Non-Euclidean principal component analysis by Hebbian learning

Lange, Mandy; Biehl, Michael; Villmann, Thomas

Published in:
Neurocomputing

DOI:
10.1016/j.neucom.2013.11.049

Neurocomputing 147 (2015) 107–119

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Non-Euclidean principal component analysis by Hebbian learning

Mandy Lange [a], Michael Biehl [b], Thomas Villmann [a,*]

[a] University of Applied Sciences Mittweida, Computational Intelligence Group, Technikumplatz 17, 09648 Mittweida, Germany
[b] University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, P.O. Box 407, 9700 AK Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 11 April 2013
Received in revised form
19 September 2013
Accepted 28 November 2013
Available online 25 June 2014

Keywords:
Principal component analysis
Hebbian learning
Kernel distances
Lp-norms
Semi-inner products

ABSTRACT

Principal component analysis based on Hebbian learning is originally designed for data processing in Euclidean spaces. We present in this contribution an extension of Oja's Hebbian learning approach for non-Euclidean spaces. We show that for Banach spaces the Hebbian learning can be carried out using the underlying semi-inner product. Prominent examples for such Banach spaces are the l_p-spaces for $p \neq 2$. For kernels spaces, as applied in support vector machines or kernelized vector quantization, this approach can be formulated as an online learning scheme based on the differentiable kernel. Hence, principal component analysis can be explicitly carried out in the respective data spaces but now equipped with a non-Euclidean metric. In the article we provide the theoretical framework and give illustrative examples.

1. Introduction

The use of unconventional distance measures or norms has gained popularity in many application domains [18,34,39,40,49,56,54]. However, a simple and reliable tool for data analysis like principal component analysis (PCA) is not available for such measures, in general. With this contribution we aim at closing this gap by providing variants of PCA which directly relate to suitable metrics such as l_p- or kernelized norms.

PCA constructs a basis of a multi-dimensional feature space, reflecting the variability observed in a given data set. It determines the linear projection of largest variance as well as orthogonal directions which are ranked according to decreasing variance [25]. Algebraic approaches to PCA, which determine directly the eigenvectors of the empirical covariance matrix, are sensitive to outliers, frequently. Iterative PCA based on Hebbian learning offers a more robust alternative as established in the pioneering work of Oja [37,38]. Several modifications and improvements of the basic idea have been proposed: while, for instance, Oja's subspace algorithm determines an arbitrary basis for the span of the leading eigenvectors [37,38], Sanger presented an extension which yields the eigenvectors ordered according to their eigenvalue, i.e. the observed empirical variance of projections [44].

A number of nonlinear extensions to the concept of PCA have been proposed in the literature. Kernel Hebbian learning was established by Kim et al. [30,31] based on the general concept of kernel PCA (KPCA) and reproducing kernel Hilbert spaces (RKHS) [20,48], which offer the possibility to capture non-linear data structures while applying PCA. This approach was further improved by Günther et al. who introduced an accelerating gain parameter [13]. Hebbian PCA for functional data using Sobolev metrics based on Euclidean norms was proposed in [56]. Other approaches for iterative PCA can be found in, for instance, [17].

The aim of this paper is to unify and generalize these approaches. In particular, we consider PCA in non-Euclidean spaces. We show that semi-inner products can be used for Hebbian PCA based on Oja's algorithm in Banach spaces. Semi-inner products are generalizations of inner products relaxing the strict properties of inner products but keeping the linear aspect. We further extend this generalization idea to kernel PCA as a non-linear kind of PCA. To this end, we revisit KPCA under the specific aspect of *differentiable kernels*.

As a result, PCA can be explicitly carried out in the data space but now equipped with non-Euclidean metrics. This allows for the adequate visualization of data in non-Euclidean spaces which becomes important when, for instance, classification is based on non-Euclidean projections or distances. These can facilitate better classification accuracy [27,14,47] or take into account application domain specific expertise and experience, e.g. the successful use of l_1-norms in image processing [50]. Yet, PCA visualization is closely related to the visualization obtained by multi-dimensional scaling (MDS, [11]) for Euclidean spaces. This remains true also for Banach spaces with a Schauder basis representation. However, PCA additionally provides the projection operator to be applied if new data become available whereas MDS has to be recalculated.

* Corresponding author.
E-mail address: thomas.villmann@hs-mittweida.de (T. Villmann).

http://dx.doi.org/10.1016/j.neucom.2013.11.049
0925-2312/

The paper is structured as follows: We start by revisiting Hebbian PCA in Euclidean spaces and extend this approach to general finite dimensional Hilbert-spaces (isomorphic to the Euclidean space). Thereafter we transfer the idea to learning in Banach spaces, like l_p-spaces, employing the concept of semi-inner products. In the last step we further extend this method to kernel spaces. Example applications and different data sets illustrate the new approaches and demonstrate their usefulness.

The paper is an extended version of the conference paper [5].

2. Hebbian learning of principal components in finite-dimensional vector spaces

In this section we discuss Hebbian learning for PCA in finite-dimensional Euclidean, Hilbert and Banach spaces, subsequently.

It is well known that any kind of normalization influences PCA in Euclidean spaces. This remains true also for general Hilbert or Banach spaces. We do not consider explicitly that point in this paper.

2.1. Hebbian PCA learning in the Euclidean space – Oja's and Sanger's rule

We consider centered n-dimensional data vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$. Hebbian principal component learning is based on a perceptron model. The model neuron has a weight vector $\mathbf{w} \in \mathbb{R}^n$ and generates the weighted output

$$O = \sum_{j=1}^{n} w_j \cdot v_j \tag{2.1}$$

for a given input $\mathbf{v}$. Mathematically, the output O is calculated as the Euclidean inner product

$$O = \langle \mathbf{v}, \mathbf{w} \rangle \tag{2.2}$$

between the weight vector $\mathbf{w}$ and the input $\mathbf{v}$ and frequently referred to as *Hebb-output* or *Hebb-response*.

Hebbian PCA learning introduced by Oja is a stochastic iteration

$$\Delta \mathbf{w} = \varepsilon \cdot O \cdot (\mathbf{v} - O \cdot \mathbf{w}) \tag{2.3}$$

using this Hebb-response O [37]. The parameter $0 < \varepsilon < 1$ is the so-called learning rate. The update (2.3) is known as *Oja's rule* in the literature [38]. Under the assumption of a slowly changing weight vector $\mathbf{w}$, i.e. $\varepsilon \ll 1$, the stationary state $\Delta \mathbf{w} = 0$ of Oja's rule corresponds to the eigenvalue equation

$$\mathbf{C}\mathbf{w} = \langle \mathbf{w}, \mathbf{C}\mathbf{w} \rangle \mathbf{w} \tag{2.4}$$

with the covariance matrix $\mathbf{C} = E[\mathbf{v}\mathbf{v}^\top]$ defined by the expectation operator $E[\cdot]$.

The stability analysis shows that the adaptation process (2.3) converges to the eigenvector corresponding to the maximum eigenvalue of $\mathbf{C}$ [37]. Therefore we denote this kind of PCA as *Hebbian PCA Learning*. Moreover, this learning scheme can be seen as a normalized stochastic gradient descent on the cost function $J(\mathbf{w}) = \mathbf{w}^\top \mathbf{C}\mathbf{w}$ [38].

The basic scheme can be extended to learn all principal components. To this end, Sanger considered n weight vectors $\mathbf{w}_i$ with Hebbian responses $O_i = \langle \mathbf{v}, \mathbf{w}_i \rangle$ and introduced the modified adaptation rule

$$\Delta \mathbf{w}_i = \varepsilon \cdot O_i \cdot \left(\mathbf{v} - \sum_{j=1}^{i} O_j \cdot \mathbf{w}_j \right). \tag{2.5}$$

Note that for $i = 1$ the update is equivalent to (2.3), Sanger's algorithm yields the eigenvectors of $\mathbf{C}$ in decreasing order with respect to the corresponding eigenvalues [44]. We denote this algorithm as *complete Hebbian PCA Learning*.

2.2. Hebbian PCA learning in general Hilbert spaces

We start considering (centered) data $\mathbf{v} = (v_1, \ldots, v_n)^\top$ in an n-dimensional Hilbert space $\mathbb{H}^n$ with the inner product $\langle \bullet, \bullet \rangle_{\mathbb{H}^n}$ defining the norm $\| \cdot \|_{\mathbb{H}^n}$. Because each n-dimensional Hilbert space $\mathbb{H}^n$ is isomorphic to the Euclidean space $\mathbb{R}^n$, there always exists an isomorphism $\Theta : \mathbb{R}^n \to \mathbb{H}^n$. Further, each linear operator constitutes a matrix $\mathbf{A}$. Application of such an operator to a vector then is defined by

$$\mathbf{A}[\mathbf{v}] = (\langle \mathbf{a}_1, \mathbf{v} \rangle_{\mathbb{H}^n}, \ldots, \langle \mathbf{a}_n, \mathbf{v} \rangle_{\mathbb{H}^n})^\top \tag{2.6}$$

where the $\mathbf{a}_i$ are the row vectors of $\mathbf{A}$.

Formally, we can now replace the Euclidean inner product in the Hebb-output (2.2) by the inner product $O_{\mathbb{H}^n} = \langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{H}^n}$ of the Hilbert space: we get

$$\Delta \mathbf{w} = \varepsilon \cdot (\mathcal{F}_{\mathbf{v}}[\mathbf{w}] - (O_{\mathbb{H}^n})^2 \cdot \mathbf{w}) \tag{2.7}$$

where

$$\mathcal{F}_{\mathbf{v}}[\mathbf{w}] = O_{\mathbb{H}^n} \cdot \mathbf{v} \tag{2.8}$$

and $\mathcal{F}_{\mathbf{v}}$ is a linear operator in the Hilbert space $\mathbb{H}^n$ because of the linearity of inner products.

In the next step we investigate the stationary state of (2.7): under the same assumption of slowly changing weight vectors as in the Euclidean case, we obtain the equation

$$\mathbf{C}_{\mathbb{H}^n}[\mathbf{w}] = \gamma \cdot \mathbf{w} \tag{2.9}$$

where $\mathbf{C}_{\mathbb{H}^n}[\mathbf{w}]$ is the expectation of $\mathcal{F}_{\mathbf{v}}[\mathbf{w}]$ taken over all $\mathbf{v}$ and $\gamma = E[(O_{\mathbb{H}^n})^2]$. In particular, $\mathbf{C}_{\mathbb{H}^n}$ plays the role of the covariance matrix (operator) in $\mathbb{H}^n$ according to the basis representation of vectors in $\mathbb{H}^n$, i.e.-

$$\begin{aligned} E[\mathcal{F}_{\mathbf{v}}[\mathbf{w}]] &= E[\mathbf{v} \cdot \langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{H}^n}] \\ &= E[\mathbf{v} \cdot \mathbf{v}^\top][\mathbf{w}] \\ &= \mathbf{C}_{\mathbb{H}^n}[\mathbf{w}] \end{aligned}$$

where the linearity of the inner product with respect to the first argument was used in the second step. The application of the operator $\mathbf{C}_{\mathbb{H}^n}[\mathbf{w}]$ has to be interpreted in the sense of (2.6) applying the considered inner product $\langle \bullet, \bullet \rangle_{\mathbb{H}^n}$.

The stability analysis of the eigenvalue equation (2.9) follows immediately from the isomorphism between $\mathbb{R}^n$ and $\mathbb{H}^n$. The extension to the Sanger-algorithm is straightforward.

These concepts can be easily transferred to infinite but *separable* Hilbert spaces $\mathcal{H}$: for those spaces, always a *countable* basis $H = \{h_k \in \mathcal{H} | k \in \mathbb{N}\}$ exists according to Zorn's-Lemma [26], with a respective unique representation $\mathbf{v} = \sum_{k=1}^{\infty} v_k \cdot h_k$ for all infinite-dimensional vectors $\mathbf{v} \in \mathcal{H}$. In this case the covariance operator $\mathbf{C}_{\mathcal{H}}$ becomes infinite-dimensional, too. Yet, it remains a linear operator, formally defined by the expectation $\mathbf{C}_{\mathcal{H}} = E[\mathbf{v} \cdot \mathbf{v}^\top]$ over infinite-dimensional vectors $\mathbf{v}$ represented according to the well-defined but infinite basis $\mathcal{B}$. The approximation property of the PCA is ensured by the Riesz representer theorem and Parseval's identity [42].

2.3. Hebbian PCA learning in Banach spaces

In the following, we study n-dimensional Banach spaces $\mathbb{B}^n$ with the norm $\| \cdot \|_{\mathbb{B}^n}$. Banach spaces have gained popularity in machine learning, recently [10,19,58,59]. Prominent n-dimensional examples are the real l_p-spaces with the Minkowski-p-norm

$$\| \mathbf{x} \|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \tag{2.10}$$

for $1 \le p \le \infty$. In particular, the frequently applied l_1-norm $\|\bullet\|_1$ constitutes a Banach space but does not form a Hilbert space. Thus, an inner product generating $\|\bullet\|_1$ does not exist.

2.3.1. Semi-inner products and Banach spaces

In the following, we briefly introduce basic concepts and properties of semi-inner products, which are important for Hebbian PCA learning in Banach spaces, neglecting details for better reading. The details are explained in the Appendix.

Semi-inner products, introduced by Lumer in 1961, can be seen as a generalization of inner products [32]:

Definition 1. A semi-inner product (SIP) $[\bullet,\bullet]$ of a vector space V is a map

$$[\bullet,\bullet] : V \times V \longrightarrow \mathbb{C} \tag{2.11}$$

with the following properties:

1. positive semi-definite
$$[\mathbf{x},\mathbf{x}] \ge 0 \tag{2.12}$$
and $[\mathbf{x},\mathbf{x}] = 0$ iff $\mathbf{x} = \mathbf{0}$
2. linear with respect to the first argument for $\xi \in \mathbb{C}$
$$\xi \cdot [\mathbf{x},\mathbf{z}] + [\mathbf{y},\mathbf{z}] = [\xi \cdot \mathbf{x} + \mathbf{y},\mathbf{z}] \tag{2.13}$$
3. Cauchy–Schwarz inequality
$$|[\mathbf{x},\mathbf{y}]|^2 \le [\mathbf{x},\mathbf{x}][\mathbf{y},\mathbf{y}] \tag{2.14}$$

We emphasize that, in contradiction to inner products, SIPs may violate the symmetry condition.

Lumer has proven that an arbitrary Banach space $\mathbb{B}$ with norm $\|\mathbf{x}\|_{\mathbb{B}}$ can be equipped with a SIP $[\bullet,\bullet]_{\mathbb{B}}$ such that

$$\|\mathbf{x}\|_{\mathbb{B}} = \sqrt{[\mathbf{x},\mathbf{x}]_{\mathbb{B}}} \tag{2.15}$$

is valid [32]. Real SIPs are continuous and generate a linear operator

$$\mathcal{F}_{\mathbf{x}}[\mathbf{y}] = [\mathbf{x},\mathbf{y}]_{\mathbb{B}} \cdot \mathbf{x} \tag{2.16}$$

according to Remark 2 in the Appendix. Further, real SIPs are unique, which follows from Corollary 5 in the Appendix.

The unique and continuous SIPs for the above-mentioned real l_p-spaces are given as

$$[\mathbf{x},\mathbf{y}]_p = \frac{1}{(\|\mathbf{y}\|_p)^{p-2}} \sum_{i=1}^{n} x_i |y_i|^{p-1} \operatorname{sgn}(y_i) \tag{2.17}$$

where $\operatorname{sgn}(x)$ is the signum function defined as

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{2.18}$$

For $p = 1$, the SIP

$$\begin{aligned}[\mathbf{x},\mathbf{y}]_1 &= \|\mathbf{y}\|_1 \sum_{i=1}^{n} x_i \cdot \operatorname{sgn}(y_i) \\ &= \|\mathbf{y}\|_1 \sum_{i=1, y_i \ne 0}^{n} x_i \cdot \frac{y_i}{|y_i|}\end{aligned} \tag{2.19}$$

is obtained [9], which generates the prominent l_1-norm. Here, $y_i \ne 0$ is assumed.

In case of the real function space $\mathcal{L}_p$ we have

$$[f,g]_p = \frac{1}{(\|g\|_p)^{p-2}} \int f \cdot |g|^{p-1} \cdot \operatorname{sgn}(g(t))\, dt \tag{2.20}$$

in analogy to (2.17). The real $\mathcal{L}_p$−space is closely related to the *Sobolev-space* $\mathcal{W}_{K,p} = \{f | D^\alpha f \in \mathcal{L}_p, |\alpha| \le K\}$ of real differentiable functions up to order K with $D^\alpha = \partial^{|\alpha|}/\partial\alpha_1 \ldots \partial\alpha_{|\alpha|}$ being the differential operator of order $|\alpha|$. Sobolev-spaces are of great interest in functional data analysis [18,40,43]. The norm of $\mathcal{W}_{K,p}$ is given by

$$\|f\|_{K,p} = \left[\sum_{|\alpha| \le K} (\|D^\alpha f\|_p)^p \right]^{1/p}, \tag{2.21}$$

and the unique SIP is

$$[f,g]_{K,p} = \frac{1}{\|g\|_{K,p}^{p-2}} \sum_{|\alpha| \le K} \int f^{(\alpha)} \cdot |g^{(\alpha)}|^{p-1} \operatorname{sgn}(g^{(\alpha)})\, dt \tag{2.22}$$

with $f^{(\alpha)} = D^\alpha f$, see Lemma 7 in the Appendix.

A generalization of SIPs can be considered, when the SIP properties are modified properly. In particular, the Cauchy–Schwarz inequality in Definition 1 can be replaced by the more general Hölder inequality as suggested in [35,36]

$$|[\mathbf{x},\mathbf{y}]| \le [\mathbf{x},\mathbf{x}]^{1/p} [\mathbf{y},\mathbf{y}]^{1/q} \tag{2.23}$$

with p and q are conjugated numbers, i.e. $1/p + 1/q = 1$. The respective SIP is denoted as *generalized* SIP of type p (gSIP(p)). It turns out that also the gSIP determines a norm via (2.15). This result was further extended by Zhang and Zhang [60].

Suppose functions $\phi, \psi : \mathbb{R}_+ \to \mathbb{R}_+$ and the map

$$[\bullet,\bullet]_\psi : V \times V \to \mathbb{C}$$

fulfills the positiveness and the linearity properties of as SIP according to Definition 1. Let further the generalized Hölder inequality

$$|[\mathbf{x},\mathbf{y}]_\psi| \le \psi([\mathbf{x},\mathbf{y}]_\psi) \cdot \phi([\mathbf{x},\mathbf{y}]_\psi)$$

be valid. Then $[\bullet,\bullet]_\psi$ is called a generalized SIP (gSIP). The gSIP reduces to gSIP(p) if we take $\psi(t) = t^p$ and $\phi(t) = t^q$ where p and q are again conjugated numbers. The gSIP generates a norm by

$$\|\mathbf{x}\|_\psi = \psi([\mathbf{x},\mathbf{x}]_\psi)$$

and, conversely, for any normed vector space exists a gSIP if the map ψ is surjective on $\mathbb{R}_+$ [60].

2.3.2. Hebbian PCA learning in separable Banach spaces

Each n-dimensional Banach space $\mathbb{B}^n$ is separable and countable with the finite basis $B = \{b_k \in \mathbb{B}^n\}$. Therefore, an unique finite basis representation $\mathbf{v} = \sum_{k=1}^{n} v_k b_k$ exists for each vector $\mathbf{v}$.

The application of a linear operator $\mathbf{A}$ in a n-dimensional Banach space $\mathbb{B}^n$ is defined via the SIP as

$$\mathbf{A}[\mathbf{v}] = ([\mathbf{a}_1,\mathbf{v}]_{\mathbb{B}^n}, \ldots, [\mathbf{a}_n,\mathbf{v}]_{\mathbb{B}^n})^\top \tag{2.24}$$

in analogy to Eq. (2.6).

For Hebbian PCA learning in Banach space, again, we suppose centered data vectors $\mathbf{v} \in \mathbb{B}^n$. Using the linear operator (2.16), we can rewrite the Euclidean Hebbian PCA learning (2.3) as

$$\Delta\mathbf{w} = \varepsilon \cdot (\mathcal{F}_{\mathbf{v}}[\mathbf{w}] - ([\mathbf{v},\mathbf{w}]_{\mathbb{B}^n})^2 \cdot \mathbf{w}) \tag{2.25}$$

replacing the Euclidean inner product by the Banach space SIP. As before, we assume slowly changing weight vectors. Then the corresponding stationary state equation reads as

$$E[\mathcal{F}_{\mathbf{v}}[\mathbf{w}]] = \gamma \cdot \mathbf{w}, \tag{2.26}$$

with the expectation $\gamma = E[([\mathbf{v},\mathbf{w}]_{\mathbb{B}^n})^2]$, which is again an eigenvalue equation. We have, in complete analogy to separable Hilbert spaces,

$$\begin{aligned}E[\mathcal{F}_{\mathbf{v}}[\mathbf{w}]] &= E[\mathbf{v} \cdot [\mathbf{v},\mathbf{w}]_{\mathbb{B}^n}] \\ &= E[\mathbf{v} \cdot \mathbf{v}^\top][\mathbf{w}] \\ &= \mathbf{C}_{\mathbb{B}^n}[\mathbf{w}]\end{aligned}$$

using the linearity of the SIP in the first argument in the second line. Thus $\mathbf{C}_{\mathbb{B}^n}$ can be interpreted as covariance matrix (operator) in the

Banach spaces $\mathbb{B}^n$ according to the given basis representation of vectors $\mathbf{v} \in \mathbb{B}^n$, i.e. $\mathbf{C}_{\mathbb{B}^n} = E[\mathbf{v} \cdot \mathbf{v}^\top]$. Yet, $\mathbf{C}_{\mathbb{B}^n}$ it is still a linear operator.

The stability analysis of conventional Euclidean Oja-learning does not rely on the sesqui-linearity[1] of the inner product but only takes the norm properties into account [37,38]. Hence, it is applicable also for semi-inner products and, therefore, the update yields the eigenvector corresponding to the largest eigenvalue also in the case of finite-dimensional Banach-spaces. Again, the extension to the Sanger-approach is straightforward.

Analogous to the Hilbert space case, we can formally extend these considerations to infinite Banach spaces $\mathbb{B}$ supposing a (countable) *Schauder basis* B_S for them, which holds for reflexive Banach spaces [24], see Appendix. The Schauder basis representation is unique and, therefore, it can serve for approximated representations [22,23,41].

Of course, generalized SIPs are also applicable in Hebbian PCA when the Hebb-output is generated by them. Yet, the respective Banach space has also to fulfill the additional constraints ensuring the separability and the existence of a (countable) Schauder basis.

3. Hebbian learning for PCA in reproducing kernel spaces

After revisiting properties of kernel spaces including both Hilbert and Banach spaces for reproducing kernel spaces, we explain in this section how the idea of iterative Hebbian PCA learning can be transferred to kernelized problems.

3.1. Kernel spaces

In the following we assume a compact metric space (V, d_V) with the vector space V equipped with a metric d_V. A function κ on V is a kernel

$$\kappa_\Phi : V \times V \to \mathbb{C}$$

if there exists a *separable* Hilbert space $\mathcal{H}$ and a map

$$\Phi : V \ni \mathbf{v} \longmapsto \Phi(\mathbf{v}) \in \mathcal{H} \tag{3.1}$$

with

$$\kappa_\Phi(\mathbf{v}, \mathbf{w}) = \langle \Phi(\mathbf{v}), \Phi(\mathbf{w}) \rangle_{\mathcal{H}} \tag{3.2}$$

for all $\mathbf{v}, \mathbf{w} \in V$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product of the Hilbert space $\mathcal{H}$. The mapping Φ is called feature map and $\mathcal{H}$ the feature space of V. Without further restrictions on the kernel κ_Φ, both, $\mathcal{H}$ and Φ are not unique. Positive kernels are of special interest because they *uniquely* correspond to a reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}$ in a canonical manner [2,33]. The kernel κ_Φ is said to be positive definite if for all finite subsets $V_m \subseteq V$ with cardinality $\#V_m = m$, the Gram-Matrix

$$\mathbf{G}_m = [\kappa(\mathbf{v}_i, \mathbf{v}_j) : i, j = 1 \ldots m] \tag{3.3}$$

is positive semi-definite [2]. The norm $\| \Phi(\mathbf{v}) \|_{\mathcal{H}} = \sqrt{\kappa_\Phi(\Phi(\mathbf{v}), \Phi(\mathbf{v}))}$ of this RKHS induces a metric

$$d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = \sqrt{\kappa_\Phi(\mathbf{v}, \mathbf{v}) - 2\kappa_\Phi(\mathbf{v}, \mathbf{w}) + \kappa_\Phi(\mathbf{w}, \mathbf{w})} \tag{3.4}$$

based on the kernel κ_Φ [46]. Steinwart has shown that continuous, universal kernels induce the continuity and separability of the corresponding feature map Φ and the image $\mathcal{I}_{\kappa_\Phi} = \Phi(V)$ is a subspace of $\mathcal{H}$ [52].

It was further shown in this Steinwart-paper that continuous, universal kernels also imply the continuity and injectivity of the map

$$\Psi : (V, d_V) \longrightarrow (V, d_{\kappa_\Phi}) \tag{3.5}$$

with $d_{\kappa_\Phi}(\mathbf{v}, \mathbf{w}) = d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w}))$ and (V, d_{κ_Φ}) is the compact vector space V with the kernel induced metric d_{κ_Φ}. It was shown in [55] that (V, d_{κ_Φ}) is isometric and isomorphic to $\mathcal{I}_{\kappa_\Phi}$.

An analogous theory can be obtained if the mapping space has weaker assumptions: Zhang et al. consider *reflexive* Banach spaces as mapping spaces [59]. As above for the Hilbert space $\mathcal{H}$, the Banach space is also assumed to be a function space, here. Consider such a reflexive function Banach space $\mathcal{B}$ over the compact metric space (V, d_V) with the SIP $[h, g]_{\mathcal{B}}$, which additionally has a reproducing property for Banach spaces (Reproducing Kernel Banach space, RKBS).

If the RKBS is Fréchet-differentiable, it is called a SIP-RKBS. Again, we consider the feature map $\Phi : V \longrightarrow \mathcal{B}$. For a SIP-RKBS $\mathcal{B}$ a unique correspondence exists between a so-called SIP-kernel γ_Φ and the map Φ with

$$\gamma_\Phi(\mathbf{v}, \mathbf{w}) = [\Phi(\mathbf{v}), \Phi(\mathbf{w})]_{\mathcal{B}} \tag{3.6}$$

based on a Banach space representation theorem [59]. If the map Φ is continuous then also γ_Φ is. Moreover, one can show that (weakly) universal SIP-kernels correspond to bijective mappings Φ [55]. Further, it turns out that the map

$$\Psi : (V, d_V) \longrightarrow (V, d_{\mathcal{B}}) \tag{3.7}$$

is also continuous and, therefore, bijective iff the SIP-kernel is (weakly) universal and continuous. In consequence, the subspace $\mathcal{I}_{\gamma_\Phi} = \Phi(V) \subseteq \mathcal{B}$ is isomorphic to $(V, d_{\mathcal{B}})$. These results are proven in [55].

3.2. Kernel principal component analysis

We start this subsection considering a RKHS $\mathcal{H}$ as a mapping space by a map Φ from a data vector space V and the corresponding kernel κ_Φ. We assume centralized kernels, i.e. $E[\Phi(\mathbf{v})] = \mathbf{0}$, which can always be achieved for arbitrary positive kernels and finite data sets [46]. We define $\mathbf{C}_\Phi = E[\Phi(\mathbf{v}) \cdot (\Phi(\mathbf{v}))^\top]$. In case of an infinite-dimensional $\mathcal{H}$, we have to interpret $\Phi(\mathbf{v}) \cdot (\Phi(\mathbf{v}))^\top$ as a linear operator $\Omega_{\mathcal{H}}$ on $\mathcal{H}$

$$\Omega_{\mathcal{H}}[\mathbf{h}] = \Phi(\mathbf{v}) \cdot \langle \Phi(\mathbf{v}), \mathbf{h} \rangle_{\mathcal{H}}. \tag{3.8}$$

Following Schölkopf et al. in [48] the respective eigen-problem $\mathbf{C}_\Phi \mathbf{g} = \lambda \mathbf{g}$ can be solved using the observation that for all $\mathbf{v} \in V$ the equation $\lambda \langle \Phi(\mathbf{v}), \mathbf{g} \rangle_{\mathcal{H}} = \langle \Phi(\mathbf{v}), \mathbf{C}_\Phi \mathbf{g} \rangle_{\mathcal{H}}$ has to be fulfilled. For a data set $D \subset V$ with m linear independent data vectors $\mathbf{v}_k$ there exists a dual representation of the eigenvectors $\mathbf{g} = \sum_{j=1}^m \alpha_j \Phi(\mathbf{v}_j)$ such that in this case $\mathbf{C}_\Phi$ becomes the Gram-matrix $\mathbf{G}_m$ from (3.3). Then the original eigen-problem can be replaced by the dual problem

$$m\lambda \boldsymbol{\alpha} = \mathbf{G}_m \boldsymbol{\alpha} \tag{3.9}$$

where $\boldsymbol{\alpha}$ is the column vector of the values α_i. According to Zhang et al., this eigen-decomposition can also be seen as an eigen-problem for a linear operator determined by

$$\langle T\mathbf{c}, \mathbf{h} \rangle_{\mathcal{H}} = \frac{1}{m} \sum_{j=1}^{m} \langle \Phi(\mathbf{v}_j), \mathbf{c} \rangle_{\mathcal{H}} \langle \Phi(\mathbf{v}_j), \mathbf{h} \rangle_{\mathcal{H}} \tag{3.10}$$

using the kernel properties [59].

It is possible to extend the RKHS approach to RKBS [59]: consider an RKBS $\mathcal{B}$ as a mapping space by a map Φ from a data vector space V and the corresponding (centralized) SIP-kernel γ_Φ. We consider again a data set $D \subset V$ with m data vectors $\mathbf{v}_k$. Let us define for an arbitrary $\mathbf{v} \in \mathcal{B}$ the complex m-dimensional vector

$$\tilde{\Phi}_{\mathcal{B}}(\mathbf{v}) = ([\Phi(\mathbf{v}), \Phi(\mathbf{v}_1)]_{\mathcal{B}}, \ldots, [\Phi(\mathbf{v}), \Phi(\mathbf{v}_m)]_{\mathcal{B}}) \tag{3.11}$$

such that a linear operator T on $\mathbb{C}^m$ can be defined by

$$T\mathbf{c} = \frac{1}{m} \sum_{j=1}^{m} (\tilde{\Phi}^*_{\mathcal{B}}(\mathbf{v}_j)\mathbf{c}) \tilde{\Phi}_{\mathcal{B}}(\mathbf{v}_j) \tag{3.12}$$

[1] Sesqui-linearity means linearity in one argument and antilinearity in the other.

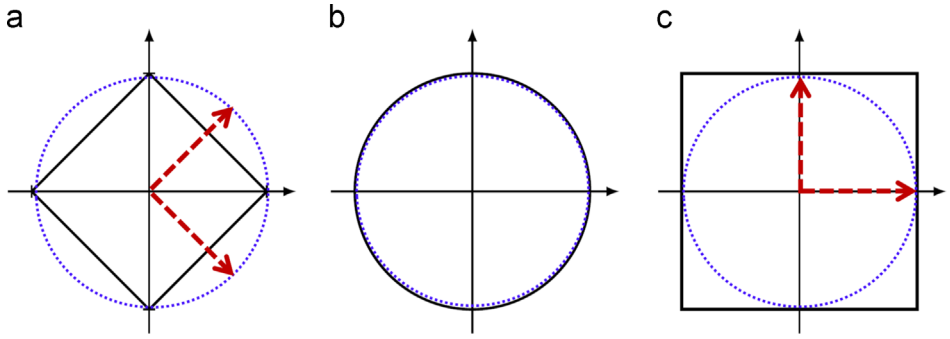

Fig. 1. Unit balls (black solid lines) and eigenvectors (red bold arrows) for circular data (blue dashed line) for several Minkowski-p-norms $\|\mathbf{x}\|_p$ from (2.10): (a) $p=1$ – the eigenvectors are in the diagonals of the rectangular axis system. (b) $p=2$ – no preferred direction. (c) $p=\infty$ – the eigenvectors coincide with the axes. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

where $\tilde{\Phi}^*_{\mathcal{B}}(\mathbf{v}_j)$ is the conjugate transpose of $\tilde{\Phi}_{\mathcal{B}}(\mathbf{v}_j)$, which corresponds to $T\mathbf{c}=\mathbf{M}_m\mathbf{c}$ with

$$\mathbf{M}_m=\frac{1}{m}(\mathbf{K}^*_m\cdot\mathbf{K}_m)^\top \tag{3.13}$$

and

$$\mathbf{K}_m=[\gamma_\Phi(\mathbf{v}_i,\mathbf{v}_j):i,j=1\ldots m] \tag{3.14}$$

is the Gram-matrix of the SIP-kernel γ_Φ. Hence, here the dual problem is

$$\mathbf{M}_m\boldsymbol{\alpha}=\lambda\boldsymbol{\alpha} \tag{3.15}$$

with the basis representation according to

$$\langle\tilde{\Phi}_{\mathcal{B}}(\mathbf{v}),\boldsymbol{\alpha}\rangle_{\mathbb{C}^m}=\sum_{j=1}^{m}\overline{\alpha_j}\gamma_\Phi(\mathbf{v},\mathbf{v}_j) \tag{3.16}$$

where $\overline{\alpha_j}$ is the conjugate-complex of α_j.

3.3. Kernel PCA and Hebbian learning

Kernel Hebbian learning based on the Oja-learning rule (2.3) was proposed in [31]. It is carried out implicitly in the Hilbert space $\mathcal{H}$ such that the coefficient vector $\boldsymbol{\alpha}$ in (3.9) is iteratively determined using the Gram-matrix $\mathbf{G}_m$ from (3.3). This approach can be transferred to the kernel Banach space problem in a straightforward manner by replacing, in the terms containing $\mathbf{G}_m$, the respective parts by $\mathbf{M}_m$ from (3.13). Due to the lack of space we drop the explicit formulation and follow a different route: we consider the mapping Ψ for RKHS and RKBS in the following.

3.3.1. Hebbian PCA learning in $(V,d_{\mathcal{H}})$

Suppose a data space V with the original data metric d_V frequently given as the Euclidean metric d_E. Now, we process PCA in the space $(V,d_{\mathcal{H}})$ from (3.5) using its isomorphism to the image space $\mathcal{I}_{\kappa_\Phi}\subseteq\mathcal{H}$ of the kernel mapping Φ such that the data remain the original ones but are equipped with the kernel metric, i.e. the relations among them are changed compared to the original data space (V,d_V).

Furthermore, we assume centralized kernels such that $E[\Psi(\mathbf{v})]=\mathbf{0}$. Now Oja's learning rule (2.3) in $(V,d_{\mathcal{H}})$ for given $\mathbf{v}\in(V,d_V)$ is given as

$$\Delta\mathbf{w}=\varepsilon\cdot O_{\mathcal{H}}\cdot(\Psi(\mathbf{v}_k)-O_{\mathcal{H}}\cdot\mathbf{w}) \tag{3.17}$$

where

$$O_{\mathcal{H}}=\kappa_\Phi(\Psi(\mathbf{v}_k),\mathbf{w}) \tag{3.18}$$

is the new non-Euclidean Hebbian response instead of the Euclidean inner product used in the original Oja's learning rule [37]. Substituting this in (3.17) we get

$$\Delta\mathbf{w}=\kappa_\Phi(\Psi(\mathbf{v}_k),\mathbf{w})\cdot\Psi(\mathbf{v}_k)-\kappa_\Phi(\Psi(\mathbf{v}_k),\mathbf{w})\kappa_\Phi(\Psi(\mathbf{v}_k),\mathbf{w})\cdot\mathbf{w}, \tag{3.19}$$

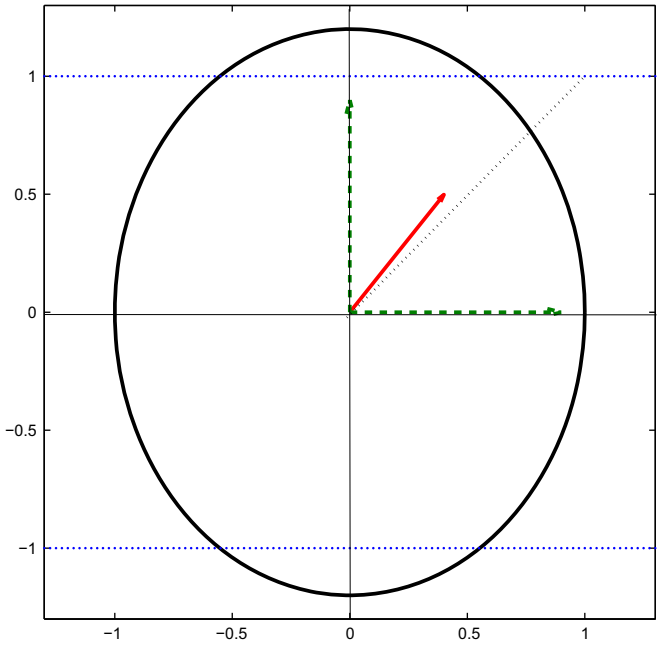

Fig. 2. Ellipsoid data set with radii $r_1=1$ and $r_2=1.2$. The Euclidean eigenvectors coincide with the coordinate axis because the symmetry of the unit ball is broken for an ellipse. The main principal vector according to the l_1-norm (red arrow) differs from the diagonal (dotted) and shifts in the direction given by the major radius r_2. It coincides with this, if $r_2>\sqrt{2}$ holds. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

which can be rewritten as

$$\Delta\mathbf{w}=\Omega[\mathbf{w}]-\kappa_\Phi(\Psi(\mathbf{v}_k),\mathbf{w})\kappa_\Phi(\Psi(\mathbf{v}_k),\mathbf{w})\cdot\mathbf{w} \tag{3.20}$$

using the linear operator $\Omega=\Psi(\mathbf{v}_k)\cdot(\Psi(\mathbf{v}_k))^\top$.[2] Here, the operator equation with

$$\Omega[\mathbf{w}]=\Psi(\mathbf{v}_k)\cdot\kappa_\Phi(\Psi(\mathbf{v}_k),\mathbf{w}) \tag{3.21}$$

is valid, which is comparable to Eq. (3.8). We remark at this point that $\Psi(\mathbf{v}_k)\in\mathcal{H}$ may be infinite dimensional vectors.

Under the usual assumption that the prototype $\mathbf{w}$ changes slowly compared to the number of presented inputs we get

$$\Delta\mathbf{w}=\mathbf{C}_\Psi[\mathbf{w}]-\lambda\mathbf{w} \tag{3.22}$$

with

$$\mathbf{C}_\Psi=E[\Omega] \tag{3.23}$$

[2] Note that $\Omega=\Psi(\mathbf{v}_k)\cdot(\Psi(\mathbf{v}_k))^\top$ is just a notation for the linear operator Ω in case of an infinite dimensional Hilbert spaces $\mathcal{H}$.

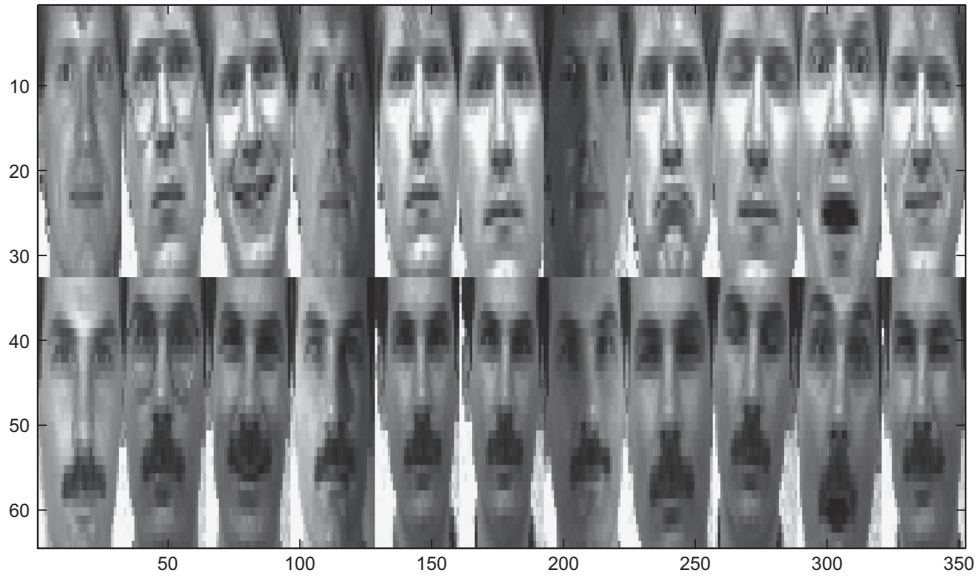

Fig. 3. Subset of the YALE face recognition data base used in the simulations.

defining the covariance in $(V, d_{\mathcal{H}})$, which reduces to

$$\mathbf{C}_{\Psi} = \frac{1}{m}\sum_{j=1}^{m} \Psi(\mathbf{v}_j)\cdot(\Psi(\mathbf{v}_j))^{\top} \tag{3.24}$$

for a finite number of data samples $D=\{\mathbf{v}_k | k=1\ldots m\}\subseteq V$.

The value λ in Eq. (3.22) is the expectation

$$\lambda = E[\kappa_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})\cdot\kappa_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})] \tag{3.25}$$

of the squared non-Euclidean Hebbian response O from (3.18). Thus, we obtain in the stationary state $\Delta\mathbf{w}=0$ an eigenvalue equation $\mathbf{C}_{\Psi}[\mathbf{w}]=\lambda\mathbf{w}$ for the operator $\mathbf{C}_{\Psi}$ for an eigenvector $\mathbf{w}\neq\mathbf{0}$ and eigenvalues $\lambda>0$. The last inequality stems from the positive definiteness of the kernel.

Because $\mathbf{w}\in(V,d_{\mathcal{H}})$, we may conclude that $\mathbf{w}\in\mathrm{span}\{\Psi(\mathbf{v}_j)|\ j=1\ldots m\}$ holds. Hence, the relation

$$\lambda\kappa_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w}) = \kappa_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{C}_{\Psi}[\mathbf{w}]) \tag{3.26}$$

must be valid for all $k=1\ldots n$. Moreover, $\mathbf{w}$ can be expressed as a linear combination

$$\mathbf{w} = \sum_{j=1}^{m}\alpha_j\Psi(\mathbf{v}_j)$$

of the images $\Psi(\mathbf{v}_k)$ of the original data vectors. Putting together the last statement with (3.26) we get

$$\lambda\sum_{j=1}^{m}\alpha_j\kappa_{\Phi}(\Psi(\mathbf{v}_k),\Psi(\mathbf{v}_j)) = \frac{1}{m}\sum_{j=1}^{m}\alpha_j\kappa_{\Phi}\left(\Psi(\mathbf{v}_k),\sum_{i=1}^{m}\Psi(\mathbf{v}_i)\cdot\kappa_{\Phi}(\Psi(\mathbf{v}_i),\Psi(\mathbf{v}_j))\right). \tag{3.27}$$

Here we have used the linearity of the kernel, interpreted as a real inner product, and the definition of $\mathbf{C}_{\Psi}$ in (3.23). If we now take into account the definition of the Gram-matrix $\mathbf{G}_n$ in (3.3), we immediately obtain

$$m\lambda\mathbf{G}_m\boldsymbol{\alpha} = \mathbf{G}_m^2\boldsymbol{\alpha} \tag{3.28}$$

where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)^{\top}$, which corresponds to the solution of the so-called dual eigen-problem (3.9) in [46], and, hence, the stability analysis can be taken from [31], which also provides the extension to the full eigen-problem and the respective Sanger-algorithm.

3.3.2. Hebbian PCA learning in $(V, d_{\mathcal{B}})$

Here we consider the space $(V,d_{\mathcal{B}})$ from (3.7) and exploit its isomorphism to the image space $\mathcal{I}_{\gamma_{\Phi}}\subseteq\mathcal{B}$ of the kernel mapping Φ for a SIP-RKBS $\mathcal{B}$. Because $\mathcal{B}$ is a RKBS it is reflexive and, therefore, possess a (countable) Schauder basis according to Remark 10 in the Appendix.

Again, we assume centralized kernels satisfying $E[\Psi(\mathbf{v})]=\mathbf{0}$. Further, we assume that the kernel γ_{Φ} takes only real values. Hence, $\mathbf{K}_m^{*}=\mathbf{K}_m^{\top}$ is valid in (3.13) which results in $\mathbf{M}_m=(1/m)(\mathbf{K}_m^{\top}\cdot\mathbf{K}_m)$ being symmetric and positive definite. The non-Euclidean Hebb-response becomes

$$O = \gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w}) \tag{3.29}$$

Substituting this in (3.17) we get in complete analogy

$$\Delta\mathbf{w} = \gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})\cdot\Psi(\mathbf{v}_k) - \gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})\gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})\cdot\mathbf{w}, \tag{3.30}$$

which reads as

$$\Delta\mathbf{w} = E[\Omega_{\mathcal{B}}[\mathbf{w}]] - \lambda\mathbf{w} \tag{3.31}$$

with the linear operator $\Omega_{\mathcal{B}}[\mathbf{w}]=\Psi(\mathbf{v}_k)\cdot\gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})$. and $\mathbf{C}_{\Psi}^{\mathcal{B}}=E[\Omega_{\mathcal{B}}]$.[3] The value λ in Eq. (3.31) is the expectation

$$\lambda = E[\gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})\cdot\gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})] \tag{3.32}$$

of the squared non-Euclidean Hebbian response O from Eq. (3.29).

Obeying the Schauder basis representation of vectors in $\mathcal{B}$ we obtain

$$\begin{aligned} E[\Omega_{\mathcal{B}}[\mathbf{w}]] &= E[\Psi(\mathbf{v}_k)\cdot\gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w})] \\ &= E[\Psi(\mathbf{v}_k)\cdot\Psi(\mathbf{v}_k)^{\top}][\mathbf{w}] \\ &= \mathbf{C}_{\Psi}^{\mathcal{B}}[\mathbf{w}] \end{aligned}$$

such that $\mathbf{C}_{\Psi}^{\mathcal{B}}=E[\Psi(\mathbf{v}_k)\cdot\Psi(\mathbf{v}_k)^{\top}]$ can be interpreted as a covariance operator. The stationary state $\Delta\mathbf{w}=0$ corresponds to the eigen-equation $\mathbf{C}_{\Psi}^{\mathcal{B}}[\mathbf{w}]=\lambda\mathbf{w}$ with eigenvector $\mathbf{w}\neq\mathbf{0}$ and eigenvalue $\lambda\neq0$.

Now we suppose data vectors $\mathbf{v}_j\in V$, $j=1\ldots m$. Because $\mathbf{w}\in(V,d_{\mathcal{B}})$, we may conclude that $\mathbf{w}\in\mathrm{span}\{\Psi(\mathbf{v}_j)|j=1\ldots m\}$ holds, because $\mathcal{B}$ is a SIP-RKBS. Hence, the relation

$$\lambda\gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{w}) = \gamma_{\Phi}(\Psi(\mathbf{v}_k),\mathbf{C}_{\Psi}^{\mathcal{B}}[\mathbf{w}]) \tag{3.33}$$

[3] We emphasize at this point that, $\Psi(\mathbf{v}_k)=\mathbf{v}_k$ is valid only numerically. Yet, $\mathbf{v}_k$ and its image $\Psi(\mathbf{v}_k)$ are objects in *different metric spaces*. Therefore, we will still use the notation $\Psi(\mathbf{v}_k)$ for the image to indicate this difference.

must be valid for all $k=1\ldots m$. Moreover, $\mathbf{w}$ can be expressed again as a linear combination $\mathbf{w}=\sum_{j=1}^{m}\beta_j\Psi(\mathbf{v}_k)$ of the images $\Psi(\mathbf{v}_k)$ of the original data vectors. Putting together the last statement together with (3.33) we get

$$\lambda\sum_{j=1}^{m}\beta_j\gamma_\Phi(\Psi(\mathbf{v}_k),\Psi(\mathbf{v}_j)) = \frac{1}{m}\sum_{j=1}^{m}\beta_j\gamma_\Phi\left(\Psi(\mathbf{v}_k),\sum_{i=1}^{m}\Psi(\mathbf{v}_i)\cdot\gamma_\Phi(\Psi(\mathbf{v}_i),\Psi(\mathbf{v}_j))\right) \tag{3.34}$$

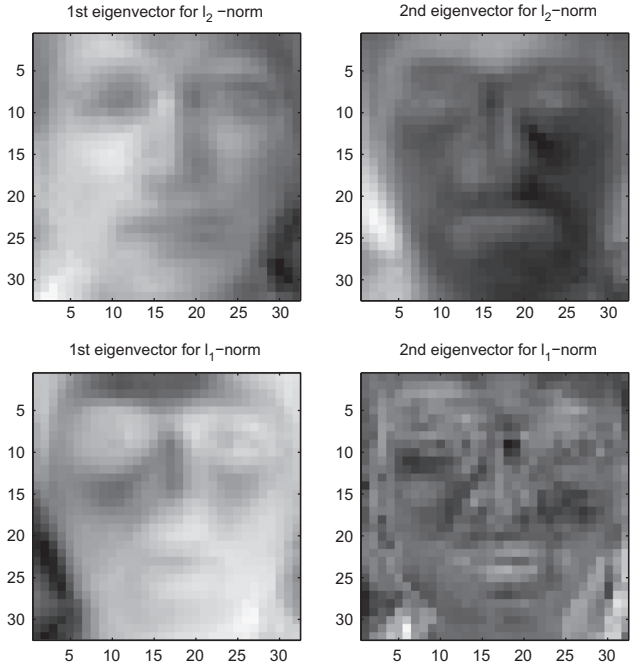

Fig. 4. First and second eigenfaces obtained for a subset of the YALE face recognition data base using the Euclidean inner product and the SIP $[\mathbf{x},\mathbf{y}]_1$ for Oja–Sanger learning (2.5).

using the linearity of the SIP-kernel in its first argument, interpreted as a real semi-inner product, and the definition of $\mathbf{C}_\Psi^{\mathcal{B}}$ as expectation. If we now take into account the definition of the Gram-matrix $\mathbf{K}_m$ in (3.14), we immediately conclude

$$m\lambda\mathbf{K}_m\boldsymbol{\beta}=\mathbf{K}_m^2\boldsymbol{\beta} \tag{3.35}$$

where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_m)^\top$ plays the same role as $\boldsymbol{\alpha}$ in (3.28). Moreover, it relates via the respective operator eigen-problems for RKHS (3.10) and RKBS (3.12) to the dual problem in case of RKBS (3.15).

As it was shown for the RKHS in [31], the stability analysis for RKBS follows analogously keeping also in mind that the original stability analysis in [37] does not require the sesqui-linearity of the inner product but only takes the resulting norm into account. Again, the extension to full PCA according to Sanger [44] is straightforward.

4. Simulations and results

In this section we present example applications and simulation results. We focus on demonstrating the different properties of the used inner products, SIPs and kernels for several data sets.

4.1. A two-dimensional toy example

This first example is an artificial one which, however, demonstrates well the aim of the non-Euclidean PCA. Here, we concentrate on l_p-norms (2.10).

We consider a circle C of radius $r=1$ in the two-dimensional plane, which is exactly the unit ball using the Euclidean distance corresponding to $p=2$ in the Minkowski-norm (2.10). However, the shape of the unit ball depends on the parameter p, see Fig. 1. Consequently, the principal directions of the circle C vary accordingly, which is also exemplified in Fig. 1.

Starting from these observations, we consider an ellipse with minor and major radius $r_1=1$ and $r_2=1.2$, respectively. Note that the corresponding principal components in Euclidean space ($p=2$) coincide with the axes. However, the principal axes for $p=1$ using

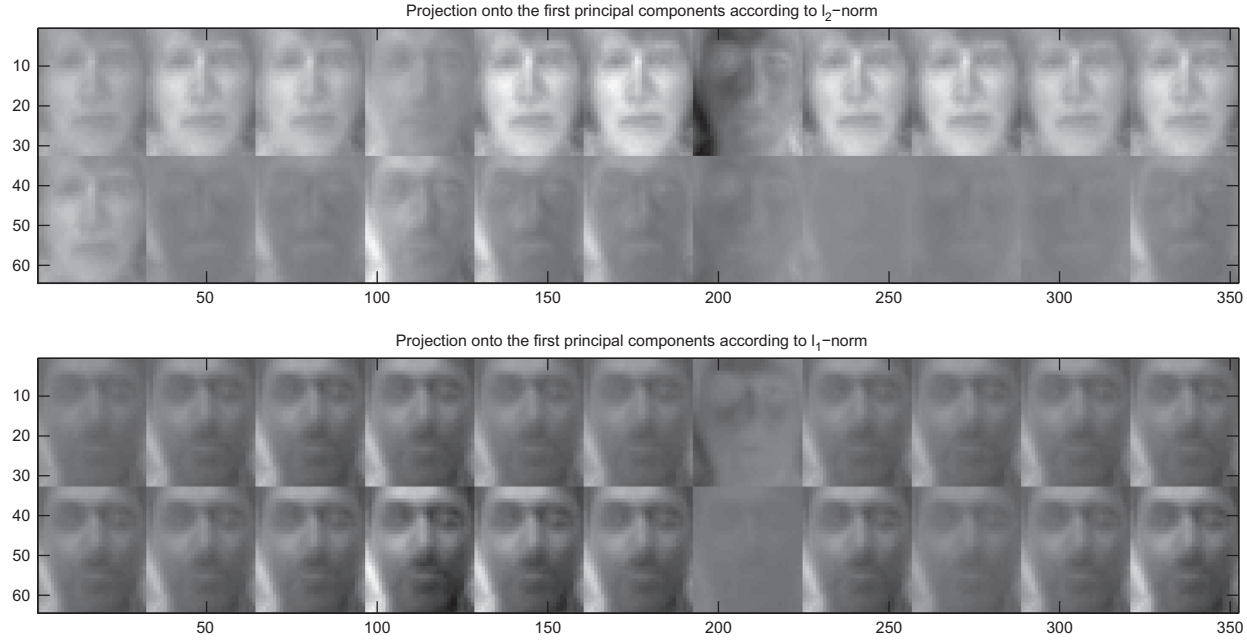

Fig. 5. Reconstruction of the original face images using only two principal components according to the Euclidean inner product (l_2-norm, top) and the SIP $[\mathbf{x},\mathbf{y}]_1$ (l_1-norm, bottom). The different behavior is obvious.

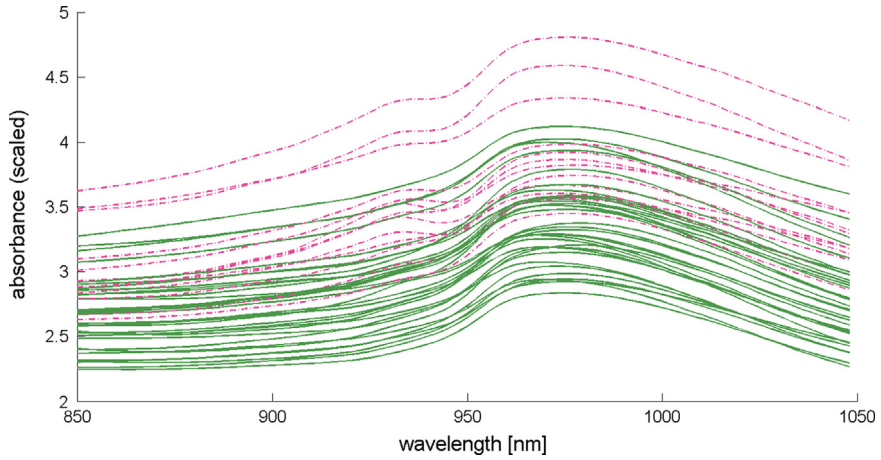

Fig. 6. Visualization of the TECATOR data set. The data vectors represent smooth spectra of meat probes with high and low fat content (two classes).

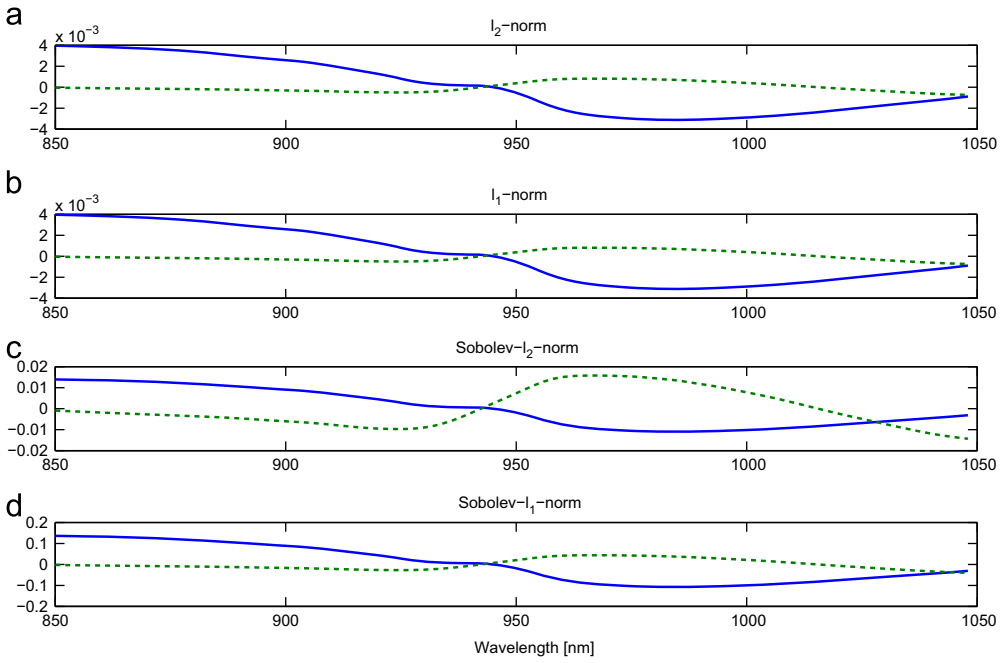

Fig. 7. Visualization of the first two eigenvectors for different norms obtained from Oja–Sanger learning using corresponding SIPs for the TECATOR data set. The vectors are normalized to unit length according to the respective norm.

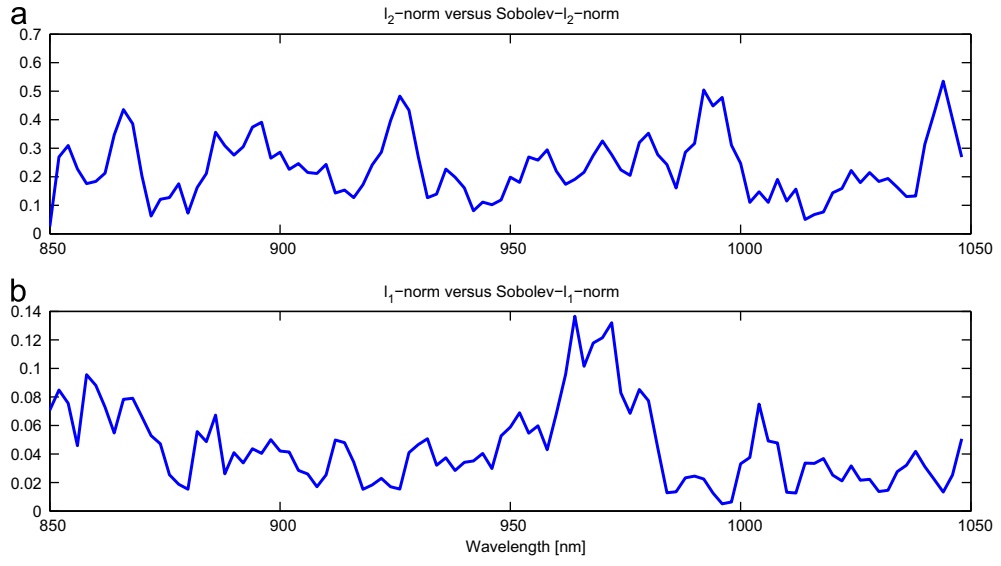

Fig. 8. Visualization of the quadratic differences for the eigenvectors of PCA between Sobolev-norms and non-functional norms for the TECATOR data set. Functional Sobolev-norms emphasize curved shapes of data.

the SIP $[\mathbf{x},\mathbf{y}]_1$ from (2.19) in Oja-learning are different from the axes provided $r_2 < \sqrt{2}$ holds true for the major radius. If $r_2 > \sqrt{2}$ is valid, the principal directions according to the l_1-norm are the same as for the Euclidean norm. Our simulations, taking the ellipse borders as inputs, show exactly this behavior, see Fig. 2.

4.2. Eigenfaces using l_p-norms

A more challenging application is the determination of eigenfaces in face recognition [1,29,50,57]. This is commonly done using standard Euclidean PCA. Yet, image processing frequently applies l_1-norms for image comparison [7,8,50,51]. Thus, non-Euclidean PCA corresponding to the l_1 norm should be useful in this domain. We illustrate the use of Oja–Sanger learning (2.5) for both Euclidean PCA and l_1-PCA applying the Euclidean inner product and the SIP, respectively.

For demonstration purposes we use a data set of 32×32 gray level images of two persons with 11 face positions/facial expressions for each [6], see Fig. 3.

This is a subset of the YALE face recognition data base [3].

Again we determined the eigenvectors according to the Euclidean inner product corresponding to the l_2-norm and the SIP $[\mathbf{x},\mathbf{y}]_1$ from (2.19) related to the l_1-norm. The resulting eigenfaces are depicted in Fig. 4.

Obviously, the eigenfaces differ significantly. This difference is also reflected when the eigenvectors are applied to approximate the original images, see Fig. 5.

Apparently, PCA according to the l_1-norm puts stronger emphasis on contours than the standard Euclidean PCA in this application.

4.3. Eigenvectors of functional data using Sobolev norms

Functional data $\mathbf{v} \in \mathbb{R}^n$ are vectorial data representing functions, i.e. $v_k = f(k)$. Frequently, these functions are assumed to be smooth and we consider differentiable functions, here. Respective dissimilarity measures including shape information are distances derived from the Sobolev-norm $\|f\|_{K,p}$ from (2.21). Hence, the related PCA is based on the corresponding SIP $[f,g]_{K,p}$ from (2.22).

In this example we consider the TECATOR-dataset [53]. The data set consists of 215 spectra obtained for several meat probes, see Fig. 6. The spectral range of wavelengths is between 850 nm and 1050 nm.

We applied Oja–Sanger learning (2.5) for the l_1- and the Euclidean norm as well as for the corresponding Sobolev norms $\|f\|_{1,1}$ and $\|f\|_{1,2}$ taking into account the first derivative. The resulting two eigenvectors according to the largest eigenvalues for each norm are plotted in Fig. 7.

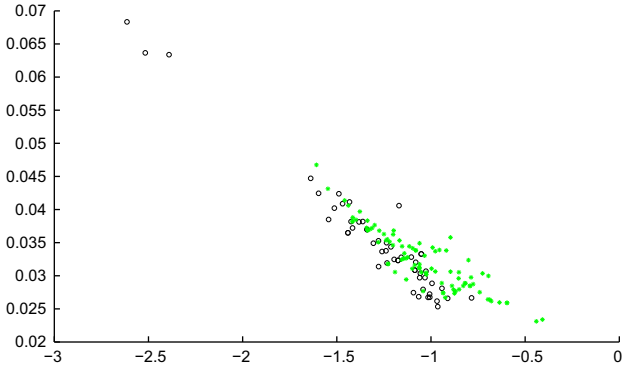

Fig. 9. Projection of the TECATOR data according to the l_1-norm PCA. The classes, i.e. low and high fat content, are displayed as green crosses and black circles, respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

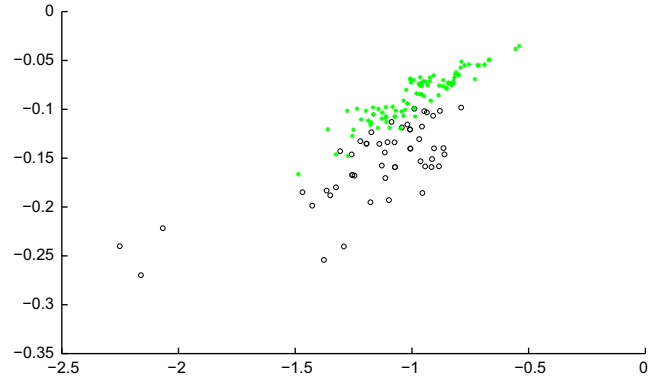

Fig. 10. Projection of the TECATOR data according to PCA based on the Sobolev norm $\|\cdot\|_{1,1}$. Green crosses and black circles correspond to low and high fat content, respectively. A clear separation of classes can be observed. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

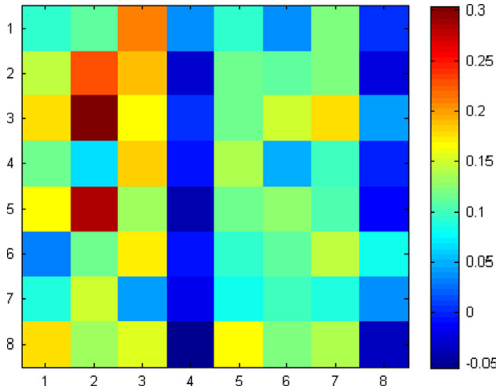

Fig. 11. Visualization of the used kernel matrix Ω in κ_Ω from (4.1) for the PIMA data set.

We can clearly observe the influence of the norms in use: Sobolev norms emphasize the spectral range around 950 nm for the l_1-variant and the ranges around 920 nm as well as 980 nm, paying attention to the derivatives, see Fig. 8.

These spectral ranges were also found to be important for classification (according to the meat's fat level) by relevance learning, in particular the range 950 nm [27]. Furthermore, it was shown that the classification of these data also benefits from the use of Sobolev norms [16]. This fact is further illustrated by the inspection of the data projections onto the respective principal components, see Figs. 9 and 10.

We find a pronounced separation of the classes when using the Sobolev norm $\|\cdot\|_{1,1}$–PCA, which confirms the findings in [16].

4.4. Eigenvectors in kernel PCA

The Indian diabetes data set (PIMA) is a standard data set from UCI which is frequently used for the comparison of classifiers [4]. It consists of 768 data vectors with 8 feature dimensions and is divided into two classes (healthy/ill). It turns out that learning the classification of this data set is relatively difficult. Application of the generalized learning vector quantization algorithm (GLVQ, [45]) using Euclidean distance achieves an accuracy of 75.1% [28]. If an adaptive exponential kernel distance is used in this method instead of the

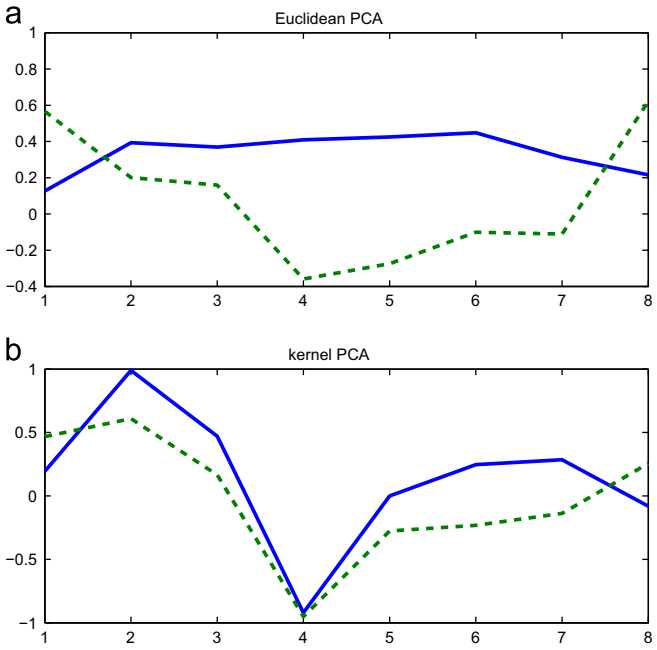

Fig. 12. Visualization of the eigenvectors of the PIMA data set according to the Euclidean inner product and the kernel κ_Ω from (4.1).

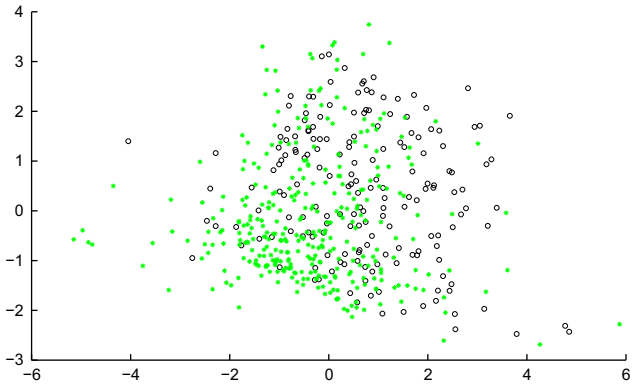

Fig. 13. Projection of the PIMA data according to the Euclidean PCA.

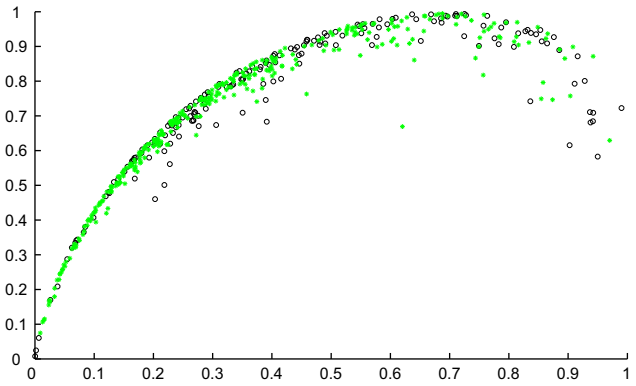

Fig. 14. Projection of the PIMA data according to Oja's kernel PCA with kernel κ_Ω from (4.1).

Euclidean, the accuracy is improved to 78.3%. There, the kernel was determined as

$$\kappa_\Omega(\mathbf{v},\mathbf{w}) = \exp(-(\Omega(\mathbf{v}-\mathbf{w}))^2) \tag{4.1}$$

with a square matrix Ω adapted during learning for optimal classification performance. The matrix Ω is displayed in Fig. 11.

As before, we also applied standard PCA to the data. Additionally, we performed Oja–Sanger learning (2.5) of the first two eigenvectors according to the kernel (4.1). The resulting eigenvectors are depicted in Fig. 12.

The related projections of the data are visualized in Figs. 13 and 14.

We may observe a slightly improved separability in case of the kernel PCA compared to the Euclidean PCA variant, which is in agreement with the improved class separability observed in [28] for kernel distance based classification learning using exactly the same matrix Ω.

5. Conclusion

In this paper we address the issue of PCA in non-Euclidean spaces. The use of non-standard distance measures shows an increased popularity to reflect the data characteristics. Yet, non-standard metrics require a consistent data processing, i.e. data analysis tools like PCA have to be performed using the same norm. With this contribution we propose variants of PCA which directly relate to suitable metrics such as l_p- or kernelized norms. In particular, we provide the framework for metrics based on norms, which are generated by semi-inner products, which are generalizations of inner products generating Banach spaces instead of Hilbert spaces for inner products. These semi-inner products (or generalizations thereof) can be directly plugged into the Hebbian PCA learning approach introduced for Euclidean inner products in usual PCA learning but now delivering principal components in non-Euclidean Banach spaces. We explain the theoretical framework for non-Euclidean PCA and prove mathematically that adaptive PCA by Hebbian learning can be done for general finite-dimensional Banach and Hilbert spaces in this context, which remains valid also for kernel metrics with underlying RKHS and RKBS. Hence, Hebbian PCA learning can also be extended to Kernel PCA learning.

Thus, generalizing the original Hebbian PCA learning in this manner we also close the gap between kernel based learning and adequate data visualization if kernel learning is done using *differentiable* kernels, which allow prototype based learning in the data space but equipped with a differentiable kernel metric as well as PCA for l_p-spaces and l_p-Sobolev spaces as prominent examples for Banach spaces.

Appendix A. Semi-inner products and Banach spaces

In this Appendix we collect important results about SIPs in Banach spaces, which are needed in the context of Hebbian learning according to Oja's learning rule.

Starting with Definition 1, we immediately observe that it is not necessarily symmetric. Furthermore Definition 1 implies semi-linearity in the second argument also called *homogeneity*, i.e.

$$[\mathbf{x}, \xi\cdot\mathbf{y}] = \overline{\xi}\cdot[\mathbf{x},\mathbf{y}] \tag{A.1}$$

for $\xi\in\mathbb{C}$ with $\overline{\xi}$ is the conjugate complex of ξ [12].[4]

Therefore, the SIPs are generally not symmetric: $[\mathbf{y},\mathbf{x}]\neq[\mathbf{x},\mathbf{y}]$, which distinguish them from inner products. Thus, the SIPs for Banach spaces are generalizations of inner products for Hilbert spaces. In contradiction to these, SIPs are not unique, in general.

Trivially, the linearity implies the continuity in the first argument. Yet, the SIP is called *continuous* if for real values $\lambda\in\mathbb{R}$ the

[4] The homogeneity together with the linearity in the first argument is also called sesqui-linearity.

real part $\Re$ of the SIP fulfills

$$\lim_{\lambda\to 0} \Re([\mathbf{y}, \mathbf{x}+\lambda\cdot\mathbf{y}]_{\mathbb{B}}) = \Re([\mathbf{y},\mathbf{x}]_{\mathbb{B}}). \tag{A.2}$$

The SIP is uniformly continuous, if the limit (A.2) is approached uniformly on $V\times V$. Obviously, the following remark is valid:

Remark 2. We consider a real SIP with $[\bullet,\bullet]_{\mathbb{B}} : V\times V \longrightarrow \mathbb{R}$ instead of (2.11). Then, the continuity is immediately given by the Cauchy–Schwarz inequality (2.14). In particular, we have linearity also in the second argument. Hence,

$$\mathcal{F}_{\mathbf{x}}[\mathbf{y}] = [\mathbf{x},\mathbf{y}]_{\mathbb{B}}\cdot\mathbf{x} \tag{A.3}$$

defines a linear operator in that case.

The norm $\|\bullet\|_{\mathbb{B}}$ from (2.15) is called *Gâteaux-differentiable* if the limit

$$D_{\mathbb{B}}(\mathbf{x}) = \lim_{\lambda\to 0}\frac{\|\mathbf{x}+\lambda\mathbf{y}\|_{\mathbb{B}} - \|\mathbf{x}\|_{\mathbb{B}}}{\lambda}$$

exists. If the limit converges uniformly, $\|\bullet\|_{\mathbb{B}}$ is denoted as *uniformly Fréchet-differentiable*. Giles has shown that in case of existence the relation

$$D_{\mathbb{B}}(\mathbf{x}) = \frac{\Re([\mathbf{y},\mathbf{x}]_{\mathbb{B}})}{\|\mathbf{x}\|_{\mathbb{B}}} \tag{A.4}$$

is valid [12]. Therefore, the following remark can be explicitly stated [59]:

Remark 3. If the norm $\|\mathbf{x}\|_{\mathbb{B}} = \sqrt{[\mathbf{x},\mathbf{x}]_{\mathbb{B}}}$ from (2.15) is Gâteaux-differentiable then the respective SIP is unique.

The continuity of the SIP can be related to the differentiability of the respective norm [12]:

Lemma 4. *The SIP $[\bullet,\bullet]_{\mathbb{B}}$ is continuous (uniformly continuous) iff the respective norm $\|\mathbf{x}\|_{\mathbb{B}} = \sqrt{[\mathbf{x},\mathbf{x}]_{\mathbb{B}}}$ is Gâteaux-differentiable (uniformly Fréchet-differentiable).*

Hence, the following conclusion can be drawn:

Corollary 5. *If the SIP $[\bullet,\bullet]_{\mathbb{B}}$ is continuous or uniformly continuous then it is also unique.*

The norms of the (complex) $\widehat{l}_p$–spaces with $1\le p\le\infty$ are Gâteaux-differentiable and their unique SIPs are given as

$$[\mathbf{x},\mathbf{y}]_p = \frac{1}{(\|\mathbf{y}\|_p)^{p-2}}\sum_{i=1}^{n} x_i\cdot\overline{y}_i\cdot|y_i|^{p-2}. \tag{A.5}$$

For real vectors $\mathbf{x}$ and $\mathbf{y}$, the SIP (A.5) becomes

$$[\mathbf{x},\mathbf{y}]_p = \frac{1}{(\|\mathbf{y}\|_p)^{p-2}}\sum_{i=1}^{n} x_i|y_i|^{p-1}\operatorname{sgn}(y_i)$$

where $\operatorname{sgn}(x)$ is the signum from (2.18) and the respective real space is denoted as l_p. Accordingly, the closely related Banach spaces $\widehat{\mathcal{L}}_p$ of complex functions are equipped with the respective SIP

$$[f,g]_p = \frac{1}{(\|g\|_p)^{p-2}}\int f\cdot\overline{g}\cdot|g|^{p-2}\,dt \tag{A.6}$$

for complex functions g and f. In case of the real function space $\mathcal{L}_p$ we have

$$[f,g]_p = \frac{1}{(\|g\|_p)^{p-2}}\int f\cdot|g|^{p-1}\cdot\operatorname{sgn}(g(t))\,dt.$$

The SIPs $[\mathbf{x},\mathbf{y}]_p$ and $[f,g]_p$ are uniformly continuous due to the Fréchet-differentiability of the p-norm $\|f\|_p = \sqrt{[f,f]_p}$ [59,15], which immediately implies the uniqueness according to Remark 3.

A representer theorem like for Hilbert spaces can be formulated for uniformly convex Banach spaces[5] [12]:

Theorem 6. *Let $\mathbb{B}$ be an uniformly convex and uniformly Fréchet-differentiable Banach space. Let f be a linear function, i.e. $f\in\mathbb{B}^*$. Then there exists a unique $\mathbf{y}\in\mathbb{B}$ such that $f(\mathbf{x}) = [\mathbf{x},\mathbf{y}]_{\mathbb{B}}$.*

It is well-known that l_p- and $\mathcal{L}_p$–spaces are uniformly convex for $1<p<\infty$.

The *Sobolev-space* $\mathcal{W}_{K,p} = \{f\,|\,D^{\alpha}f\in\mathcal{L}_p, |\alpha|\le K\}$ of (real) differentiable functions up to order K with $D^{\alpha} = \partial^{|\alpha|}/\partial\alpha_1\ldots\partial\alpha_{|\alpha|}$ being the differential operator has the norm

$$\|f\|_{K,p} = \left[\sum_{|\alpha|\le K}(\|D^{\alpha}f\|_p)^p\right]^{1/p}$$

i.e. the Sobolev-norm is based on the $\mathcal{L}_p$–norm. It is well-known that $\mathcal{W}_{K,p}$ and $\mathcal{L}_p$ are Hilbert spaces only for $p=2$. We can state the following lemma:

Lemma 7. *The unique SIP of $\mathcal{W}_{K,p}$ is given as*

$$[f,g]_{K,p} = \frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\int f^{(\alpha)}\cdot|g^{(\alpha)}|^{p-1}\operatorname{sgn}(g^{(\alpha)})\,dt$$

with $f^{(\alpha)} = D^{\alpha}f$.

Proof. (a) *SIP properties*: The properties (1)–(3) of a SIP according to Definition 1 are obviously fulfilled. The remaining property to show is the Cauchy–Schwarz inequality. We suppose $1<p<\infty$ and consider

$$\begin{aligned}|[f,g]_{K,p}| &= \left|\frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\int f^{(\alpha)}\cdot|g^{(\alpha)}|^{p-1}\operatorname{sgn}(g^{(\alpha)})\,dt\right| \\ &\le \frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\int|f^{(\alpha)}|\cdot|g^{(\alpha)}|^{p-1}\,dt,\end{aligned} \tag{A.7}$$

where the triangle inequality was applied. Using the Hölder inequality for integrals we obtain

$$\begin{aligned}&\frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\int|f^{(\alpha)}|\cdot|g^{(\alpha)}|^{p-1}\,dt \\ &\le \frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\left(\int|f^{(\alpha)}|^p\,dt\right)^{1/p}\cdot\left(\int|g^{(\alpha)}|^{(p-1)q}\,dt\right)^{1/q}\end{aligned} \tag{A.8}$$

for $1/p+1/q=1$. Hence, we have

$$q = \frac{p}{p-1}, \tag{A.9}$$

such that the right-hand term in (A.8) can be rewritten as

$$\begin{aligned}&\frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\left(\int|f^{(\alpha)}|^p\,dt\right)^{1/p}\cdot\left(\int|g^{(\alpha)}|^{(p-1)q}\,dt\right)^{1/q} \\ &= \frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\|f^{(\alpha)}\|_p\cdot\|g^{(\alpha)}\|_p^{p-1},\end{aligned} \tag{A.10}$$

which can be further majorized by

$$\begin{aligned}&\frac{1}{\|g\|_{K,p}^{p-2}}\sum_{|\alpha|\le K}\|f^{(\alpha)}\|_p\cdot\|g^{(\alpha)}\|_p^{p-1} \\ &\le \frac{1}{\|g\|_{K,p}^{p-2}}\left(\sum_{|\alpha|\le K}\|f^{(\alpha)}\|_p^p\right)^{1/p}\cdot\left(\sum_{|\alpha|\le K}\|g^{(\alpha)}\|_p^{(p-1)q}\right)^{1/q}\end{aligned} \tag{A.11}$$

[5] A Banach space with norm $\|\bullet\|_{\mathbb{B}}$ is uniformly convex if for each $\varepsilon>0$ exists a $\delta>0$ such that $\|\mathbf{x}+\mathbf{y}\|_{\mathbb{B}}\le 2-\delta$ holds if $\|\mathbf{x}-\mathbf{y}\|_{\mathbb{B}}\ge\varepsilon$ is valid.

again applying the Hölder inequality but now for sums. We detect the equivalence

$$\frac{1}{\|g\|_{K,p}^{p-2}}\left(\sum_{|\alpha|\le K}\|g^{(\alpha)}\|_p^{(p-1)q}\right)^{1/q}=\|g\|_{K,p} \tag{A.12}$$

paying again attention to the relation (A.9). Thus we conclude

$$\frac{1}{\|g\|_{K,p}^{p-2}}\left(\sum_{|\alpha|\le K}\|f^{(\alpha)}\|_p^{p}\right)^{1/p}\cdot\left(\sum_{|\alpha|\le K}\|g^{(\alpha)}\|_p^{(p-1)q}\right)^{1/q} =\|f\|_{K,p}\cdot\|g\|_{K,p}, \tag{A.13}$$

which gives the desired relation for the Cauchy–Schwarz inequality.

(b) *Uniqueness*: The Sobolev space $\mathcal{W}_p^K$ can be seen as the Cartesian product

$$\mathcal{W}_p^K=\mathcal{L}_p^{(0)}\otimes\mathcal{L}_p^{(1)}\otimes\cdots\otimes\mathcal{L}_p^{(K)}$$

of $\mathcal{L}_p$−spaces $\mathcal{L}_p^{(k)}$ where k denotes the order of the derivative. Thus, we have the sum of uniformly convex spaces, which is uniformly convex itself. Then, Remark 3 ensures the uniqueness. □

We emphasize the following remark about orthogonality with respect to SIPs:

Remark 8. Consider two vectors $\mathbf{v}$ and $\mathbf{w}$ in a Banach space $\mathbb{B}$. The vector $\mathbf{v}$ is *normal* to the vector $\mathbf{w}$ and the vector $\mathbf{w}$ is *transversal* to the vector $\mathbf{v}$ iff $[\mathbf{v},\mathbf{w}]_{\mathbb{B}}=0$, i.e. the orthogonality relation is not symmetric.

Last but not least we collect some properties regarding the separability of Banach spaces. Unfortunately, for *infinite-dimensional* Banach spaces $\mathbb{B}$, the separability property is not sufficient for a *countable* basis. However, the following statement can be made:

Remark 9. If a countable set of elements $B_S=\{b_k\in\mathbb{B}|k\in\mathbb{N}\}$ exists and B_S is dense in $\mathbb{B}$ then it is called a *Schauder-basis*, implying the separability of $\mathbb{B}$ and a respective unique vector representation $\mathbf{v}=\sum_{k=1}^{\infty}v_kb_k$ for all infinite-dimensional vectors $\mathbf{v}\in\mathbb{B}$ [26].

If the representation $\mathbf{v}=\sum_{k=1}^{\infty}v_kb_k$ converges unconditionally then the basis is called *unconditional*.

The Banach spaces l_p with the SIP (A.5) have a Schauder basis for $1\le p<\infty$ as well as the space $\mathcal{L}_p(\mathcal{K})$ over a compact set $\mathcal{K}\subset\mathbb{R}^n$ with the SIP (A.6). The same is valid for the real counterparts with SIPs (2.17) and (2.20), respectively. The latter one also implies a Schauder basis for the *Sobolev-space* $\mathcal{W}_{K,p}(\mathcal{K})$ with the SIP (2.22).

Let $\mathbb{B}^*$ be the dual space of linear functionals over $\mathbb{B}$ with Schauder basis $B_S=\{b_k\in\mathbb{B}|k\in\mathbb{N}\}$ and an arbitrary subspace $\mathbb{B}_n\subset\mathbb{B}$ spanned by $b_1,\ldots,b_n$ with dual $\mathbb{B}_n^*$. Consider a function $f\in\mathbb{B}^*$ and $f|_{\mathbb{B}_n}\in\mathbb{B}_n^*$ its restriction. The basis B_S is called *shrinking* if $\lim_{n\to\infty}\|f|_{\mathbb{B}_n}\|=0$ is valid.

Remark 10. According to a theorem provided by James, a Banach space is reflexive iff it has an unconditional shrinking Schauder basis [24]. Hence, we can always assume a Schauder basis for reflexive Banach spaces.

These mathematical considerations for SIPs remain also valid for generalized SIPs as introduced in Section 2.3.1, in particular the statements about uniqueness, existence and approximation capability based on the Schauder basis theory for Banach spaces. For a detailed mathematical analysis we refer to [60,21].

References

[1] J.L. Alba, A. Pujol, J.J. Villanueva, Separating geometry from texture to improve face analysis, in: IEEE International Conference on Image Processing, vol. 2, 2001, pp. 673–676.

[2] N. Aronszajn, Theory of reproducing kernels, Trans. Am. Math. Soc. 68 (1950) 337–404.

[3] A.S. Georghiades, YALE face data set. Available at: ⟨http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html⟩.

[4] A. Asuncion, D. Newman, Indian diabetes data set (PIMA) ⟨http://archive.ics.uci.edu/ml/⟩.

[5] M. Biehl, M. Kästner, M. Lange, T. Villmann, Non-Euclidean principal component analysis and Oja's learning rule – theoretical aspects, in: P. Estevez, J. Principe, P. Zegers (Eds.), Advances in Self-Organizing Maps: 9th International Workshop WSOM 2012 Santiage de Chile, Advances in Intelligent Systems and Computing, vol. 198, Springer, Berlin, 2013, pp. 23–34.

[6] D. Cai, Subset of YALE face data set. ⟨http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html⟩. dencai@gmail.com.

[7] N.W. Campbell, B.T. Thomas, T. Troscianko, Automatic segmentation and classification of outdoor images using neural networks, Int. J. Neural Syst. 8 (1) (1997) 137–144.

[8] J.C.W. Chan, K.P. Chan, A.G.O. Yeh, Detecting the nature of change in an urban environment: a comparison of machine learning algorithms, Photogramm. Eng. Remote Sens. 67 (February (2)) (2001) 213–225.

[9] J. Chmieliński, On an ϵ-Birkhoff orthogonality, J. Inequal. Pure Appl. Math. 6 (3) (2005) 1–7, Article 79.

[10] R. Der, D. Lee, Large-margin classification in Banach spaces, in: JMLR Workshop and Conference Proceedings, AISTATS, vol. 2, 2007, pp. 91–98.

[11] R. Duda, P. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[12] J. Giles, Classes of semi-inner-product spaces, Trans. Am. Math. Soc. 129 (1967) 436–446.

[13] S. Günter, N. Schraudolph, S. Vishwanathan, Fast iterative kernel principal component analysis, J. Mach. Learn. Res. 8 (2007) 1893–1918.

[14] B. Hammer, T. Villmann, Generalized relevance learning vector quantization, Neural Netw. 15 (8–9) (2002) 1059–1068.

[15] O. Hanner, On the uniform convexity of L^p and l^p, Ark. Mat. 3 (19) (1956) 239–244.

[16] C. Harth, Erweiterung von Generalized [Relevance | Matrix] LearningVector Quantization zur Anwendung auf funktionale Daten (Master's thesis), University of Applied Sciences Mittweida, Mittweida, Saxony, Germany, 2012, see ⟨http://www.mni.hs-mittweida.de/webs/villmann.html⟩.

[17] S. Haykin, Neural Networks. A Comprehensive Foundation, Macmillan, New York, 1994.

[18] G. Hébrail, B. Hugueney, Y. Lechevallier, F. Rossi, Exploratory analysis of functional data via clustering and optimal segmentation, Neurocomputing 73 (2010) 1125–1141.

[19] M. Hein, O. Bousquet, B. Schölkopf, Maximal margin classification for metric spaces, J. Comput. Syst. Sci. 71 (2005) 333–359.

[20] T. Hoffmann, B. Schölkopf, A. Smola, Kernel methods in machine learning, Ann. Stat. 36 (3) (2008) 1171–1220.

[21] A. Horváth, Semi-infinite inner product and generalized Minkowski spaces, J. Geom. Phys. 60 (2010) 1190–1208.

[22] P. Jain, K. Ahmad, Unconditional Schauder basis and best approximations in Banach spaces, Indian J. Pure Appl. Math. 12 (12) (1981) 1456–1467.

[23] P. Jain, K. Ahmad, Schauder decomposition and best approximations in Banach spaces, Port. Math. 44 (1) (1987) 25–39.

[24] R. James, Bases in Banach spaces, Am. Math. Mon. 89 (1982) 625–640.

[25] I. Joliffe, Principal Component Analysis, 2nd ed., Springer, Berlin-Heidelberg, 2002.

[26] I. Kantorowitsch, G. Akilow, Funktionalanalysis in normierten Räumen, 2nd, revised ed., Akademie-Verlag, Berlin, 1978.

[27] M. Kästner, B. Hammer, M. Biehl, T. Villmann, Functional relevance learning in generalized learning vector quantization, Neurocomputing 90 (9) (2012) 85–95.

[28] M. Kästner, D. Nebel, M. Riedel, M. Biehl, T. Villmann, Differentiable kernels in generalized matrix learning vector quantization, in: Proceedings of the International Conference on Machine Learning Applications (ICMLA'12), IEEE Computer Society Press, Los Alamitos, 2012, pp. 1–6.

[29] G.A. Khuwaja, An adaptive combined classifier system for invariant face recognition, Digital Signal Process.: A Rev. J. 12 (January (1)) (2002) 21–46.

[30] K. Kim, M. Franz, B. Schölkopf, Kernel Hebbian Algorithm for Iterative Kernel Principal Component Analysis, Technical Report 109, Max-Planck-Institute for Biological Cybernetics, June 2003.

[31] K. Kim, M. Franz, B. Schölkopf, Iterative kernel principal component analysis for image modelling, IEEE Trans. Pattern Anal. Mach. Intell. 27 (9) (2005) 1351–1366.

[32] G. Lumer, Semi-inner-product spaces, Trans. Am. Math. Soc. (1961) 29–43.

[33] J. Mercer, Functions of positive and negative type and their connection with the theory of integral equations, Philos. Trans. R. Soc. Lond. A 209 (1909) 415–446.

[34] E. Mwebaze, P. Schneider, F.-M. Schleif, S. Haase, T. Villmann, M. Biehl, Divergence based learning vector quantization, in: M. Verleysen (Ed.), Proceedings of European Symposium on Artificial Neural Networks (ESANN'2010), Evere, Belgium, 2010, pp. 247–252 d-side publications.

[35] B. Nath, On a generalization of semi-inner product spaces, Math. J. Okayama Univ. 15 (1/1) (1971) 1–6.

[36] B. Nath, Topologies on generalized semi-inner product spaces, Compos. Math. 23 (3) (1971) 309–316.

[37] E. Oja, Neural networks, principle components and subspaces, Int. J. Neural Syst. 1 (1989) 61–68.

[38] E. Oja, Nonlinear pca: algorithms and applications, in: Proceedings of the World Congress on Neural Networks Portland, Portland, 1993, pp. 396–400.
[39] E. Pekalska, R. Duin, The Dissimilarity Representation for Pattern Recognition: Foundations and Applications, World Scientific, Singapore, 2006.
[40] J. Ramsay, B. Silverman, Functional Data Analysis, 2nd ed., Springer Science+Media, New York, 2006.
[41] J. Retherford, R. James, Unconditional bases and best approximation in Banach spaces, Bull. Am. Math. Soc. 75 (1) (1969) 108–112.
[42] F. Riesz, B. Sz.-Nagy, Vorlesungen über Functionalanalysis, 4th ed., Verlag Harri Deutsch, Frankfurt/M., 1982.
[43] F. Rossi, N. Delannay, B. Conan-Gueza, M. Verleysen, Representation of functional data in neural networks, Neurocomputing 64 (2005) 183–210.
[44] T. Sanger, Optimal unsupervised learning in a single-layer linear feedforward neural network, Neural Netw. 12 (1989) 459–473.
[45] A. Sato, K. Yamada, Generalized learning vector quantization, in: D.S. Touretzky, M.C. Mozer, M.E. Hasselmo (Eds.), Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference, MIT Press, Cambridge, MA, USA, 1996, pp. 423–429.
[46] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, 2002.
[47] P. Schneider, B. Hammer, M. Biehl, Adaptive relevance matrices in learning vector quantization, Neural Comput. 21 (2009) 3532–3561.
[48] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 14 (7) (1998) 1299–1319.
[49] B. Silverman, Smoothed functional principal components analysis by the choice of norm, Ann. Stat. 24 (1) (1996) 1–24.
[50] M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis and Machine Vision, 2nd ed., Brooks Publishing, Baltimore, 1998.
[51] M. Soriano, L. Garcia, C. Saloma, Fluorescent image classification by major color histograms and a neural network, Opt. Express 8 (February (5)) (2001) 271–277.
[52] I. Steinwart, On the influence of the kernel on the consistency of support vector machines, J. Mach. Learn. Res. 2 (2001) 67–93.
[53] H. Thodberg, Tecator meat sample dataset, Available on: ⟨http://lib.stat.cmu.edu/datasets/tecator⟩.
[54] T. Villmann, S. Haase, Divergence based vector quantization, Neural Comput. 23 (5) (2011) 1343–1392.
[55] T. Villmann, S. Haase, A note on gradient based learning in vector quantization using differentiable kernels for Hilbert and Banach spaces, Mach. Learn. Rep. 6 (MLR-02-2012) (2012) 1–29, ISSN:1865-3960 ⟨http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_02_2012.pdf⟩.
[56] T. Villmann, B. Hammer, Functional principal component learning using Ojas method and Sobolev norms, in: J. Principe, R. Miikkulainen (Eds.), Advances in Self-Organizing Maps – Proceeding of the Workshop on Self-Organizing Maps (WSOM), Springer, Berlin, 2009, pp. 325–333.
[57] T. Villmann, M. Kästner, D. Nebel, M. Riedel, ICMLA face recognition challenge – results of the team Computational Intelligence Mittweida, in: Proceedings of the International Conference on Machine Learning Applications (ICMLA'12), IEEE Computer Society Press, Los Alamitos, 2012, pp. 7–10.
[58] U. von Luxburg, O. Bousquet, Distance-based classification with Lipschitz functions, J. Mach. Learn. Res. 5 (2004) 669–695.
[59] H. Zhang, Y. Xu, J. Zhang, Reproducing kernel Banach spaces for machine learning, J. Mach. Learn. Res. 10 (2009) 2741–2775.
[60] H. Zhang, J. Zhang, Generalized semi-inner products with applications to regularized learning, J. Math. Anal. Appl. 372 (2010) 181–196.

Mandy Lange received both, her diploma in Applied Mathematics in 2008 and her M.Sc. degree in Discrete and Computer-oriented Mathematics in 2011, from the University of Applied Sciences in Mittweida, Germany. Since 2011 she is a Ph.D. student and a member of the Computational Intelligence Group at this university. Her research focus is in vector quantization and non-standard metrics.

Michael Biehl received a Ph.D. in Physics from the University of Giessen, Germany, in 1992 and the Venia Legendi in Theoretical Physics from the University of Würzburg, Germany 1996. He is currently Associate Professor with Tenure in Computer Science at the Johann Bernoulli Institute, University of Groningen, The Netherlands. His main research interest is in the theory of machine learning techniques and their application in the life sciences. He is furthermore active in the modelling and simulation of complex physical system. He has co-authored more than 100 publications in international journals and conferences, preprint versions and further information can be obtained from http://www.cs.rug.nl/~biehl.

Thomas Villmann holds a diploma degree in Mathematics, received his Ph.D. in Computer Science in 1996 and his venia legendi in the same subject in 2005, all from University of Leipzig/Germany. From 1997 to 2009 he led the Computational Intelligence Group of the Clinic for Psychotherapy at Leipzig University. Since 2009 he is a full Professor for Technomathematics/Computational Intelligence at the University of Applied Sciences Mittweida (Saxonia) Germany. He is founding member of the German chapter of the European Neural Network Society (ENNS) and acts as president since 2011. He serves as an Associate Editor for Neural Processing Letters. His research focus includes the theory of prototype based clustering and classification, non-standard metrics and relevance learning, information theory and their application in pattern recognition for use in medicine, bioinformatics, remotesensing, hyperspectral analysis and others. He co- /authored more than 75 articles in scientific journals and more than 200 contributions in books and reviewed conference proceedings.