Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. 126

Hilbertian Metrics and Positive Definite Kernels on Probability Measures

Matthias Hein[1], Olivier Bousquet[1]

July 2004

[1] Department Schölkopf, email: matthias.hein;olivier.bousquet@tuebingen.mpg.de

Hilbertian Metrics and Positive Definite Kernels on Probability Measures

Matthias Hein and Olivier Bousquet

Abstract. We investigate the problem of defining Hilbertian metrics resp. positive definite kernels on probability measures, continuing the work in [1]. This type of kernels has shown very good results in text classification and has a wide range of possible applications. In this paper we extend the two-parameter family of Hilbertian metrics of Topsøe such that it now includes all commonly used Hilbertian metrics on probability measures. This allows us to do model selection among these metrics in an elegant and unified way. Second we investigate further our approach to incorporate similarity information of the probability space into the kernel. The analysis provides a better understanding of these kernels and gives in some cases a more efficient way to compute them. Finally we compare all proposed kernels in two text and one image classification problem.

1 Introduction

Kernel Methods have shown in the last years that they are one of the best and generally applicable tools in machine learning. Their great advantage is that positive definite (pd) kernels can be defined on every set. Therefore they can be applied to data of any type. Nevertheless in order to get good results the kernel should be adapted as well as possible to the underlying structure of the input space. This has led in the last years to the definition of kernels on graphs, trees and manifolds. Kernels on probability measures also belong to this category but they are already one level higher since they are not defined on the structures directly but on probability measures on these structures. In recent time they have become quite popular due to the following possible applications:

- Direct application on probability measures e.g. histogram data of text [2] and colors [3].
- Given a statistical model for the data one can first fit the model to the data and then use the kernel to compare two fits, see [2, 4]. Thereby linking parametric and non-parametric models.
- Given a bounded probability space $\mathcal{X}$ one can use 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 e.g. sequence data in bioinformatics.

In this paper we consider Hilbertian metrics and pd kernels on $\mathcal{M}_+^1(\mathcal{X})$[1]. In a first section we will summarize the close connection between Hilbertian metrics and pd kernels so that in general statements for one category can be easily transferred to the other one.
We will consider two types of kernels on probability measures. The first one is general covariant that means that arbitrary smooth coordinate transformations of the underlying probability space will have no influence on the kernel. Such kernels can be applied if only the probability measures itself are of interest but not the space they are defined on. We introduce and extend a two parameter family of covariant pd kernels which encompasses all previously used kernels of this type. Despite the great success of these general covariant kernels in text and image classification, they have some shortcomings. For example for some applications we might have a similarity measure resp. a pd kernel on the probability space which we would like to use for the kernel on probability measures. In the second part we further investigate a type of kernel on probability measures which incorporates such a similarity measure, see [1]. This will yield on the one hand a better understanding of this type of kernels and on the other hand gives an efficient way of computing these kernels in some cases. Finally we apply these kernels on two text and one image classification tasks, namely the Reuters, the WebKB data set and the Corel14 data set.

[1] $\mathcal{M}_+^1(\mathcal{X})$ denotes the set of positive measures μ on $\mathcal{X}$ with $\mu(\mathcal{X}) = 1$

2 Hilbertian Metrics versus Positive Definite Kernels

It is a well-known fact that a pd kernel $k(x, y)$ 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) kernels is less well known. Nevertheless this class is of great interest since Schölkopf showed in [5] that all translation invariant kernel methods can also use the bigger class of cpd kernels. Therefore we give a short summary of this type of kernels and their connection to Hilbertian metrics[2].

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

Note that every pd kernel is also cpd. The close connection between the two classes is shown by the following lemma:

Lemma 2.1 *[6] Let k be a kernel defined as $k(x, y) = \hat{k}(x, y) - \hat{k}(x, x_0) - \hat{k}(x_0, y) + \hat{k}(x_0, x_0)$, where $x_0 \in \mathcal{X}$. Then k is pd if and only if $\hat{k}$ is cpd.*

Similar to pd kernels one can also characterize cpd kernels. Namely one write all cpd kernels in the form: $k(x, y) = -\left\|\phi_x - \phi_y\right\|_{\mathcal{H}} + f(x) + f(y)$. The cpd kernels corresponding to Hilbertian (semi)-metrics are characterized by $f(x) = 0$ for all $x \in \mathcal{X}$, whereas if k is pd it follows that $f(x) \sim k(x, x) \geq 0$. We refer to [6, 3.2] and [5] for more on this topic. We also would like to point out that for SVM's the class of Hilbertian (semi)-metrics is in a sense more important than the class of pd kernels. Namely one can show, see [7], that 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 semi-metric. In order to avoid confusion we will in general speak of Hilbertian metrics since, using Lemma 2.1, one can always define a corresponding pd kernel. Nevertheless for the convenience of the reader we will often explicitly state the corresponding pd kernels.

3 Hilbertian Metrics and Positive Definite Kernels on $\mathbb{R}_+$[3]

The class of Hilbertian metrics on probability measures we consider in this paper are all based on a pointwise comparison of the densities $p(x)$ with a Hilbertian metric on $\mathbb{R}_+$. Therefore Hilbertian metrics on $\mathbb{R}_+$ are the basic ingredient of our approach. In principle we could use any Hilbertian metric on $\mathbb{R}_+$, but as we will explain later we require the metric on probability measures to have a certain property. This in turn requires that the Hilbertian metric on $\mathbb{R}_+$ is γ-homogeneous[4]. The class of γ-homogeneous Hilbertian metrics on $\mathbb{R}_+$ was recently characterized by Fuglede:

Theorem 3.1 (Fuglede [8]) *A symmetric function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $d(x, y) = 0 \iff x = y$ is a γ-homogeneous, continuous Hilbertian metric d on $\mathbb{R}_+$ if and only if there exists a (necessarily unique) non-zero bounded measure $\rho \geq 0$ on $\mathbb{R}_+$ such that d^2 can be written as*

$$d^2(x, y) = \int_{\mathbb{R}_+} \left| x^{(\gamma + i\lambda)} - y^{(\gamma + i\lambda)} \right|^2 d\rho(\lambda) \tag{1}$$

Using Lemma 2.1 we define the corresponding class of pd kernels on $\mathbb{R}_+$, where we choose $x_0 = 0$. We will see later that this corresponds to choosing the zero-measure as origin of the RKHS.

Corollary 3.1 *A symmetric function $k : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $k(x, x) = 0 \iff x = 0$ is a 2γ-homogeneous continuous pd kernel 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 k can be written as*

$$k(x, y) = \int_{\mathbb{R}} x^{(\gamma + i\lambda)} y^{(\gamma - i\lambda)} \, d\kappa(\lambda) \tag{2}$$

[2] A (semi)-metric $d(x, y)$ (A semi-metric $d(x, y)$ fulfills the conditions of a metric except that $d(x, y) = 0$ does not imply $x = y$.) is called Hilbertian if one can embed the (semi)-metric space $(\mathcal{X}, d)$ isometrically into a Hilbert space. A (semi)-metric d is Hilbertian if and only if $-d^2(x, y)$ is cpd. That is a classical result of Schoenberg.

[3] $\mathbb{R}_+$ is the positive part of the real line with 0 included

[4] A symmetric function k is γ-homogeneous if $k(c\,x, c\,y) = c^\gamma k(x, y)$ for all $c \in \mathbb{R}_+$

Proof: If k has the form given in (2), then it is obviously 2γ-homogeneous and since $k(x,x) = x^{2\gamma}\kappa(\mathbb{R})$ we have $k(x,x) = 0 \iff x = 0$. The other direction follows by first noting that $k(0,0) = \langle v_0, v_0\rangle = 0$ and then by applying theorem 3.1, where κ is the symmetrized version of ρ around the origin, together with lemma 2.1 and

$$k(x,y) = \langle v_x, v_y\rangle = \frac{1}{2}\left(-d^2(x,y) + d^2(x,0) + d^2(y,0)\right).$$

□

At first glance Theorem 3.1, though mathematical beautiful, seems to be not very helpful from the viewpoint of applications. But as we will show in the section on structural pd kernels on $\mathcal{M}_+^1(\mathcal{X})$ this result will allow to compute this class of kernels very efficiently.

Recently Topsøe and Fuglede proposed an interesting two-parameter family of Hilbertian metrics on $\mathbb{R}_+$ [9, 8]. We will now extend the parameter range of this family. This will allow us in the next section to recover all previously used Hilbertian metrics on probability measures from this family.

Theorem 3.2 *The function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined as:*

$$d^2_{\alpha|\beta}(x,y) = \frac{2^{\frac{1}{\beta}}\left(x^\alpha + y^\alpha\right)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}}\left(x^\beta + y^\beta\right)^{\frac{1}{\beta}}}{2^{\frac{1}{\alpha}} - 2^{\frac{1}{\beta}}} \tag{3}$$

is a $1/2$-homogeneous Hilbertian metric on $\mathbb{R}_+$, if $\alpha \in [1,\infty]$, $\beta \in [\frac{1}{2}, \alpha]$ or $\beta \in [-\infty, -1]$. Moreover the pointwise limit for $\alpha \to \beta$ is given as:

$$\begin{aligned}\lim_{\alpha\to\beta} d^2_{\alpha|\beta}(x,y) &= \frac{\beta^2 2^{1/\beta}}{\log(2)}\frac{\partial}{\partial\beta}\left(\frac{x^\beta + y^\beta}{2}\right)^{(1/\beta)} \\ &= \frac{\left(x^\beta + y^\beta\right)^{\frac{1}{\beta}}}{\log(2)}\left[\frac{x^\beta}{x^\beta + y^\beta}\log\left(\frac{2x^\beta}{x^\beta + y^\beta}\right) + \frac{y^\beta}{x^\beta + y^\beta}\log\left(\frac{2y^\beta}{x^\beta + y^\beta}\right)\right]\end{aligned}$$

Note that $d^2_{\alpha|\beta} = d^2_{\beta|\alpha}$. We need the following lemmas in the proof:

Lemma 3.1 *[6, 2.10] If $k : \mathcal{X}\times\mathcal{X}$ is cpd and $k(x,x) \le 0$, $\forall x \in \mathcal{X}$ then k^γ is also cpd for $0 < \gamma \le 1$.*

Lemma 3.2 *If $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_-$[5] is cpd, then $-1/k$ is pd.*

Proof: It follows from theorem 2.3 in [6] that if $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_-$ is cpd, then $1/(t-k)$ is pd for all $t > 0$. The pointwise limit of a sequence of cpd resp. pd kernels is cpd resp. pd if the limit exists, see e.g. [10]. Therefore $\lim_{t\to 0} 1/(t-k) = -1/k$ is positive definite if k is strictly negative. □

We can now prove Theorem 3.2:

Proof: The proof for the symmetry, the limit $\alpha \to \beta$ and the parameter range $1 \le \alpha \le \infty$, $1/2 \le \beta \le \alpha$ can be found in [8]. We prove that $-d^2_{\alpha|\beta}$ is cpd for $1 \le \alpha \le \infty$, $-\infty \le \beta \le -1$. First note that $k(x,y) = -(f(x) + f(y))$ is cpd on $\mathbb{R}_+$, for any function $f : \mathbb{R}_+ \to \mathbb{R}_+$ and satisfies $k(x,y) \le 0$, $\forall\, x, y \in \mathcal{X}$. Therefore by Lemma 3.1, $-(x^\alpha + y^\alpha)^{1/\alpha}$ is cpd for $1 \le \alpha < \infty$. The pointwise limit $\lim_{\alpha\to\infty} -(x^\alpha + y^\alpha)^{1/\alpha} = -\max\{x,y\}$ exists, therefore we can include the limit $\alpha = \infty$. Next we consider $k(x,y) = -(x+y)^{1/\beta}$ for $1 \le \beta \le \infty$ which is cpd as we have shown and strictly negative if we restrict k to $\{x \in \mathbb{R} \,|\, x > 0\}$. Then all conditions for lemma 3.2 are fulfilled, so that $k(x,y) = (x+y)^{-1/\beta}$ is pd. But then also $k(x,y) = (x^{-\beta} + y^{-\beta})^{-1/\beta}$ is pd. Moreover k can be continuously extended to 0 by $k(x,y) = 0$ for $x = 0$ or $y = 0$. Multiplying the first part with $(2^{(1/\alpha - 1/\beta)} - 1)^{-1}$ and the second one with $(1 - 2^{(1/\beta - 1/\alpha)})^{-1}$ and adding them gives the result. □

4 Covariant Hilbertian Metrics on $\mathcal{M}_+^1(\mathcal{X})$

In this section we define Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ by comparing the densities pointwise with a Hilbertian metric on $\mathbb{R}_+$ and integrating these distances over $\mathcal{X}$. Since densities can only be defined with respect to a dominating measure[6] our definition will at first depend on the choice of the dominating measure. This dependence would restrict the applicability of our approach. For example if we had $\mathcal{X} = \mathbb{R}^n$ and chose μ to be the Lebesgue

[5] $\mathbb{R}_- = \{x \in \mathbb{R} \,|\, x < 0\}$

[6] A measure μ dominates a measure ν if $\mu(E) > 0$ whenever $\nu(E) > 0$ for all sets $E \subset \mathcal{X}$. In $\mathbb{R}^n$ the dominating measure μ is usually the Lebesgue measure.

measure, then we could not deal with Dirac measures δ_x since they are not dominated by the Lebesgue measure. Therefore we construct the Hilbertian metric such that it is independent of the dominating measure. This justifies the term covariant since independence from the dominating measure also yields invariance from arbitrary one-to-one coordinate transformations. In turn this also implies that all structural properties of the probability space will be ignored so that the metric on $\mathcal{M}_+^1(\mathcal{X})$ only depends on the probability measures. As an example take the color histograms of images. Covariance here means that the choice of the underlying color space say RGB, HSV or CIE Lab does not influence our metric, since these color spaces are all related by one-to-one transformations. Note however that in practice the results will usually slightly differ due to different discretizations of the color space.
In order to simplify the notation we define $p(x)$ to be the Radon-Nikodym derivative $(dP/d\mu)(x)$ [7] of P with respect to the dominating measure μ.

Proposition 4.1 *Let P and Q be two probability measures on $\mathcal{X}$, μ an arbitrary dominating measure*[8] *of P and Q and $d_{\mathbb{R}_+}$ a $1/2$-homogeneous Hilbertian metric on $\mathbb{R}_+$. Then $D_{\mathcal{M}_+^1(\mathcal{X})}$ defined as*

$$D^2_{\mathcal{M}_+^1(\mathcal{X})}(P,Q) := \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(p(x),q(x))d\mu(x)\,, \tag{4}$$

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

For a proof of this proposition, see [1]. We can now apply this principle of building covariant Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ and use the family of 1/2-homogeneous Hilbertian metrics $d^2_{\alpha|\beta}$ on $\mathbb{R}_+$ from the previous section. This yields as special cases the following well-known measures on probability distributions.

$$\begin{aligned}
D^2_{\frac{1}{2}|1}(P,Q) &= \int_{\mathcal{X}} (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x), \quad D^2_{\infty|1}(P,Q) = \int_{\mathcal{X}} |p(x)-q(x)|d\mu(x) \\
D^2_{1|1}(P,Q) &= \frac{1}{\log(2)} \int_{\mathcal{X}} p(x) \log\left(\frac{2p(x)}{p(x)+q(x)}\right) + q(x)\log\left(\frac{2q(x)}{p(x)+q(x)}\right) d\mu(x), \\
D^2_{1|-1}(P,Q) &= \int_{\mathcal{X}} \frac{(p(x)-q(x))^2}{p(x)+q(x)} d\mu(x)
\end{aligned}$$

$D_{\frac{1}{2}|1}$ is the Hellinger distance, $D^2_{1|-1}$ is the symmetric χ^2-measure, $D^2_{\infty|1}$ is the total variation and $D^2_{1|1}$ is the Jensen-Shannon divergence. The Hellinger metric is well known in the kernel community and was for example used in [4], the symmetric χ^2-metric was for a long time wrongly assumed to be pd and is new in this family due to our extension of $d^2_{\alpha|\beta}$ to negative values of β. The total variation was implicitly used in SVM's through a pd counterpart which we will give below. The Jensen-Shannon divergence is very interesting since it is a symmetric and smoothed variant of the Kullback-Leibler divergence. Instead of the work in [11] where they have a heuristic approach to get from the Kullback-Leibler divergence to a pd matrix, the Jensen-Shannon divergence is a theoretically sound alternative. Note that the family $d^2_{\alpha|\beta}$ is designed in such a way that the maximal distance of $D^2_{\alpha|\beta}$ is 2, $\forall\, \alpha, \beta$. For completeness we also give the corresponding pd kernels on $\mathcal{M}_+^1(\mathcal{X})$, where we take in Lemma 2.1 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 the whole framework presented in this paper can easily be extended to all finite, positive measures on $\mathcal{X}$. For this set the zero measure is a natural choice of the origin.

$$\begin{aligned}
K_{\frac{1}{2}|1}(P,Q) &= \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x), \qquad K_{1|-1}(P,Q) = \int_{\mathcal{X}} \frac{p(x)q(x)}{p(x)+q(x)} d\mu(x) \\
K_{1|1}(P,Q) &= \frac{-1}{\log(2)} \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{p(x)+q(x)}\right) + q(x)\log\left(\frac{q(x)}{p(x)+q(x)}\right) d\mu(x), \\
K_{\infty|1}(P,Q) &= \int_{\mathcal{X}} \min\{p(x),q(x)\} d\mu(x)
\end{aligned}$$

The astonishing fact is that we find the four previously used Hilbertian metrics resp. pd kernels on $\mathcal{M}_+^1(\mathcal{X})$ as special cases of a two-parameter family of Hilbertian metrics resp. pd kernels on $\mathcal{M}_+^1(\mathcal{X})$. Due to the symmetry of $d^2_{\alpha|\beta}$ (which implies symmetry of $D^2_{\alpha|\beta}$) we can even see all of them as special cases of the family restricted to

[7] In case of $\mathcal{X} = \mathbb{R}^n$ and when μ is the Lebesgue measure we can think of $p(x)$ as the normal density function.
[8] Such a dominating measure always exists take e.g. $M = (P+Q)/2$

$\alpha = 1$. This on the one hand shows the close relation of these metrics and on the other hand gives us the opportunity to do model selection in a one-parameter family. So that we can treat the four known cases and intermediate ones in a very elegant way.

5 Structural Positive Definite Kernels

The covariant Hilbertian metrics proposed in the last section have the advantage that they only compare the probability measures, thereby ignoring all structural properties of the probability space. On the other hand there exist cases where we have a reasonable similarity measure on the probability space, which we would like to be incorporated into the metric. As an example this is helpful when we compare probability measures with disjoint support, since for the covariant metrics disjoint measures have always maximal distance, irrespectively how "close" or "far" their support is. Obviously if our training set consists only of disjoint measures learning is not possible with covariant metrics. We have proposed in [1] a positive definite kernel which incorporates a given similarity measure, namely a pd kernel, on the probability space. The only disadvantage is that this kernel is not invariant with respect to the dominating measure. That means we can only define it for the subset $\mathcal{M}_+^1(\mathcal{X}, \mu) \subset \mathcal{M}_+^1(\mathcal{X})$, that are the measures dominated by μ. On the other hand if one can define a kernel on $\mathcal{X}$, then one can build e.g. by the induced semi-metric a uniform measure μ on $\mathcal{X}$, which is then a natural choice for the dominating measure.

Theorem 5.1 (Structural Kernel) *Let k be a bounded PD kernel on $\mathcal{X}$ and $\hat{k}$ a bounded PD kernel on $\mathbb{R}_+$. Then*

$$K(P, Q) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, \hat{k}(p(x), q(y))\, d\mu(x)\, d\mu(y) \tag{5}$$

is a pd kernel on $\mathcal{M}_+^1(\mathcal{X}, \mu) \times \mathcal{M}_+^1(\mathcal{X}, \mu)$.

The structural kernel generalizes previous work done by Suquet, see [12], where $\hat{k}(p(x), q(y)) = p(x)q(y)$. The advantage of this choice for $\hat{k}$ is that $K(P, Q)$ is independent of the dominating measure. In fact it is easy to see that among the family of structural kernels $K(P, Q)$ this kernel is the only one independent of the dominating measure.

The structural kernel has the disadvantage that the computational cost increases dramatically, since we have to integrate twice over $\mathcal{X}$. An implementation seems therefore only be possible for either very located probability measures or a sharply concentrated similarity function e.g. a compactly supported radial basis function on $\mathbb{R}^n$. We will now give an equivalent representation of this kernel which provides a better understanding and shows that in some cases the computational cost can be reduced roughly to that of the covariant kernel.

Proposition 5.1 *The structural kernel $K(P, Q)$ can be equivalently written as the inner product in $L_2(T \times \mathbb{R}, \omega \otimes \kappa)$:*

$$K(P, Q) = \int_T \int_{\mathbb{R}} \phi_P(t, \lambda) \overline{\phi_Q(t, \lambda)}\, d\kappa(\lambda)\, d\omega(t)$$

for some set T with the feature map:

$$\phi : \mathcal{M}_+^1(\mathcal{X}, \mu) \to L_2(T \times \mathbb{R}, \omega \otimes \kappa), \quad P \to \phi_P(t, \lambda) = \int_{\mathcal{X}} \Gamma(x, t) p(x)^{(1+i\lambda)} d\mu(x).$$

Proof: First note that one can write every pd kernel in the form : $k(x, y) = \langle \Gamma(x, \cdot), \Gamma(y, \cdot) \rangle_{L_2(T,\omega)} = \int_T \Gamma(x, t) \overline{\Gamma(y, t)} d\omega(t)$, where $\Gamma(x, \cdot) \in L_2(T, \mu)$ for each $x \in \mathcal{X}$. In general the space T is very big, since one can show that such a representation always exists in $L_2(\mathbb{R}^{\mathcal{X}}, \mu)$, see e.g. [13]. For the kernel $\hat{k}$ we have such a representation on $L_2(\mathbb{R}, \kappa)$ from Corollary 3.1. Since for any finite measure space $(\mathcal{Y}, \mu)$ one has $L_2(\mathcal{Y}, \mu) \subset L_1(\mathcal{Y}, \mu)$ we can apply Fubini's theorem and interchange the integration order. The definition of the feature map then follows easily. □

This representation has several advantages. First the functions $\Gamma(x, t)$ tell us which properties of the measure P are used. Second in the case where T is of the same or smaller size than $\mathcal{X}$ we can decrease the computation cost, since we now have to do only an integration over $T \times \mathbb{R}$ instead of an integration over $\mathcal{X} \times \mathcal{X}$. Last this representation is a good starting point if one wants to approximate the structural kernel. Since any discretization of $T, \mathbb{R}$, or $\mathcal{X}$ or integration over smaller subsets, will nevertheless give a pd kernel in the end.

We illustrate this result with a simple example. We take $\mathcal{X} = \mathbb{R}^n$ and $k(x, y) = k(x - y)$ to be a translation invariant kernel, furthermore we take $\hat{k}(p(x), q(y)) = p(x)q(y)$. The characterization of translation invariant kernels is a classical result due to Bochner:

Table 1: The table shows the test errors for the covariant and structural kernels for each data set. The first column shows the test error and the α-value of the kernel with the best cross-validation error over the family $d^2_{\alpha|1}$ and penalty constants C. The next four columns provide the results for the special cases in $d^2_{\alpha|1}$. The last column gives the test error for the non-covariant L_2-kernel as a comparison to covariant kernels. For the structural kernel we already have the L_2-kernel as $K_{1|1/2}$ since we do the transformation $p(x) \to p(x)^2$.

		Best α		$\alpha = -1$	$\alpha = \frac{1}{2}$	$\alpha = 1$	$\alpha = \infty$	L_2
Reuters	cov	**1.36**	-1	**1.36**	1.42	**1.36**	1.79	1.98
	str	**1.91**	1/2	2.04	1.91	1.98	**1.91**	–
WebKB	cov	4.88	16	4.76	**4.40**	4.52	4.64	7.85
	str	**6.18**	1	6.54	6.54	**6.18**	6.54	–
Corel14	cov	12.86	-1	12.86	20.71	15.71	**12.50**	30.00
	str	**20.00**	-1	20.00	28.57	**20.00**	20.36	–

Theorem 5.2 *A continuous function $k(x, y) = k(x - y)$ is pd on $\mathbb{R}^n$ if and only if $k(x - y) = \int_{\mathbb{R}^n} e^{i\langle t, x-y\rangle} d\omega(t)$, where μ is a finite non-negative measure on $\mathbb{R}^n$*

Obviously we have in this case $T = \mathbb{R}^n$. Then the above proposition tells us that we are effectively computing the following feature vector for each P, $\phi_P(t) = \int_{\mathbb{R}^n} e^{i\langle x,t\rangle} p(x) d\mu(x) = E_P e^{i\langle x,t\rangle}$. Finally the structural kernel can in this case be equivalently written as $K(P, Q) = \int_{\mathbb{R}^n} E_P e^{i\langle x,t\rangle} E_Q e^{i\langle x,t\rangle} d\omega(t)$. That means the kernel is in this case nothing else than the inner product between the characteristic functions of the measures in $L_2(\mathbb{R}^n, \omega)$[9]. Moreover the computational cost has decreased dramatically, since we only have to integrate over $T = \mathbb{R}^n$ instead of $\mathbb{R}^n \times \mathbb{R}^n$. Therefore in this case the kernel computation has the same computational complexity as in the case of the covariant kernels. The calculation of the features, here the characteristic functions, can be done as a preprocessing step for each measure.

6 Experiments

We evaluated the quality of the proposed metrics/kernels in three classification tasks. First the *Reuters* text data set. Here the documents are represented as term histograms. Following [2] we used the five most frequent classes *earn*, *acq*, *moneyFx*, *grain* and *crude*. We excluded documents that belong to more than one of theses classes. This resulted in a data set with 8085 examples of dimension 18635. Second the *WebKB* web pages data set. The documents are also represented as term histograms. We used the four most frequent classes *student*, *faculty*, *course* and *project*. 4198 documents remained each of dimension 24212, see [2]. For both text data sets we took the correlation matrix in the bag of documents representation as a pd kernel on the probability space of terms. Third the *Corel* image data base. We chose the categories Corel14 as in [3]. The Corel14 has 14 classes each with 100 examples. As reported in [3] the classes are very noisy, especially the bear and polar bear classes. We performed a uniform quantization of each image in the RGB color space, taking 16 bins per color, yielding 4096 dimensional histograms. For the Corel14 data set we took as a similarity measure on the euclidean RGB color space, the compactly supported RBF kernel $k(x, y) = (1 - \|x - y\| / d_{max})^2_+$, with $d_{max} = 0.15$.
All data sets were split into a training (80%) and a test (20%) set. The multi-class problem was solved by one-vs-all with SVM's. We did all experiments with the family $d^2_{\alpha|1}$, once for the covariant Hilbertian metrics and once for the structural kernels. As stated in section 3 the squared metrics $d^2_{\alpha|1}$ are one-homogeneous. For the structural kernels we plugged the squared densities into the pd counterparts of $d^2_{\alpha|1}$ yielding a two-homogeneous family of pd kernels as desired for the structural kernels.
For the penalty constant we chose from $C = 10^k, k = -1, 0, 1, 2, 3, 4$ and for α from $\alpha = 1/2, \pm 1, \pm 2, \pm 4, \pm 16, \infty$. The case $\alpha = -\infty$ coincides with $\alpha = \infty$. In order to find the best parameters for C and α we performed 10-folds cross validation. For the best parameter among α and C we evaluated the test error. In order to show the results of the previously used kernels we also give the test errors for the kernels corresponding to $\alpha = -1, 1/2, 1, \infty$. The results are shown in table 1. The test errors show that model selection among the family $d^2_{\alpha|1}$ gives usually close to optimal results. Besides the test errors of the covariant kernels are

[9] Note that ω is not the Lebesgue measure.

always better than the non-covariant L_2. This indicates that for these problems only the probabilities as "geometric" objects mattered. The structural kernels were always worse than the covariant ones. This could have two reasons. Either all the similarity measures we use on the probability spaces are not suited to the problem or all the considered problems can be solved more efficiently using only the probabilities. Nevertheless we would like to note that the structural kernels with the L_2-kernel, that is $K_{1|\frac{1}{2}}$ were always better than the L_2-kernel alone. From this point of view the similarity information helped to improve the L_2-kernel.

7 Conclusion

We went on with the work started in [1] on Hilbertian metrics resp. pd kernels on $\mathcal{M}_+^1(\mathcal{X})$. We could extend a family of Hilbertian metrics proposed by Topsøe, so that the all previously used measures on probabilities are included now in this family. Moreover we gave an equivalent representation for our structural kernels on $\mathcal{M}_+^1(\mathcal{X})$, which on the one hand gives a more direct access to the way they capture structure of the probability measures and on the other hand gives in some cases a more efficient way to compute it. Finally we could show that doing model selection in $d^2_{\alpha|1}$ gives almost optimal results for covariant and structural kernels. In all three tasks the covariant kernels were better than the structural ones. It remains an open problem if one can improve the results of the structural kernels by taking other similarity kernels on the probability space.

Acknowledgements

We would like to thank Guy Lebanon for kindly providing us with the WebKB and Reuters data set in preprocessed from. Furthermore we are thankful to Flemming Topsøe and Bent Fuglede for providing us with preprints of their papers [9, 8]. Financial support was provided by the PASCAL network.

References

[1] M. Hein, T. N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in SVM's. Accepted at DAGM, 2004.

[2] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. Technical Report CMU-CS-04-101, School of Computer Science, Carnegie Mellon University, Pittsburgh, 2004.

[3] O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10:1055–1064, 1999.

[4] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *16th Annual Conference on Learning Theory (COLT)*, 2003.

[5] B. Schölkopf. The kernel trick for distances. *NIPS*, 13, 2000.

[6] J. P. R. Christensen C. Berg and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, New York, 1984.

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

[8] B. Fuglede. Spirals in Hilbert space. With an application in information theory. To appear in Expositiones Mathematicae, 2004.

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

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

[11] P. J. Moreno, P. P. Hu, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *NIPS*, 16, 2003.

[12] C. Suquet. Distances euclidiennes sur les mesures signées et application à des théorèmes de Berry-Esséen. *Bull. Belg. Math. Soc. Simon Stevin*, 2:161–181, 1995.

[13] S. Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, Cambridge, 1997.