Hilbertian Metrics on Probability Measures and their Application in SVM's

Matthias Hein, Thomas Navin Lal and Olivier Bousquet

Max Planck Institute for Biological Cybernetics
Spemannstr. 38
72076 Tuebingen, Germany
{matthias.hein, navin.lal, olivier.bousquet}@tuebingen.mpg.de

Abstract. In this article we investigate the field of Hilbertian metrics on probability measures. Since they are very versatile and can therefore be applied in various problems they are of great interest in kernel methods. Quit recently Topsøe and Fuglede introduced a family of Hilbertian metrics on probability measures. We give basic properties of the Hilbertian metrics of this family and other used metrics in the literature. Then we propose an extension of the considered metrics which incorporates structural information of the probability space into the Hilbertian metric. Finally we compare all proposed metrics in an image and text classification problem using histogram data.

1 Introduction

Recently the need for specific design of kernels for a given data structure has been recognized by the kernel community. One type of structured data are probability measures $\mathcal{M}_+^1(\mathcal{X})$[1]on a probability space $\mathcal{X}$. The following examples show the wide range of applications of this class of kernels:

- Direct application on probability measures e.g. histogram data [1].
- Having 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 [5, 4].
- Given a bounded probabiliy space $\mathcal{X}$ one can use the kernel to compare sets in that space, by putting e.g. the uniform measure on each set.

In this article we study instead of positive definite (PD) kernels the more general class of conditionally positive definite (CPD) kernels. Or to be more precise we concentrate on Hilbertian metrics, that are metrics d which can be isometrically embedded into a Hilbert space, that is $-d^2$ is CPD. This choice can be justified by the fact that the support vector machine (SVM) only uses the metric information of the CPD[2] kernel, see [3], and that every CPD kernel is generated by a Hilbertian metric.
We propose a general method to build Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ from Hilbertian metrics on $\mathbb{R}_+$. Then we completely characterize the Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ which are invariant under the change of the dominating measure

[1] $\mathcal{M}_+^1(\mathcal{X})$ denotes the set of positive measures μ on $\mathcal{X}$ with $\mu(\mathcal{X}) = 1$
[2] Note that every PD kernel is a CPD kernel.

using results of Fuglede. As a next step we introduce a new family of Hilbertian metrics which incorporates similarity information of the probability space. Finally we support the theoretical analysis by two experiments. First we compare the performance of the basic metrics on probability measures in an image and text classification problem. Second we do the image classification problem again but now using similarity information of the color space.

2 Hilbertian Metrics

An interesting subclass of metrics is the class of Hilbertian metrics, that are metrics which can be isometrically embedded into a Hilbert space. In order to characterize this subclass of metrics, we first introduce the following function class:

Definition 1. *A real valued function k on $\mathcal{X} \times \mathcal{X}$ is positive definite (PD) (resp. conditionally positive definite (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 $\sum_{i}^{n} c_i = 0$).*

The following theorem describes the class of Hilbertian metrics:

Theorem 1 (Schoenberg [6]). *A metric space $(\mathcal{X}, d)$ can be embedded isometrically into a Hilbert space if and only if $-d^2(x, y)$ is CPD.*

What is the relevance of this notion for the SVM? Schölkopf showed that the class of CPD kernels can be used in SVM's due to the translation invariance of the maximal margin problem in the RKHS, see [7]. Furthermore it is well known that the maximal margin problem is equivalent to the optimal separation of the convex hulls of the two classes. This was used in [3] to show that the properties of the SVM only depend on the Hilbertian metric. That is all CPD kernels are generated by a Hilbertian metric $d(x, y)$ through $k(x, y) = -d^2(x, y) + g(x) + g(y)$ where $g : \mathcal{X} \to \mathbb{R}$ and the solution of the SVM only depends on the Hilbertian metric $d(x, y)$.

3 Hilbertian Metrics on Probability Measures

It would be very ambitious to address the question of all possible Hilbertian metrics on probability measures. Instead we restrict ourselves to a special family. Nevertheless this special case encompasses almost all measures previously used in the machine learning community. In the first section we use recent results of Fuglede and Topsøe, which describe all α-homogeneous[3], continuous Hilbertian (semi)-metrics on $\mathbb{R}_+$[4]. Using these results it is straightforward to characterize all Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ of a certain from. In the second part we extend the framework and incorporate similarity information of $\mathcal{X}$.

[3] That means $d^2(c\,p, c\,q) = c^\alpha d^2(p, q)$ for all $c \in \mathbb{R}_+$

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

3.1 Hilbertian Metrics on Probability Measures derived from Hilbertian metrics on $\mathbb{R}_+$

For simplicity we will first only treat the case of discrete probability measures on $D = \{1, 2, \ldots, N\}$, where $1 \leq N \leq \infty$. Given a Hilbertian metric d on $\mathbb{R}_+$ it is easy to see that the metric $d_{\mathcal{M}_+^1}$ 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 metric on a probability space $\mathcal{X}$. In order to simplify the notation we define $p(x)$ to be the Radon-Nikodym derivative $(dP/d\mu)(x)$ [5] of P with respect to the dominating measure μ.

Proposition 1. *Let P and Q be two probability measures on $\mathcal{X}$, μ an arbitrary dominating measure[6] 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*

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

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

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

$$\int_{\mathcal{X}} d^2_{\mathbb{R}_+}(\frac{dP}{d\mu}, \frac{dQ}{d\mu}) d\mu = \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(\frac{dP}{d\nu}\frac{d\nu}{d\mu}, \frac{dQ}{d\nu}\frac{d\nu}{d\mu})\frac{d\mu}{d\nu} d\nu = \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(\frac{dP}{d\nu}, \frac{dQ}{d\nu}) d\nu$$

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

It is in principle very easy to construct Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ using an arbitrary Hilbertian metric on $\mathbb{R}_+$ and plugging it into the definition (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. For example the euclidean norm on $\mathbb{R}_+$ will yield a metric on $\mathcal{M}_+^1(\mathcal{X})$ but it is not invariant with respect to arbitrary coordinate transformations. We think that this could be the reason why the naive application of the linear or the Gaussian kernel yields worse results than Hilbertian metrics resp. kernels which are invariant, see [1, 5].
Quite recently Fuglede completely characterized the class of homogeneous Hilbertian metrics on $\mathbb{R}_+$. The set of all $1/2$-homogeneous Hilbertian metrics on $\mathbb{R}_+$ characterizes then all invariant Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ of the form (1).

[5] In $\mathbb{R}^n$ the dominating measure μ is usually the Lebesgue measure. In this case we can think of $p(x)$ as the normal density function.

[6] Such a dominating measure always exists take, e.g. $M = (P+Q)/2$.

Theorem 2 (Fuglede [2]). *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 $\mu \geq 0$ on $\mathbb{R}_+$ such that d^2 can be written as*

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

Topsøe proposed the following family of 1/2-homogeneous Hilbertian metrics.

Theorem 3 (Topsøe,Fuglede). *The function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined as:*

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

is a 1/2-homogeneous Hilbertian metric on $\mathbb{R}_+$, if $1 \leq \alpha \leq \infty$, $1/2 \leq \beta \leq \alpha$. Moreover $-d^2$ is strictly CPD except when $\alpha = \beta$ or $(\alpha,\beta) = (1, 1/2)$.

Obviously one has $d^2_{\alpha|\beta} = d^2_{\beta|\alpha}$. Abusing notation we denote in the following the final metric on $\mathcal{M}^1_+(\mathcal{X})$ generated using (1) by the same name $d^2_{\alpha|\beta}$. The following special cases are interesting:

$$d^2_{\infty|1}(P,Q) = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)| d\mu(x), \quad d^2_{\frac{1}{2}|1}(P,Q) = \frac{1}{4}\int_{\mathcal{X}} (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x)$$

$$d^2_{1|1}(P,Q) = \frac{1}{2}\int_{\mathcal{X}} p(x)\log\left(\frac{2p(x)}{p(x)+q(x)}\right) + q(x)\log\left(\frac{2q(x)}{p(x)+q(x)}\right) d\mu(x)$$

$d^2_{\infty|1}$ is the total variation[7]. $d^2_{\frac{1}{2}|1}$ is the square of the Hellinger distance. It is induced by the positive definite Bhattacharyya kernel, see [4]. $d^2_{1|1}$ can be derived by a limit process, see [2]. It was not used in the machine learning literature before. Since it is the proper version of a Hilbertian metric which corresponds to the Kullback-Leibler divergence $D(P||Q)$, it is especially interesting. In fact it can be written with $M = (P+Q)/2$ as $d^2_{1|1}(P,Q) = \frac{1}{2}\left(D(P||M) + D(Q||M)\right)$. For an interpretation from information theory, see [9]. We did not consider other metrics from this family since they all have similar properties as we show later. Another 1/2-homogeneous Hilbertian metric previously used in the machine learning literature is the modified χ^2-distance : $d^2_{\chi^2}(P,Q) = \sum_{i=1}^n \frac{(p_i-q_i)^2}{p_i+q_i}$. $d^2_{\chi^2}$ is not PD, as often wrongly assumed in the literature, but CPD. See [8] for a proof and also for the interesting upper and lower bounds on the considered metrics:

$$d^2_{\frac{1}{2}|1} \leq d^2_{\alpha|\beta} \leq d^2_{\infty|1} \leq \frac{1}{2} d_{\frac{1}{2}|1}, \quad 4d^2_{\frac{1}{2}|1} \leq d^2_{\chi^2} \leq 8d^2_{\frac{1}{2}|1}, \quad 2d^4_{\infty|1} \leq d^2_{\chi^2} \leq 2d^2_{\infty|1}$$

In order to compare all different kinds of metrics resp. kernels on $\mathcal{M}^1_+(\mathcal{X})$ which were used in the kernel community, we also considered the geodesic distance of

[7] This metric was implicitly used before, since it is induced by the positive definite kernel $k(P,Q) = \sum_{i=1}^n \min(p_i, q_i)$.

the multinomial statistical manifold used in [5]: $d_{geo}(P,Q) = \arccos(\sum_{i=1}^{N} \sqrt{p_i q_i})$. We could not prove that it is Hilbertian. In [5] they actually use the kernel $\exp(-\lambda d^2(P,Q))$ as an approximation to the first order parametrix of the heat kernel of the multinomial statistical manifold. Despite the mathematical beauty of this approach, there remains the problem that one can only show that this kernel is PD for $\lambda < \epsilon$[8]. In practice ϵ is not known which makes it hard to judge when this approach may be applied.
It is worth mentioning that all the Hilbertian metrics explicitly mentioned in this section can be written as f-divergences. It is a classical result in information geometry that all f-divergences induce up to scaling the Fisher metric. In this sense all considered metrics are locally equivalent. Globally we have the upper and lower bounds introduced earlier. Therefore we expect in our experiments relatively small deviations in the results of the different metrics.

3.2 Hilbertian Metrics on Probability Measures Incorporating Structural Properties of the Probability Space

If the probability space $\mathcal{X}$ is a metric space $(\mathcal{X}, d_{\mathcal{X}})$ one can use $d_{\mathcal{X}}$ to derive a metric on $\mathcal{M}_+^1(\mathcal{X})$. One example of this kind is the Kantovorich metric:

$$d_K(P,Q) = \inf_{\mu} \{ \int_{\mathcal{X}\times\mathcal{X}} d(x,y) d\mu(x,y) \Big| \mu \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{X}), \pi_1(\mu) = P, \pi_2(\mu) = Q \}$$

where π_i denotes the marginal with respect to i-th coordinate. When $\mathcal{X}$ is finite, the Kantovorich metric gives the solution to the mass transportation problem. In a similar spirit we extend the generation of Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ based on (1) by using similarity information of the probability space $\mathcal{X}$. That means we do not only compare the densities pointwise but also the densities of distinct points weighted by a similarity measure $k(x,y)$ on $\mathcal{X}$. The only requirement we need is that we are given a similarity measure on $\mathcal{X}$, namely a positive definite kernel $k(x,y)$[9]. The disadvantage of our approach is that we are not anymore invariant with respect to the dominating measure. 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 $\mathcal{X}$ and use this as a dominating measure. We denote in the following by $\mathcal{M}_+^1(\mathcal{X},\mu)$ all probability measure which are dominated by μ.

Theorem 4. *Let k be a PD kernel on $\mathcal{X}$ and $\hat{k}$ a PD kernel on $\mathbb{R}_+$ such that $\int_{\mathcal{X}} \sqrt{k(x,x)\,\hat{k}(q(x),q(x))}\, d\mu(x) < \infty$, $\forall\, q \in \mathcal{M}_+^1(\mathcal{X},\mu)$. Then*

$$K(P,Q) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y)\, \hat{k}(p(x),q(y))\, d\mu(x)\, d\mu(y) \tag{3}$$

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

[8] which does not imply that $-d_{geo}^2$ is CPD.
[9] Note that a positive definite kernel k on $\mathcal{X}$ always induces a semi-metric on $\mathcal{X}$ by $d_{\mathcal{X}}^2(x,y) = k(x,x) + k(y,y) - 2k(x,y)$.

Proof. Note first that the product $k(x,y)\hat{k}(r,s)$ $(x,y \in \mathcal{X}, r,s \in \mathbb{R}_+)$ is a positive definite kernel on $\mathcal{X} \times \mathbb{R}_+$. The corresponding RKHS $\mathcal{H}$ is the tensor product of the RKHS $\mathcal{H}_k$ and $\mathcal{H}_{\hat{k}}$, that is $\mathcal{H} = \mathcal{H}_k \otimes \mathcal{H}_{\hat{k}}$. We denote the corresponding feature map by $(x,r) \to \phi_x \otimes \psi_r$. Now let us define a linear map $L_q : \mathcal{H} \to \mathbb{R}$ by

$$\begin{aligned} L_q : \phi_x \otimes \psi_r \longrightarrow \int_{\mathcal{X}} k(x,y)\hat{k}(r,q(y))d\mu(y) &= \int_{\mathcal{X}} \langle \phi_x, \phi_y \rangle_{\mathcal{H}_k} \left\langle \psi_r, \psi_{q(y)} \right\rangle_{\mathcal{H}_{\hat{k}}} d\mu(y) \\ &\leq \|\phi_x \otimes \psi_r\|_{\mathcal{H}} \int_{\mathcal{X}} \left\|\phi_y \otimes \psi_{q(y)}\right\|_{\mathcal{H}} d\mu(y) \end{aligned}$$

Therefore by the assumption L_q is continuous. By the Riesz lemma, there exists a vector u_q such that $\forall v \in \mathcal{H}$, $\langle u_q, v \rangle_{\mathcal{H}} = L_q(v)$. It is obvious from

$$\begin{aligned} \langle u_p, u_q \rangle_{\mathcal{H}} &= \int_{\mathcal{X}} \left\langle u_p, \phi_y \otimes \psi_{q(y)} \right\rangle_{\mathcal{H}} d\mu(y) = \int_{\mathcal{X}^2} \left\langle \phi_x \otimes \psi_{p(x)}, \phi_y \otimes \psi_{q(y)} \right\rangle_{\mathcal{H}} d\mu(y) d\mu(x) \\ &= \int_{\mathcal{X}^2} k(x,y)\, \hat{k}(p(x),q(y))\, d\mu(x)\, d\mu(y) \end{aligned}$$

that K is positive definite.

The induced Hilbertian metric D of K is given by

$$\begin{aligned} D^2(P,Q) &= \int_{\mathcal{X}^2} k(x,y) \left[\hat{k}(p(x),p(y)) + \hat{k}(q(x),q(y)) - 2\hat{k}(p(x),q(y))\right] d\mu(x) d\mu(y) \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y) \left\langle \psi_{p(x)} - \psi_{q(x)}, \psi_{p(y)} - \psi_{q(y)} \right\rangle d\mu(x) d\mu(y). \end{aligned} \tag{4}$$

4 Experiments

The performance of the following Hilbertian metrics on probability distributions

$$\begin{aligned} d^2_{geo}(P,Q) &= \arccos^2\Big(\sum_{i=1}^{N} \sqrt{p_i}\sqrt{q_i}\Big), \qquad d^2_{\chi^2}(P,Q) = \sum_{i=1}^{N} \frac{(p_i - q_i)^2}{p_i + q_i} \\ d^2_H(P,Q) &= \frac{1}{4} \sum_{i=1}^{N} (\sqrt{p_i} - \sqrt{q_i})^2, \qquad d^2_{TV}(P,Q) = \frac{1}{2} \sum_{i=1}^{N} |p_i - q_i| \\ d^2_{JS}(P,Q) &= \frac{1}{2} \sum_{i=1}^{N} p_i \log\left(\frac{2p_i}{p_i + q_i}\right) + q_i \log\left(\frac{2q_i}{p_i + q_i}\right) \end{aligned} \tag{5}$$

respectively of the transformed "Gaussian" metrics

$$d^2_{exp}(P,Q) = 1 - \exp(-\lambda d^2(P,Q)) \tag{6}$$

was evaluated in three multi-class classification tasks:
The *Reuters* data set. The documents are represented as term histograms. Following [5] 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. The *WebKB* web pages data set. The documents are also represented as histograms. We used the four most frequent classes *student, faculty, course* and *project*. 4198 documents remained each of dimension 24212 (see [5]). The *Corel* image data base. We chose the data set Corel14 as in [1], which has 14 classes. Two different features were used. First the histogram was computed directly from the RGB data second from the CIE Lab color space, which has the advantage that the euclidean metric in that space locally discriminates colors according to the human vision uniformly over the whole space. Therefore the quantization process is more meaningful in CIE Lab than in RGB space[10]. In both spaces we used 16 bins per dimension, yielding a 4096-dimensional histogram. All the 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 using the CPD kernels $K = -d^2$. For each metric d from (5) we either used the metric directly with varying penalty constants C in the SVM, or we used the transformed metric d_{exp} defined in (6) again with different penalty constants C and λ. The best parameters were found using 10-folds cross-validation from the set $C \in \{10^k \mid k = -2, -1, ..., 4\} =: R_C$ respectively $(C, \lambda) \in R_C \times \frac{1}{\sigma}\{2, 1, \frac{2}{3}, \frac{1}{2}, \frac{2}{5}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{7}, \frac{1}{10}\}$, where σ was set to $\{\frac{\pi}{4}, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, \frac{\sqrt{\log 2}}{2}, \frac{\sqrt{2}}{2}\}$ to compensate for the different maximal distances of $d_{geo}, d_{\chi^2}, d_H, d_{JS}, d_{TV}$ respectively. For the best parameters the classifier was trained then on the whole training set and its error evaluated on the test set. The results are shown in Table 1. In a second experiment we used (4) for the Corel data[11]. We employ the euclidean CIE 94 distance on the color space since it models the color perception of humans together with the compactly supported RBF $k(x, y) = (1 - \|x - y\|)_+^2$, see e.g. [10], to generate a similarity kernel for the color space. Then the same experiments are done again for the RGB histograms and the CIE histograms with all the distances except the geodesic one, since it is not of the form (1). The results are shown in rows CIE CIE94 and RGB CIE94.

Table 1: The table shows the test errors with the optimal values of the parameters of C resp. C, λ found from 10-fold cross-validation. The first row of each data set is obtained using the metric directly, the second row shows the errors of the transformed metric (6).

	Geodesic			χ^2			Hellinger			JS			Total Var.		
	error	C	σ	error	C	σ	error	C	σ	error	C	σ	error	C	σ
Reuters	0.015	1		0.016	1		0.016	10^2		0.014	10		0.018	10^2	
	0.015	10	1/10	0.015	10	1/7	0.016	10	1/10	0.015	10	1/5	0.019	10^3	1/13
WebKB	0.052	1		0.046	1		0.046	1		0.045	10		0.052	1	
	0.045	10	1/2	0.048	10^3	2/5	0.044	10	1/2	0.049	10^4	2/3	0.050	10	1/10
Corel RGB	0.254	1		0.171	1		0.225	10		0.171	10		0.161	10	
	0.171	10^2	1/2	0.157	10^2	1	0.154	10^2	1	0.161	10^2	1/2	0.161	10^2	1/5
Corel CIE	0.282	1		0.179	10		0.200	10		0.196	10^3		0.186	10^2	
	0.154	10	1	0.146	10	2/5	0.139	10	2/3	0.146	10^2	2/3	0.171	10	2/3

[10] In principle we expect no difference in the results of RGB and CIE Lab when we use invariant metrics. The differences in practice come from the different discretizations.

[11] The geodesic distance cannot be used since it cannot be written in appropriate form.

RGB CIE94		0.161 1 0.157 10 1/4	0.214 10 0.164 100 1/2	0.168 10 0.161 100 2/3	0.168 10 0.157 100 1/5
CIE CIE94		0.161 10 0.154 10 2/5	0.182 10 0.143 10^2 2/5	0.150 10^2 0.146 10^2 2/3	0.193 10 0.179 10 2/5

The results show that there is not a "best" metric. It is quite interesting that the result of the direct application of the metric are comparable to that of the transformed "Gaussian" metric. Since the "Gaussian" metric requires an additional search for the optimal width parameter, in the case of limited computational resources the direct application of the metric seems to yield a good trade-off.

5 Conclusion

We presented a general method to build Hilbertian metrics on probability measures from Hilbertian metrics on $\mathbb{R}_+$. Using results of Fuglede we characterized the class of Hilbertian metrics on probability measures generated from Hilbertian metrics on $\mathbb{R}_+$ which are invariant under the change of the dominating measure. We then generalized this framework by incorporating a similarity measure on the probability space into the Hilbertian metric. Thus adding structural information of the probability space into the distance. Finally we compared all studied Hilbertian metrics in two text and one image classification tasks.

Acknowledgements We would like to thank Guy Lebanon for kindly providing us with the WebKB and Reuters dataset in preprocessed from. Furthermore we are thankful to Flemming Topsøe and Bent Fuglede for providing us with preprints of their papers [9, 2], to Olivier Chapelle for his help with the experiments in the early stages of this article and finally to Jeremy Hill, Frank Jäkel and Felix Wichmann for helpful suggestions on color spaces and color distances.

References

1. O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10:1055–1064, 1999.
2. B. Fuglede. Spirals in Hilbert space. With an application in information theory. To appear in Expositiones Mathematicae, 2004.
3. M. Hein and O. Bousquet. Maximal margin classification for metric spaces. In *16th Annual Conference on Learning Theory (COLT)*, 2003.
4. T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *16th Annual Conference on Learning Theory (COLT)*, 2003.
5. 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.
6. I. J. Schoenberg. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, 44:522–536, 1938.
7. B. Schölkopf. The kernel trick for distances. *NIPS*, 13, 2000.
8. F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Th.*, 46:1602–1609, 2000.
9. F. Topsøe. Jenson-shannon divergence and norm-based measures of discrimination and variation. Preprint, 2003.
10. H. Wendland. Piecewise polynomial, positive definite and compactly supported radial basis functions of minimal degree. *Adv. Comp. Math.*, 4:389–396, 1995.