\mathcal{S}$. The inner product in Eq. (4) can be evaluated in terms of the kernel $k(\cdot, \cdot)$ even though $\mathbf{c}_1$ and $\mathbf{c}_2$ are not in $\phi(\mathcal{X})$ since $\mathbf{c}_1 = \mathbf{w}_1 / ||\mathbf{w}_1||_{\mathcal{H}}$ and $\mathbf{c}_2 = \mathbf{w}_1 / ||\mathbf{w}_2||_{\mathcal{H}}$:

$$d_{\text{arc}}(\mathbf{c}_1, \mathbf{c}_2) \;=\; \arccos \frac{\boldsymbol{\alpha}_1^{\mathrm{T}} K_{12} \boldsymbol{\alpha}_2}{\sqrt{\boldsymbol{\alpha}_1^{\mathrm{T}} K_{11} \boldsymbol{\alpha}_1} \sqrt{\boldsymbol{\alpha}_2^{\mathrm{T}} K_{22} \boldsymbol{\alpha}_2}} \qquad (5)$$

where $\boldsymbol{\alpha}_1$ (resp. $\boldsymbol{\alpha}_2$) is the column vector which entries are the α_1^i, $i = 1, \ldots, m_1$ (resp. α_2^i, $i = 1, \ldots, m_2$). The kernel matrix K_{uv}, $(u, v) \in \{1, 2\} \times \{1, 2\}$ has entries at row #i and column #j given by $k(x_u^i, x_v^j)$ where we recall that x_u^i is training vector #i in the set $\boldsymbol{x}_u$. Similar calculation can be applied to $\mathbf{c}_1\widehat{\mathbf{p}}_1$ and $\mathbf{c}_2\widehat{\mathbf{p}}_2$, and yields:

$$d_{\text{arc}}(\mathbf{c}_1, \mathbf{p}_1) \;=\; \arccos \frac{\rho_1}{\sqrt{\boldsymbol{\alpha}_1^{\mathrm{T}} K_{11} \boldsymbol{\alpha}_1}} \qquad (6)$$

Note that $\mathcal{D}_{\mathcal{S}}(\cdot, \cdot)$ can be equivalently defined by replacing $\mathbf{p}_1$ in Eq. (3) by any Margin SV (denoted $\mathbf{x}_1^{\text{MSV}} = \phi(x_1^{\text{MSV}})$) which might not be in $\mathcal{P}$. The arc distance $d_{\text{arc}}(\mathbf{c}_1, \mathbf{p}_1)$ equals $d_{\text{arc}}(\mathbf{c}_1, \mathbf{x}_1^{\text{MSV}})$, because $\mathbf{p}_1$ and $\mathbf{x}_1^{\text{MSV}}$ both are located on the intersection of $\mathcal{S}$ with $\mathcal{W}_1$, and have the same arc distance to $\mathbf{c}_1$. The same reasoning also holds for $\mathbf{p}_2$ and $\mathbf{x}_2^{\text{MSV}}$.

3. CONNECTION WITH $\mathcal{D}_F$ IN INPUT SPACE

In this section, we show that the dissimilarity measure $\mathcal{D}_{\mathcal{S}}(\cdot, \cdot)$ defined in Section 2 is asymptotically equivalent to the Fisher ratio $\mathcal{D}_F(\cdot, \cdot)$ in the case of radial Gaussian distributions. The asymptotic behavior of $\boldsymbol{c}_1$ (resp. $\boldsymbol{c}_2$) is described by the following theorem:

Theorem 1. *Let $\boldsymbol{x}$ be a set of m vectors i.i.d. sampled from a Gaussian pdf $p(\cdot)$ with mean μ and covariance matrix $\sigma^2 \mathbf{I}$. Let $k(\cdot, \cdot)$ be a kernel such that*

$$k(x, x') \;=\; q(||x - x'||^2_{\mathbf{I}, \mathcal{X}}) \;\; \textit{for all} \;\; (x, x') \in \mathcal{X} \times \mathcal{X} \quad (7)$$

[1] These are mild conditions over kernels, which are verified, e.g., by the Gaussian kernel $k(x, y) = \exp\left(-\frac{||x - y||^2_{\mathcal{X}}}{2\sigma^2}\right)$, with $\sigma > 0$.

Then, with probability one, for $m \to \infty$, the center $\mathbf{c}$ yielded by the ν-one class SVM converges to $\boldsymbol{\mu} = \phi(\mu)$.

(The proof for Theorem 1 is not exposed here for brevity reasons). If we assume that the underlying pdfs p_1 and p_2 are Gaussian with means μ_1 and μ_2 and covariance matrices $\sigma_1^2\mathbf{I}$ and $\sigma_2^2\mathbf{I}$, then

$$\mathcal{D}_{\mathcal{S}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \underset{m_1,m_2\to\infty}{\longrightarrow} \frac{d_{\text{arc}}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)}{d_{\text{arc}}(\boldsymbol{\mu}_1, \mathbf{x}_1^{\text{MSV}}) + d_{\text{arc}}(\boldsymbol{\mu}_2, \mathbf{x}_2^{\text{MSV}})} \tag{8}$$

or, equivalently $\mathcal{D}_{\mathcal{S}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ converges to:

$$\frac{g\big(||\mu_1 - \mu_2||^2_{\mathbf{I},\mathcal{X}}\big)}{g\big(||\mu_1 - x_1^{\text{MSV}}||^2_{\mathbf{I},\mathcal{X}}\big) + g\big(||\mu_2 - x_2^{\text{MSV}}||^2_{\mathbf{I},\mathcal{X}}\big)} \tag{9}$$

where $g(u) = \arccos\big(q(u)\big)$. As $||\mu_i - x_i^{\text{MSV}}||^2_{\mathbf{I},\mathcal{X}}$ is asymptotically proportional to σ_i^2 $(i = 1, 2)$, we finally have

$$\mathcal{D}_{\mathcal{S}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \underset{m_1,m_2\to\infty}{\longrightarrow} \frac{g\big(||\mu_1 - \mu_2||^2_{\mathbf{I},\mathcal{X}}\big)}{g\big(\beta\sigma_1^2\big) + g\big(\beta\sigma_2^2\big)} \tag{10}$$

where $\beta > 0$ is a constant. We note that $g(\cdot)$ is a monotonically increasing function with $g(0) = 0$ (in fact, it is shown in Section 4 that $(x, x') \mapsto g\big(||x - x'||^2_{\mathbf{I},\mathcal{X}}\big)$ is a metric in $\mathcal{X}$). This result is quite important because it shows that, for sets sampled according to radial Gaussian distributions, $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ behaves like the standard Fisher ratio $\mathcal{D}_F(\cdot,\cdot)$ in input space.

Remark 2. *Generalisation. This result can be generalised to Gaussian distributions with proportional covariance matrices $\sigma_i^2\Sigma$ $(i = 1, 2)$ when using the Mahalanobis Gaussian kernel $k(x, x') = q(||x - x'||^2_{\Sigma,\mathcal{X}})$. The norm in Eq.* (10) *is then Mahalanobis $||\cdot||_{\Sigma,\mathcal{X}}$ instead of the Euclidean norm $||\cdot||_{\mathbf{I},\mathcal{X}}$.*

As it is defined in feature space where the shape of the mapped quantile estimate is always a segment of the hypersphere $\mathcal{S}$, $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ is well-suited to situations where p_1 and p_2 are not Gaussian, and in particular when $\mathcal{R}^{\mathcal{X}}_{\boldsymbol{x}_1}$ and $\mathcal{R}^{\mathcal{X}}_{\boldsymbol{x}_2}$ have complicated, possibly non-connected, shapes. Simulations on toy examples in Section 5 show that $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ has the expected behavior, namely it is small for similar training sets $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and large when the training sets do not occupy the same region in $\mathcal{X}$. This illustrates the interest of $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ in general situations.

4. MORE DISSIMILARITY MEASURES

The dissimilarity measure $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$, as defined in Section 2, is based on the arc distance in feature space. A key question is: is it possible to define other dissimilarity measures $\mathcal{D}_f(\cdot,\cdot)$ in feature space, with the same interesting properties, but based on other distances in $\mathcal{H}$?

Building on the results of the previous section, Proposition 3 describes how it is possible to generalise $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ to a large class of dissimilarity measures in feature space:

Proposition 3. *Let $k(\cdot,\cdot)$ be a translation invariant kernel, i.e., $k(x, x') = q(||x - x'||_{\Sigma,\mathcal{X}})$. Consider a function $f : \mathbb{R} \to \mathbb{R}$ such that: $1/$ $f(< \mathbf{a}, \mathbf{b} >_{\mathcal{H}})$ is a metric for all $\mathbf{a}, \mathbf{b}$ in the positive orthant of $\mathcal{S}$, $2/$ $f\big(q(||x - x'||_{\mathcal{X}})\big)$ is a metric for all (x, x') in $\mathcal{X}$. Then the dissimilarity measure*

$$\mathcal{D}_f(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{f(< \mathbf{c}_1, \mathbf{c}_2 >_{\mathcal{H}})}{f(< \mathbf{c}_1, \mathbf{p}_1 >_{\mathcal{H}}) + f(< \mathbf{c}_2, \mathbf{p}_2 >_{\mathcal{H}})} \tag{11}$$

behaves asymptotically like the Fisher ratio in the case of Gaussian distributions with covariance matrices proportional to Σ.

The $\arccos$ function used in the definition of $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ verifies these two conditions, as $f(< \cdot, \cdot >_{\mathcal{H}}) = d_{\text{arc}}(\cdot,\cdot)$ and the function $g(u) = \arccos\big(q(u)\big)$ is *metric preserving*. We now recall definition and sufficient conditions about metric preserving functions, before using them to build functions $f(\cdot)$ fitting to the conditions of Proposition 3.

Definition 4. *A function $f : [0,\infty) \to [0,\infty)$ is metric-preserving if for all metric spaces $(\mathcal{A}, d)$, $f\big(d(\cdot,\cdot)\big)$ is a metric.*

The following proposition holds (see, e.g., [5]):

Proposition 5. *If $f : [0;\infty) \to [0;\infty)$ is concave, and such that $f^{-1}(0) = \{0\}$, then $f(\cdot)$ is metric-preserving.*

Weaker conditions do exist, but the sufficient condition given in Proposition 5 is particularly easy to check. If one chooses $f(\cdot)$ such that $u \mapsto f(1 - 1/2u^2)$ and $u \mapsto f(q(u))$ both are metric-preserving, then the required conditions of Proposition 3 are fulfilled. An example is : $f(u) = (1 - u^n)^{1/n}$, which is well defined as the points we consider in feature space are all located in the positive orthant of $\mathcal{S}$, or $f(u) = (1 - u)^{1/n}$.

Yet, the property of metric-preservingness is stronger than what is actually needed: for example, the function $u \mapsto \arccos(1 - u^2/2)$ is not metric preserving, but $\arccos(1 - ||\mathbf{a} - \mathbf{b}||^2_{\mathcal{H}}/2)$ is still a distance between $\mathbf{a}$ and $\mathbf{b}$ (it is the arc distance $d_{\text{arc}}(\mathbf{a}, \mathbf{b})$ on the hypersphere $\mathcal{S}$). This can easily be explained as in Proposition 3, condition 1 (resp. 2) is to be verified for the induced metric $||\cdot - \cdot||_{\mathcal{H}}$ (resp. $||\cdot - \cdot||_{\mathcal{X}}$), and need not be true for *any* metric in $\mathcal{H}$ (resp. $\mathcal{X}$).

In fact, once a function satisfying Proposition 3 is found (possibly a non-metric preserving function), it is possible to build other valid functions $f(\cdot)$ by composition with any metric-preserving functions, such as, e.g., $g_1(u) = au$ with $a > 0$, $g_2(u) = \log_b(1 + u)$, $g_3(u) = \frac{u}{1+u}$, $g_4(u) = max(u, c)$ with $c > 0$, $g_5(u) = u^d$ with $0 < d < 1$. The corresponding dissimilarity measure $\mathcal{D}_f(\cdot,\cdot)$ defined by Eq. (11) has the same asymptotic behavior as $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$; it also shares $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ relevant behavior when the quantiles to estimate have complicated shapes.

5. SIMULATION RESULTS

In this section, we compare on toy examples the behavior of some dissimilarity measures $\mathcal{D}_f(\cdot,\cdot)$ built using sections 2 and 4 to the Fisher ratio. We select three dissimilarity measures $\mathcal{D}_{f_i}$, $i = 1,\ldots,4$, defined using Eq. (11) and $f_1(u) = \arccos(u)$, $f_2(u) = (1-u^2)^{1/2}$ and $f_3(u) = \frac{u}{1+u}$. The kernel we use is the Gaussian kernel $k(x,y) = \exp\left(-||x-y||^2_{\mathcal{X}}/2\sigma^2\right)$ with $\sigma = 2.5$.

The first toy data set is composed of two 2-D Gaussian populations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (with variance $\mathbf{I}$), and with mean $\mu_1(t) = \mu_1 = 0$ and $\mu_2(t) = \mu_1 + 0.25t$, where t denotes the time instant. Both sets size is 30. We see in Fig. 2 that all the dissimilarity measures have the same behavior: they are small when $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are very similar, and large when $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ occupy the same spatial location. In the second toy

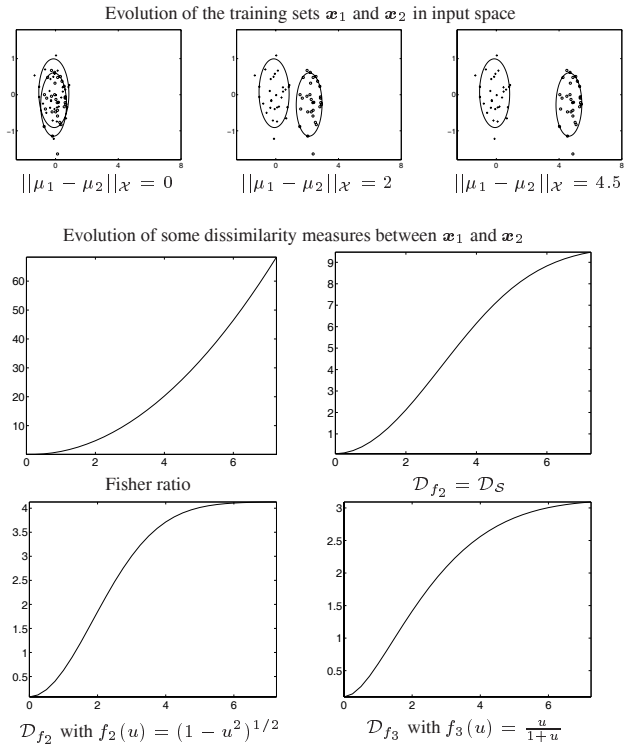

Fig. 2. When the data sets are drawn from Gaussian distributions, the Fisher ratio and dissimilarity measures (plotted as functions of the distance between μ_1 and μ_2) which are built according to Section 4 all have a relevant behavior.

data set, $\boldsymbol{x}_1$ again is sampled from a Gaussian distribution with fixed mean $\mu_1(t) = \mu_1$ and variance $\sigma^2 = \mathbf{I}$; the set $\boldsymbol{x}_2$ is sampled from a mixture of two Gaussian distributions with mean $\mu_2(t) = \mu_1 - 0.25t$ and ${\mu_2}'(t) = \mu_1 + 0.25t$, and with variance $\sigma^2/2$. The size of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is 30. In Fig. 3, all the dissimilarity measures built according to the method proposed in Section 4 have the expected behavior (they increase as the training sets are better separated), whereas the Fisher ratio decreases.

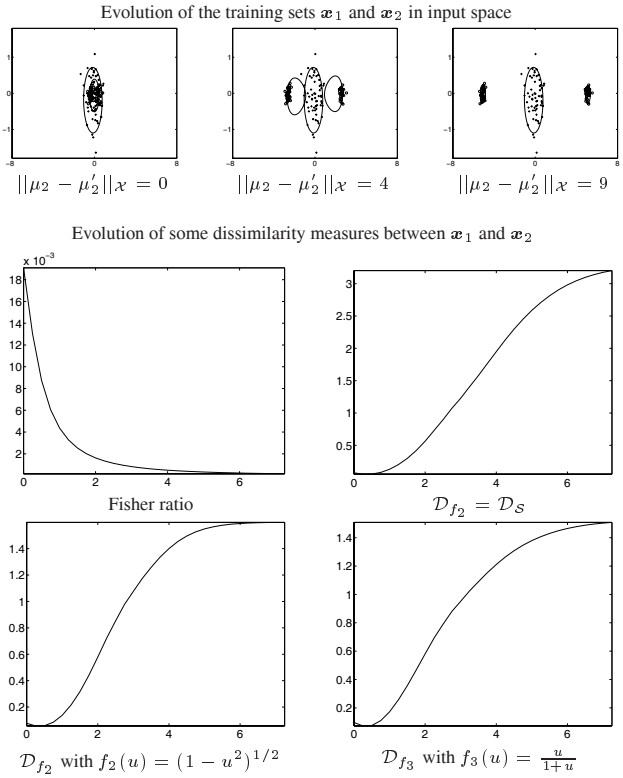

Fig. 3. Though $\mu_1 = \mu_2$, the sets $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ do not occupy the same region in $\mathcal{X}$. The dissimilarity measures built according to Section 4 still have a relevant behavior, but the Fisher ratio does not.

6. CONCLUSION

In this paper, we presented the statistical study of the dissimilarity measure $\mathcal{D}_{\mathcal{S}}(\cdot,\cdot)$ introduced in [1]; in particular, we showed that it generalised the Fisher ratio, and extended its range. It then helped to define a large class of dissimilarity measures $\mathcal{D}_f(\cdot,\cdot)$. Future direction research include the study of further connections with classic dissimilarity measures met in Signal Processing.

7. REFERENCES

[1] Desobry, F. and Davy, M., "Support Vector-Based online Detection of Abrupt Changes," in *Proc. IEEE ICASSP*, Hong Kong, China, Apr. 2003.

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.

[3] Michèle Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, Dec. 1989.

[4] A. Smola and B. Schölkopf, *Learning with Kernels*, MIT press, 2002.

[5] P. Corazza, "Introduction to Metric-Preserving Functions," *Am. Math. Month.*, vol. 104, no. 4, pp. 309–323, Apr. 1999.