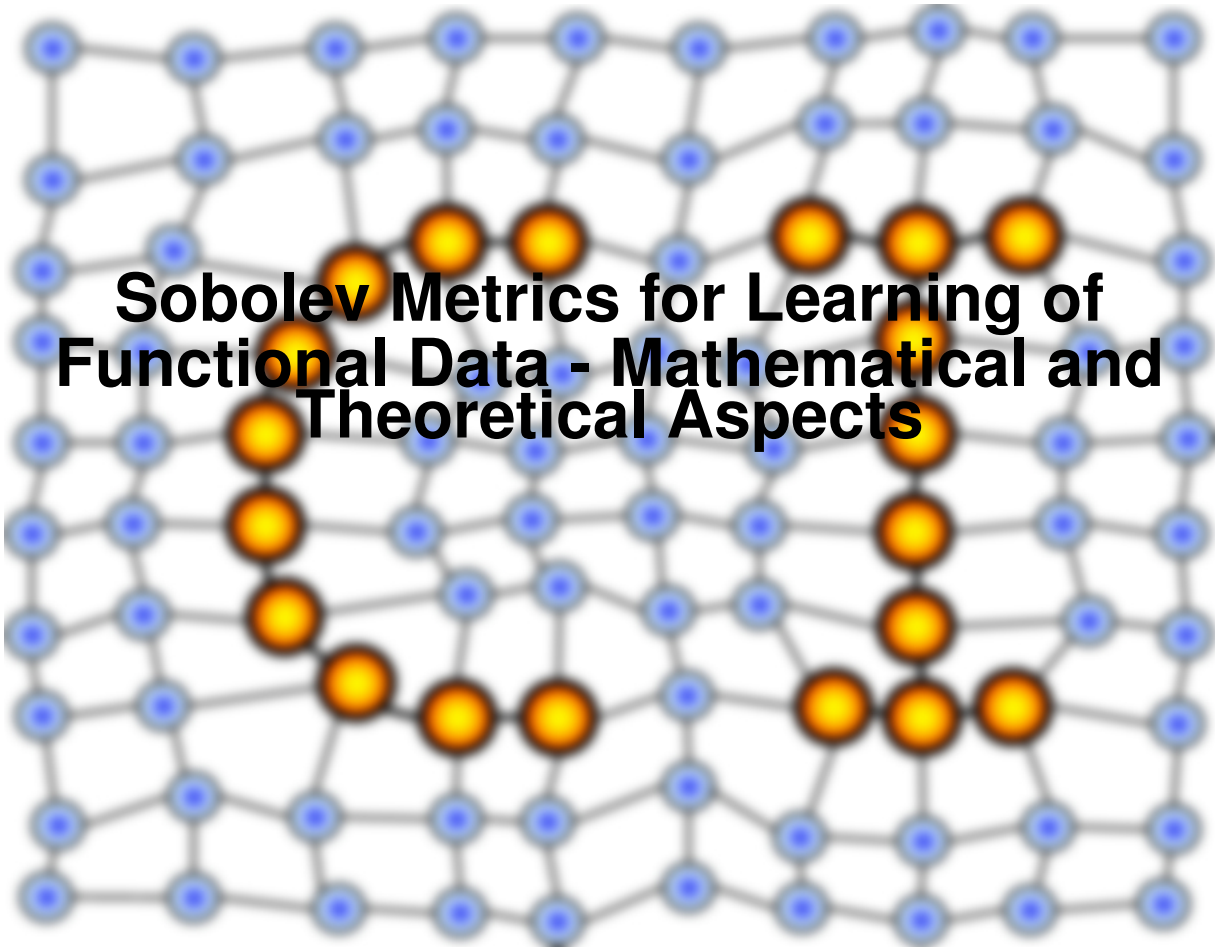
MACHINE LEARNING REPORTS



Sobolev Metrics for Learning of Functional Data - Mathematical and Theoretical Aspects

Report 03/2007

Thomas Villmann

- submitted 26th November 2007, published 7th December 2007 -

Abstract

We study the utilization of functional metrics for learning of functional data. In particular we investigate the metrics based on the Sobelev metric which can be related top a respective inner product. This offers capabilities for adequate data processing of functional data taking into acccount the dependencies within the functional data vectors. We outline these possibilities and give the mathematical derivations as well as the theoretical basis for two basic applications: functional principal component analysis based on Oja's algorithm and prototype based vector quantization.

1 Introduction

Data processing of functional data is a challenging topic in machine learning data analysis [RS06]. There is a broad area of application: biomedicine, chemometrics and chemistry, physics and astrophysics as well as geosciences and remote sensing analysis, to name just a few. The problems to be solved ranges from time series analysis and prediction, identification of characteristic patterns and classification to spectral data analysis.

Usually the functional data are given as high-dimensional vectors $\mathbf{v}$ with components $v_i = f(x_i)$, $x_i \in \mathbb{R}$. The characteristic feature which distinguishes usual vectorial data from functional once is that the vector components v_i are not independent. However, there exist only few methods in machine learning which take into account this property [LV05],[RDCGV05].

In this work we investigate the usability of *Sobolev-metrics* for adequate handling of functional data in data analysis. The main advantage of this metric in comparison to other methods is that it can be derived from a inner product defined for a special function class which has special assumptions on differentiability. We show, how this methodology can me plugged into machine learning methods. As basic examples we demonstrate this for two important methods: functional principal component analysis (FPCA) based on the Oja's algorithm and prototype based vector quantization.

The paper is organized as follows: First we investigate functional metrics in the light of applicability for adaptive learning methods. In particular we will concentrate on Sobolev-metrics and norms. Subsequently, we will outline the application of the Sobolev-inner-product, which is in direct dependence on the Sobolev-norm, for FPCA. We give the mathematical derivations and foundations for the incorporation of both, inner product and norm, into PCA-learning and prototype based vector quantization. A short conclusion concludes this paper.

2 Functional metrics, norms and inner products

In this chapter we provide all the ingredients which are needed for the application of functional norms and related inner products for machine learning algorithms.

2.1 Functional norm according to LEE & VERLEYSEN

There exist only few methods which are specifically designed to process functional data paying attention to the special property of inherent dependencies. Most of them deal with the function description in terms of basis functions like Fourier-, Laplace-, wavelet expansions or others, such that methods can be applied to the respective coordinate space. An interesting alternative was proposed by LEE&VERLEYSEN in [LV05]. It is based on the usual *Minkowski-p-norm*

$$\|f\|_p = \sqrt{\left(\int |f(x) \cdot f(x)|^p \, dx\right)^{\frac{1}{p}}} \tag{2.1}$$

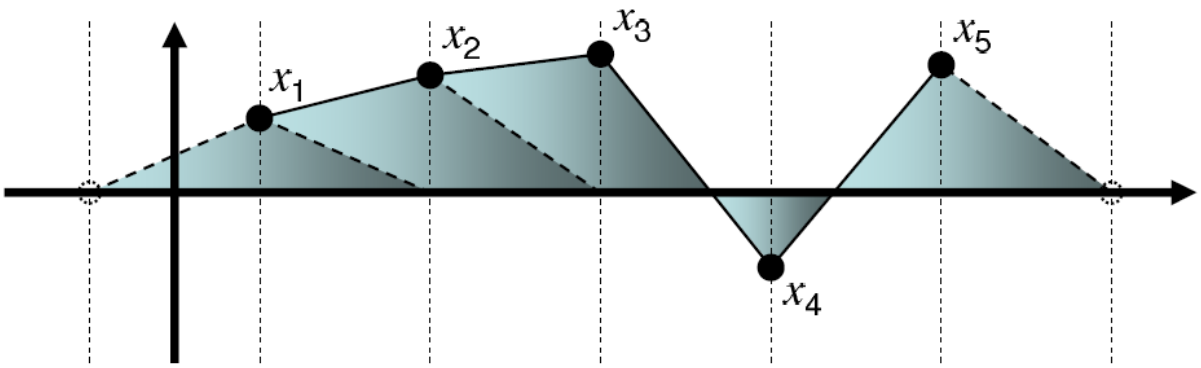


Figure 1: Illustration of the $\|\mathbf{f}\|_p^{fc}$ −norm. The function graph is given as $f_k = f(x_k)$. The norm involves the areas of the triangles located on the left and right sides of each coordinate. (Figure from [LV05])

or in its vectorial form

$$\|\mathbf{f}\|_p = \sqrt{\left(\sum_{k=1}^{D} |f_k \cdot f_k|^p\right)^{\frac{1}{p}}} \tag{2.2}$$

with $\mathbf{f} = (f_1, \ldots, f_D)$ and $f_k = f(x_k)$, $x_k \in X \subseteq \mathbb{R}$ whereby we assume w. l. o. g. that $x_k < x_{k+1}$ for all k. The functional norm by LEE&VERLEYSEN motivated by geometrical considerations is defined as

$$\|\mathbf{f}\|_p^{fc} = \left(\sum_{k=1}^{D} \left(B_k(\mathbf{f}) + B_k(\mathbf{f})\right)^p\right)^{\frac{1}{p}} \tag{2.3}$$

with

$$A_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2}\frac{v_k^2}{|v_k|+|v_{k-1}|} & \text{if } 0 > v_k v_{k-1} \end{cases} \quad \text{and } B_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2}\frac{v_k^2}{|v_k|+|v_{k+1}|} & \text{if } 0 > v_k v_{k+1} \end{cases} \tag{2.4}$$

and the usual choice $\tau = 1$. The functional dependencies are involved by areas of the triangles located on the left and right sides of each coordinate, see Fig. 1.

The $\|\mathbf{f}\|_p^{fc}$-norm is a generalization of the usual $\|\mathbf{f}\|_p$-norm. As usual for every norm, an accompanying distance measure can be defined by

$$\delta_p^{fc}(\mathbf{f}, \mathbf{g}) = \|\mathbf{f} - \mathbf{g}\|_p^{fc} \tag{2.5}$$

with $\delta_p^{fc}(\mathbf{f}, \mathbf{g}) \leq \|\mathbf{f} - \mathbf{g}\|_p$ as it was shown in [LV05]. In particular, one has

$$\|\mathbf{f}\|_p^{fc} = \|\mathbf{f}\|_p \text{ iff } \forall k\ f_k \geq 0 \text{ or } \forall k\ f_k \leq 0. \tag{2.6}$$

From a machine learning point of view, it is interesting that the quadratic functional metric $\left(\delta_2^{fc}(\mathbf{f}, \mathbf{g})\right)^2$ is 'differentiable' (in the sense of difference quotients) for the choice $p = 2$.

$$\frac{\partial\left(\delta_2^{fc}(\mathbf{f}, \mathbf{g})\right)^2}{\partial g_k} \overset{def.}{=} \frac{1}{2}\left(2 - U_{k-1} - U_{k+1}\right)\left(V_{k-1} + V_{k+1}\right)\triangle_k \tag{2.7}$$

with

$$
\begin{aligned}
U_{k-1} &= \begin{cases} 0 & \text{if } 0 \leq \triangle_k \triangle_{k-1} \\ \left(\frac{\triangle_{k-1}}{|\triangle_k|+|\triangle_{k-1}|}\right)^2 & \text{if } 0 > \triangle_k \triangle_{k-1} \end{cases} \\
U_{k+1} &= \begin{cases} 0 & \text{if } 0 \leq \triangle_k \triangle_{k+1} \\ \left(\frac{\triangle_{k+1}}{|\triangle_k|+|\triangle_{k+1}|}\right)^2 & \text{if } 0 > \triangle_k \triangle_{k+1} \end{cases} \\
V_{k-1} &= \begin{cases} 1 & \text{if } 0 \leq \triangle_k \triangle_{k-1} \\ \frac{|\triangle_k|}{|\triangle_k|+|\triangle_{k-1}|} & \text{if } 0 > \triangle_k \triangle_{k-1} \end{cases} \\
V_{k+1} &= \begin{cases} 1 & \text{if } 0 \leq \triangle_k \triangle_{k+1} \\ \frac{|\triangle_k|}{|\triangle_k|+|\triangle_{k+1}|} & \text{if } 0 > \triangle_k \triangle_{k+1} \end{cases}
\end{aligned}
$$

and $\triangle_j = f_k - g_k$ [LV05].

However, the $\|\mathbf{f}\|_p^{fc}$-norm has a disadvantage. It can not be derived from an inner product. We will see later for FPCA (see Sec.3) that this feature can be used for adaptive FPCA. We will proof the following lemma:

Lemma 1 *The functional norm $\|\cdot\|_p^{fc}$ cannot be related to an inner product.*

Proof. The proof stresses the parallelogram equation. A norm $\|\cdot\|$ can be derived from an inner product iff the parallelogram equation

$$\|\mathbf{f}-\mathbf{g}\|^2 + \|\mathbf{f}+\mathbf{g}\|^2 = 2\left(\|\mathbf{f}\|^2 + \|\mathbf{g}\|^2\right) \tag{2.8}$$

is fulfilled. We will see that this is not the case for the $\|\mathbf{f}\|_p$-norm: For this purpose, we consider to vectorial functions $\mathbf{f}$ and $\mathbf{g}$ with $\forall k\ f_k \geq 0$ and $g_k \geq 0$. Thus $\|\mathbf{f}\|_p^{fc} = \|\mathbf{f}\|_p$ and $\|\mathbf{g}\|_p^{fc} = \|\mathbf{g}\|_p$ according to (2.6). Obviously, $f_k + g_k \geq 0$ for all k and, hence, $\|\mathbf{f}+\mathbf{g}\|_p^{fc} = \|\mathbf{f}+\mathbf{g}\|_p$. We further assume that there exist exactly one k^* such that $f_{k^*} < g_{k^*}$ whereas $f_k > g_k$ holds otherwise. It can easily be computed that for the difference vector $\mathbf{z} = \mathbf{f} - \mathbf{g}$ the inequality $\|\mathbf{z}\|_p^{fc} \neq \|\mathbf{z}\|_p$ is valid because of $A_{k^*}(\mathbf{z}) + B_{k^*}(\mathbf{z}) \neq |z_{k^*}|^p$ whereas $A_k(\mathbf{z}) + B_k(\mathbf{z}) = |z_k|^p$ otherwise. This completes the proof. ■

Beside this impossibility it is difficult to generalize the $\|\mathbf{f}\|_p^{fc}$-norm to integrable functions. Therefore, we now focus on a norm which is naturally defined by an inner product but also involving the functional dependency. The idea of application of functional norms for data analysis was demonstrated for FPCA in [Sil96].

2.2 Minkowski–metrics and respective Sobolev-metrics: the function spaces $\mathcal{L}_p$ and $\mathcal{S}_p$

We start with the usual $p-$*Minkowski-inner-product* ($p-$MIP). Let f, g be real-valued functions over $X \subseteq \mathbb{R}$. Then the inner product $p-$MIP is defined as

$$\langle f, g\rangle_p = \left(\int_X |f(x)\, g(x)|^p\, dx\right)^{\frac{1}{p}}. \tag{2.9}$$

for real-valued absolute-integrable functions f and g. The accompanied $p-$*Minkowski-norm* is ($p-$MN)

$$\|f\|_p = \sqrt{\langle f, f\rangle_p} \tag{2.10}$$

For discrete representations we analogously have

$$\langle f, g\rangle_p = \left(\sum_{k=1}^{n} |f_k \cdot g_k|^p\right)^{\frac{1}{p}} \tag{2.11}$$

The respective function space is $\mathcal{L}_p(X)$, which forms a Hilbert-space [Tri89]. The special case $p = 2$ can be seen as the Euclidean inner product

$$\begin{aligned} \langle f, g\rangle_{\mathrm{E}} &= \int_X f(x)\, g(x)\, dx && (2.12)\\ &= \langle f, g\rangle_1 && (2.13) \end{aligned}$$

We now introduce the $p-$*Sobolev-inner-product* ($p-$SIP) of degree k with parameter $\alpha > 0$. Let $f, g \in \mathcal{C}^K(X)$ be $K-$times continuous-differentiable integrable functions (in the Lebesgue sense) over X. Then an inner product can be defined by

$$\langle f, g\rangle^{\mathcal{S}}_{p,\alpha} = \langle f, g\rangle_p + \alpha \left\langle D^{(k)} f, D^{(k)} g\right\rangle_p \tag{2.14}$$

with $D^{(k)}$ being the kth differential operator and $D^{(0)} = \mathrm{id}$ is the identity [KF75]. The requirements for an inner product are obviously fulfilled due to the linearity of the differential operator $D^{(k)}$. The accompanied $p-$Sobolev-norm ($p-$SN) of degree k is

$$\|f\|^{\mathcal{S}}_{p,k,\alpha} = \sqrt{\langle f, f\rangle^{\mathcal{S}}_{p,\alpha}} \tag{2.15}$$

which defines the Sobolev distance of degree k

$$s^{\mathcal{S}}_{p,k,\alpha}(f, g) = \|f - g\|^{\mathcal{S}}_{p,\alpha}. \tag{2.16}$$

The space $\mathcal{C}^K(X)$ together with the norm (2.15) forms a *Banach-space* $\mathcal{S}_{p,k,\alpha}$ and obviously one has $\mathcal{S}_{p,k,\alpha} \subset \mathcal{L}_p$.[1] For the special case $p = 2$ the space $\mathrm{S}_{2,\alpha} = \mathrm{S}_\alpha$ becomes a Hilbert-space [KA78]. Moreover, for this case an interesting connection to the Fourier-analysis can be made using the *Parsevals-equation*: Let $\hat{f}$ be the Fourier-transform of f

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \exp(-i\omega x)\, dx \tag{2.17}$$

or for discrete valued functions g given in vectorial form

$$\hat{g}(\omega_k) = \sum_{j=1}^{N-1} g_j \exp\left(-i2\pi \frac{k \cdot j}{N}\right) \tag{2.18}$$

with $\omega_k = \frac{2\pi k}{N}$. Then, the norm $\|\cdot\|^{\mathcal{S}}_k = \|\cdot\|^{\mathcal{S}}_{2,k,1}$ can be written as

$$\|f\|^{\mathcal{S}}_k = \sqrt{\int_{-\infty}^{\infty} (1+\omega)^k \left|\hat{f}(\omega)\right|^2 d\omega} \tag{2.19}$$

[1] Yet, there are more general definitions possible. We here restrict ourself to this simplification which are sufficient for the most applications of machine learning problems. For a further reading we refer to [KA78] or [KF75].

or in its dicrete form

$$\|\mathbf{g}\|_k^{\mathcal{S}} = \sqrt{\sum_{j=1}^{N-1} (1+\omega_j)^k \left|\hat{g}(\omega_j)\right|^2} \tag{2.20}$$

Clearly, all the other definitions can also be transferred to vectorial representations of functions replacing the integrals by sums and the differential operators $D^{(k)}$ by difference operators $\triangle^{(k)}$.

2.3 Some remarks about statistical values and (discrete) inner products

We return to the Euclidean inner product (2.12) in the discrete form

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\mathrm{E}} = \sum_{k=1}^{D} f_k g_k \tag{2.21}$$

Let $\mathbf{1}$ be the vector $\mathbf{1} = (1, \ldots, 1)$. Then

$$\bar{f} = \langle \mathbf{f}, \mathbf{1} \rangle / D \tag{2.22}$$

is the mean of $\mathbf{f}$ and

$$\sigma_{\mathbf{f}} = \left\langle \mathbf{f} - \bar{f}\mathbf{1}, \mathbf{f} - \bar{f}\mathbf{1} \right\rangle / D \tag{2.23}$$

$$= \left\| \mathbf{f} - \bar{f}\mathbf{1} \right\|^2 \tag{2.24}$$

its variance, whereby $\|\cdot\|$ is the usual quadratic Euclidean norm. Analogously one gets

$$\sigma_{\mathbf{f},\mathbf{g}} = \left\langle \mathbf{f} - \bar{f}\mathbf{1}, \mathbf{g} - \bar{g}\mathbf{1} \right\rangle / D \tag{2.25}$$

for the covariance.

3 Functional principal component analysis (FPCA)

In this chapter we will give two approaches for FPCA. The first method uses the function representation in terms of orthogonal basis functions, whereas the second approach utilizes the Sobolev-inner-product 2-SIP.

3.1 FPCA based on orthogonal basis functions

In this section we assume that the real function f, g over $X \subseteq \mathbb{R}$ can be represented by orthogonal basis functions ϕ_k which form a basis of the functional space containing f and g. Thereby, orthogonality is defined by $\langle \phi_k, \phi_j \rangle_{\mathrm{E}} = \delta_{k,j}$. The basis may contain a infinite number of basis functions. Prominent examples are the the set of monomials $1, x, x^2, \ldots, x^k, \ldots$ or the Fourier-system of $\sin(k\omega x)$,$\cos(k\omega x)$ with $k = 0, 1, 2, \ldots$.

Using a basis system of K linear independent functions an arbitrary function f can be approximated by

$$f(x) = \sum_{k=1}^{K} \alpha_k \phi_k(x) \tag{3.1}$$

which can be seen as a discrete Euclidean inner product $\langle \alpha, \phi(x) \rangle_{\mathrm{E}}$ of the coordinate vector $\alpha = (\alpha_1, \ldots, \alpha_k)^\top$ with the function vector $\phi = (\phi_1(x), \ldots, \phi_k(x))^\top$. If the basis functions are the Fourier functions and f given as functional vector $\mathbf{f}$, then the Sobolev-norm $\|\mathbf{f}\|_k^S$ can be immediately computed via (2.20).

We denote by $\mathcal{A}$ the function space spanned by all basis functions ϕ_k:

$$\mathcal{A} = span(\phi_1, \ldots, \phi_k). \tag{3.2}$$

Following the suggestions in [RS06] and [RDCGV05] to transfer the ideas of usual multivariate PCA to FPCA. We consider the Euclidean inner product

$$\begin{aligned} \langle f, g \rangle_{\mathrm{E}} &= \int_X f(x)\, g(x)\, dx && (3.3)\\ &= \sum_{k=1}^{K} \sum_{j=1}^{K} \alpha_k \beta_j \int_X \phi_k(x)\, \phi_j(x)\, dx && (3.4)\\ &= \sum_{k=1}^{K} \sum_{j=1}^{K} \alpha_k \beta_j \langle \phi_k, \phi_j \rangle_{\mathrm{E}} && (3.5) \end{aligned}$$

whereby in the second line the Fubini-lemma was used to exchange the integral and the sums. Let $\mathbf{\Phi}$ be the symmetric matrix spanned by $\Phi_{k,j} = \langle \phi_k, \phi_j \rangle_{\mathrm{E}}$ using the symmetry of an inner product. Using this definition, the last equation can be rewritten as $\langle f, g \rangle_{\mathrm{E}} = \langle f, g \rangle_{\mathbf{\Phi}}$ with the new inner product

$$\langle f, g \rangle_{\mathbf{\Phi}} = \alpha^\top \mathbf{\Phi} \beta \tag{3.6}$$

We remark that $\mathbf{\Phi}$ is independent of both f and g. If the basis is orthogonal, $\mathbf{\Phi}$ is diagonal with entries $\Phi_{k,k} = 1$. Thus, the inner product of functions is reduced to the inner product of the coordinate vectors

$$\langle f, g \rangle_{\mathrm{E}} = \langle \alpha, \beta \rangle_{\mathrm{E}} \tag{3.7}$$

For handling non-orthogonal basis systems we refer to [RDCGV05].

Looking at (3.7) we see that performing elementary vector operations on the coordinate vectors in the Euclidean space $\mathbb{R}^K$ equipped with the (discrete) Euclidean inner product (2.12) is equivalent to the respective operations in the inner product space $\mathcal{A}$ with the Euclidean inner product (3.3). This statement allows a straightforward application to FPCA: FPCA can performed on a set $F = \{f_k\}_{k=1\ldots N}$ of functions f_k by usual vectorial PCA analysis of the respective set of coordinate vectors α_k as explained in [RS06].

3.2 Oja's PCA-learning for functional data

E. OJA developed an online-learning algorithm to determine the first principal component of data vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$ adaptively [Oja89],[Oja93]. The first principal component $\mathbf{w}$ related to the maximum eigen value for the data set V is obtained by the stochastic adaptation. For a single input the learning rule is

$$\triangle \mathbf{w} = \epsilon O(\mathbf{v} - O\mathbf{w}) \tag{3.8}$$

with O being the output

$$O = \mathbf{v}^\top \cdot \mathbf{w} \tag{3.9}$$
$$= \mathbf{w}^\top \cdot \mathbf{v} \tag{3.10}$$

which judges the correlation strength according to the Hebbian postulate of coincidence for neural connections between nerve fibres $\mathbf{v}$ and the neural dendrites realized by the connection strength $\mathbf{w}$ [Hay94],[KS91],[RMS92]. We obtain

$$\triangle \mathbf{w} = \epsilon \left(\mathbf{v}^\top \cdot \mathbf{w}\right) \left(\mathbf{v} - \left(\mathbf{v}^\top \cdot \mathbf{w}\right) \mathbf{w}\right) \tag{3.11}$$
$$= \epsilon \mathbf{v}\mathbf{v}^\top \cdot \mathbf{w} - \left(\mathbf{w}^\top \cdot \mathbf{v}\right) \left(\mathbf{v}^\top \cdot \mathbf{w}\right) \mathbf{w} \tag{3.12}$$
$$= \epsilon \mathbf{w} \cdot \mathbf{v}^\top \mathbf{v} - \mathbf{w} \cdot \left(\mathbf{v}^\top \mathbf{v}\right) \cdot \mathbf{w}^\top \mathbf{w} \tag{3.13}$$

We remark here that the output O can also be written as the Euclidean inner product $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathrm{E}}$

$$\triangle \mathbf{w} = \epsilon \langle \mathbf{w}, \mathbf{v} \rangle_{\mathrm{E}} \left(\mathbf{v} - \left(\langle \mathbf{w}, \mathbf{x} \rangle_{\mathrm{E}}\right)^2 \mathbf{w}\right). \tag{3.14}$$

Obviously, this algorithm can be immediately applied to the above outlined approach of FPCA based on function representations using orthogonal basis functions. However, if the functional data are given in vectorial form, there exist an interesting alternative. Instead of the Euclidean scalar product we formally plug the $p-$SSP of degree k (2.14) into the basic Oja-learning rule (3.14).

In particular we focus on the $1-$SIP

$$\langle \mathbf{w}, \mathbf{v} \rangle_{1,\alpha}^{\mathcal{S}} = \langle \mathbf{w}, \mathbf{v} \rangle_1 + \alpha \left\langle D^{(k)}\mathbf{w}, D^{(k)}\mathbf{v} \right\rangle_1 \tag{3.15}$$
$$= . \langle \mathbf{w}, \mathbf{v} \rangle_{\mathrm{E}} + \alpha \left\langle D^{(k)}\mathbf{w}, D^{(k)}\mathbf{v} \right\rangle_{\mathrm{E}} \tag{3.16}$$

. of degree k. We denote by υ the vector $D^{(k)}\mathbf{v}$ and by ω the vector $D^{(k)}\mathbf{w}$ and replace in the original learning rule (3.14) $\langle \mathbf{w}, \mathbf{v} \rangle_{\mathrm{E}}$ by $\langle \mathbf{w}, \mathbf{v} \rangle_{1,\alpha}^{\mathcal{S}}$. Then we obtain

$$\triangle \mathbf{w} = \epsilon \left[\langle \mathbf{v}, \mathbf{w} \rangle_{1,\alpha}^{S} \left(\mathbf{v} - \langle \mathbf{v}, \mathbf{w} \rangle_{1,\alpha}^{S} \mathbf{w}\right)\right] \tag{3.17}$$
$$= \epsilon \left[\left(\langle \mathbf{w}, \mathbf{v} \rangle_{\mathrm{E}} + \alpha \langle \omega, \upsilon \rangle_{\mathrm{E}}\right) \left(\mathbf{v} - \left(\langle \mathbf{w}, \mathbf{v} \rangle_{\mathrm{E}} + \alpha \langle \omega, \upsilon \rangle_{\mathrm{E}}\right) \mathbf{w}\right)\right] \tag{3.18}$$
$$= \epsilon \left[\mathbf{v} \left(\mathbf{v}^\top \cdot \mathbf{w} + \alpha \upsilon^\top \cdot \omega\right) - \left(\mathbf{w}^\top \cdot \mathbf{v} + \alpha \omega^\top \cdot \upsilon\right) \left(\mathbf{v}^\top \cdot \mathbf{w} + \alpha \upsilon^\top \cdot \omega\right) \mathbf{w}\right] \tag{3.19}$$
$$= \epsilon \left[\begin{array}{c} \mathbf{v} \left(\mathbf{v}^\top \cdot \mathbf{w} + \alpha \upsilon^\top \cdot \omega\right) - \\ \left(\mathbf{w}^\top \cdot \mathbf{v}\mathbf{v}^\top \cdot \mathbf{w} + \mathbf{w}^\top \cdot \mathbf{v}\alpha\upsilon^\top \cdot \omega + \alpha\omega^\top \cdot \upsilon\mathbf{v}^\top \cdot \mathbf{w} + \alpha\omega^\top \cdot \upsilon\alpha\upsilon^\top \cdot \omega\right) \mathbf{w} \end{array}\right] \tag{3.20}$$
$$= \epsilon \left[\begin{array}{c} \mathbf{v}\mathbf{v}^\top \cdot \mathbf{w} - \left(\mathbf{w}^\top \cdot \mathbf{v}\mathbf{v}^\top \cdot \mathbf{w}\right) \mathbf{w} \\ + \alpha \mathbf{v}\upsilon^\top \cdot \omega - \left(\alpha^2 \omega^\top \cdot \upsilon\upsilon^\top \cdot \omega\right) \mathbf{w} \\ - \alpha \left(\omega^\top \cdot \upsilon\mathbf{v}^\top \cdot \mathbf{w} + \mathbf{w}^\top \cdot \mathbf{v}\upsilon^\top \cdot \omega\right) \mathbf{w} \end{array}\right] \tag{3.21}$$

as new update rule for one given data vector. We denote the final vector $\mathbf{w}^*$ of this dynamic as *'functionally modified (first) principal component'* (FMPCA).

It is clear, for small (vanishing) values of α, the original Oja-rule is preserved. For non-vanishing α with relevant magnitude compared to the variance of the covariance matrix $\mathbf{C}$ of the data vectors $\mathbf{v}$, the influence of the second term of the Sobolev metric becomes significantly such that a deviation to the usual fist principal component is observed for the final weight vector $\mathbf{w}^*$. If the choice of the degree of the differential operator is $k = 2$ there it has the following interpretation: According to [RS06] (p.41), $D^{(2)}f$ measures the *curvature* of the function f. Smooth functions have low curvature. Hence, the higher the curvature of the functional vectors of a data set, the more the FMPCA deviates from the usual first principal component, i.e. it is more adapted to the curvature in comparison to the non-functional result.

4 Functional vector quantization using the Sobolev distance

In this section we focus of the application of the Sobolev distance $s^{\mathcal{S}}_{p,k,\alpha}$ from (2.16) to prototype based vector quantization algorithms. Well known robust vector quantizers are KOHONEN'S self-organizing map (SOM) [Koh95] (also in the variant proposed by HESKES, [HES99]) or the neural gas (NG) provided by MARTINETZ [MBS93] for unsupervised learning or fuzzy-labeled NG (FLNG, [VHS+06]) and fuzzy-labeled SOM (FLSOM, [VSMH07]) for semi-supervised vector quantization. Usually, all these algorithms try to minimize some variants of the quadratic Euclidean error between the prototypes $\mathbf{w}_k \in \mathbb{R}^n$ and the data vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$ by stochastic gradient decsent. However, the applications of non-standard metrics became a challenging topic which is still under ongoing research [HV05],[VSMH07].

Two basic ingredients have to be considered in this line: the determination of the best matching prototype $\mathbf{w}_{k^*}$ for a given data vector $\mathbf{v}$ and the adaptation rule for the prototypes, both based on the quadratic Euclidean distance in the non-functional algorithms. We now replace this distance by the quadratic Sobolev distance of degree 2 (consistently to the above Oja' algorithm)

$$\begin{aligned} s_\alpha(\mathbf{f}, \omega_k) &= \left(s^{\mathcal{S}}_{2,2,\alpha}(\mathbf{f}, \omega_k)\right)^2 && (4.1)\\ &= \langle \mathbf{f}-\omega_k, \mathbf{f}-\omega_k\rangle_2 + \alpha \left\langle D^{(2)}(\mathbf{f}-\omega_k), D^{(2)}(\mathbf{f}-\omega_k)\right\rangle_2 && (4.2)\\ &= \|\mathbf{f}-\omega_k\|_{\mathrm{E}} + \alpha \left\|D^{(2)}(\mathbf{f}-\omega_k)\right\|_{\mathrm{E}} && (4.3)\end{aligned}$$

for functional vectors $\mathbf{f} \in F \subseteq \mathbb{R}^n$ and functional prototypes ω_k. $\|\cdot\|_{\mathrm{E}}$ is the usual Euclidean norm. Thus, stochastic gradient descent on any cost function based on the distance $s_\alpha(\mathbf{f}, \omega_k)$ involves the second derivative

$$\frac{\partial s_\alpha(\mathbf{f}, \omega_k)}{\partial \omega_k} = \frac{\partial \|\mathbf{f}-\omega_k\|_{\mathrm{E}}}{\partial \omega_k} + \alpha \frac{\partial \left\|D^{(2)}(\mathbf{f}-\omega_k)\right\|_{\mathrm{E}}}{\partial \omega_k}. \tag{4.4}$$

We investigate for an arbitrary prototype $\omega = (\omega_1, \ldots, \omega_n)$ the single dimension j:

$$\frac{\partial s_\alpha(\mathbf{f}, \omega)}{\partial \omega_j} = \frac{\partial \|\mathbf{f}-\omega\|_{\mathrm{E}}}{\partial \omega_j} + \alpha \frac{\partial \left\|D^{(2)}(\mathbf{f}-\omega)\right\|_{\mathrm{E}}}{\partial \omega_j} \tag{4.5}$$

The first term can easily be computed by

$$\begin{aligned} \frac{\partial \|\mathbf{f}-\omega\|_{\mathrm{E}}}{\partial \omega_j} &= \frac{\partial\left(\sqrt{\sum_{l=1}^n (f_l-\omega_l)^2}\right)}{\partial \omega_j} && (4.6)\\ &= \frac{-2(f_j-\omega_j)}{\|\mathbf{f}-\omega\|_{\mathrm{E}}} && (4.7)\end{aligned}$$

The second term $\frac{\partial \left\|D^{(2)}(\mathbf{f}-\omega)\right\|_{\mathrm{E}}}{\partial \omega_j}$ contains the differential operator $D^{(2)} = D^{(1)} \circ D^{(1)}$.

Now, we assume that the function z is sampled and given in vectorial form $\mathbf{z} = (z_1, \ldots, z_n)$: $z_k = z(x_k)$, $x_k \in X \subseteq \mathbb{R}$ and $x_k < x_{k+1} = x_k + \triangle x$ is equidistantly distributed. Then $D^{(1)}\mathbf{z}$ can be approximated by the central difference

$$D^{(1)}\mathbf{z}\big|_j \approx \frac{z_{j+1} - z_{j-1}}{2 \triangle x} \tag{4.8}$$

and, hence, the second derivate is consistently approximated by the second central difference

$$D^{(2)}\mathbf{z}\big|_j \approx \frac{z_{j-2} - 2z_j + z_{j+2}}{4\left(\triangle x\right)^2} \tag{4.9}$$

$$= : d_j \tag{4.10}$$

Setting now $\mathbf{z} = \mathbf{f} - \omega$, the term $\frac{\partial \left\| D^{(2)}(\mathbf{f}-\omega)\right\|_{\mathrm{E}}}{\partial \omega_j}$ in (4.5) can be rewritten as

$$\frac{\partial \left\| D^{(2)}\left(\mathbf{f}-\omega\right)\right\|_{\mathrm{E}}}{\partial \omega_j} \approx \frac{1}{\sqrt{\sum_{l=1}^{n}\left(d_l\right)^2}} \cdot \frac{\partial S}{\partial \omega_j} \tag{4.11}$$

with the notation $S = \sqrt{\sum_{l=1}^{n}\left(d_l\right)^2}$. We consider the derivative

$$\frac{\partial S}{\partial \omega_j} = 2d_{j-2}\frac{\partial d_{j-2}}{\partial \omega_j} + 2d_j\frac{\partial d_j}{\partial \omega_j} + 2d_{j+2}\frac{\partial d_{j+2}}{\partial \omega_j} \tag{4.12}$$

with

$$\frac{\partial d_{j+2}}{\partial \omega_j} = \frac{\partial d_{j-2}}{\partial \omega_j} = \frac{\partial z_j}{\partial \omega_j} = \frac{-1}{4\left(\triangle x\right)^2}, \frac{\partial d_j}{\partial \omega_j} = \frac{-2\partial z_j}{\partial \omega_j} = \frac{2}{4\left(\triangle x\right)^2} \tag{4.13}$$

and obtain

$$\frac{\partial S}{\partial \omega_j} = \frac{-z_{j-4} + 3z_{j-2} - 4z_j + 3z_{j+2} - z_{j+4}}{8\left(\triangle x\right)^4} \tag{4.14}$$

which gives

$$\frac{\partial \left\| D^{(2)}\left(\mathbf{f}-\omega\right)\right\|_{\mathrm{E}}}{\partial \omega_j} \approx \frac{1}{\sqrt{\sum_{l=1}^{n}\left(d_l\right)^2}} \cdot \frac{-z_{j-4} + 3z_{j-2} - 4z_j + 3z_{j+2} - z_{j+4}}{8\left(\triangle x\right)^4} \tag{4.15}$$

Thus we have the final result that a prototype adaptation rule according to functional data can be obtained by replacing the derivative of the Euclidean distance between prototypes and data vectors by its functional counterpart for any of the above mentioned algorithms:

$$\frac{\partial s_\alpha\left(\mathbf{f},\omega\right)}{\partial \omega_j} \approx \frac{-2\left(f_j - \omega_j\right)}{\left\|\mathbf{f}-\omega\right\|_{\mathrm{E}}} + \alpha\frac{-z_{j-4} + 3z_{j-2} - 4z_j + 3z_{j+2} - z_{j+4}}{\left\|\mathbf{d}\right\|_{\mathrm{E}}\, 8\left(\triangle x\right)^4} \tag{4.16}$$

with $\mathbf{d} = (d_1, \ldots, d_l)$.

As consequence, a prototype based vector quantizer modified in this manner is emphasizing the curvature of the functional data, than its usual Euclidean counterpart.

5 Conclusion

In this report we provide theoretical aspects and the mathematical theory for the application of Sobolev-inner-products and -metrics for learning of functional data. Both concepts pay attention to the curvature of the functional data, i.e. its spatial dependencies within the data curve by incorporation of differential operators. Thus, approaches utilizing this aspect are more sensitive to the inherent functional structure of the functional

vectors than standard methods using the Euclidean counterparts. We showed that this idea can be applied in several domains: functional principal component analysis by means of Oja's-online-learning can be reformulated in terms of the Sobolev-inner-product whereas prototype based vector quantization like neural gas or self-organizing maps can be adapted for Sobolev-metrics, straightforward.

A remaining problem of the utilization of the Sobolev-inner-products and -metrics for learning of functional data is the determination of an adequate value α, which controls the influence of the differential operators. This parameter has to be chosen problem dependent but now theoretical suggestion can be made in general.

References

[Hay94] HAYKIN, Simon: *Neural Networks - A Comprehensive Foundation*. New York : IEEE Press, 1994

[Hes99] HESKES, T.: Energy functions for self-organizing maps. In: OJA, E. (Hrsg.); KASKI, S. (Hrsg.): *Kohonen Maps*. Amsterdam : Elsevier, 1999, S. 303–316

[HV05] HAMMER, B.; VILLMANN, Th.: Classification using non-standard metrics. In: VERLEYSEN, M. (Hrsg.): *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2005)*. Brussels, Belgium : d-side publications, 2005, S. 303–316

[KA78] KANTOROWITSCH, I.W.; AKILOW, G.P.: *Funktionalanalysis in normierten Räumen*. 2nd, revised. Berlin : Akademie-Verlag, 1978

[KF75] KOLMOGOROV, A.N.; FOMIN, S.V.: *Reelle Funktionen und Funktionalanalysis*. Berlin : VEB Deutscher Verlag der Wissenschaften, 1975

[Koh95] KOHONEN, Teuvo: *Springer Series in Information Sciences*. Bd. 30: *Self-Organizing Maps*. Berlin, Heidelberg : Springer, 1995. – (Second Extended Edition 1997)

[KS91] KANDEL, E.R.; SCHWARTZ, J.H.: *Principle of Neural Science*. 3rd. ed. New York : Elsevier, 1991

[LV05] LEE, J.; VERLEYSEN, M.: Generalization of the L_p norm for time series and its application to self-organizing maps. In: COTTRELL, M. (Hrsg.): *Proc. of Workshop on Self-Organizing Maps (WSOM) 2005*. Paris, Sorbonne, 2005, S. 733–740

[MBS93] MARTINETZ, Thomas M.; BERKOVICH, Stanislav G.; SCHULTEN, Klaus J.: 'Neural-Gas' Network for Vector Quantization and its Application to Time-Series Prediction. In: *IEEE Trans. on Neural Networks* 4 (1993), Nr. 4, S. 558–569

[Oja89] OJA, E.: Neural Networks, Principle Components And Suspaces. In: *International Journal of Neural Systems* 1 (1989), S. 61–68

[Oja93] OJA, E.: Nonlinear PCA: Algorithms and Applications. In: *Proc. Of the World Congress on Neural Networks Portland*. Portland, 1993, S. 396–400

[RDCGV05] ROSSI, F.; DELANNAY, N.; CONAN-GUEZA, B.; VERLEYSEN, M.: Representation of functional data in neural networks. In: *Neurocomputing* 64 (2005), S. 183–210

[RMS92] RITTER, Helge; MARTINETZ, Thomas; SCHULTEN, Klaus: *Neural Computation and Self-Organizing Maps: An Introduction*. Reading, MA : Addison-Wesley, 1992

[RS06] RAMSAY, J.O.; SILVERMAN, B.W.: *Functional Data Analysis*. 2nd. New York : Springer Science+Media, 2006

[Sil96] SILVERMAN, B.W.: Smoothed functional principal components analysis by the choice of norm. In: *The Annals of Statistics* 24 (1996), Nr. 1, S. 1–24

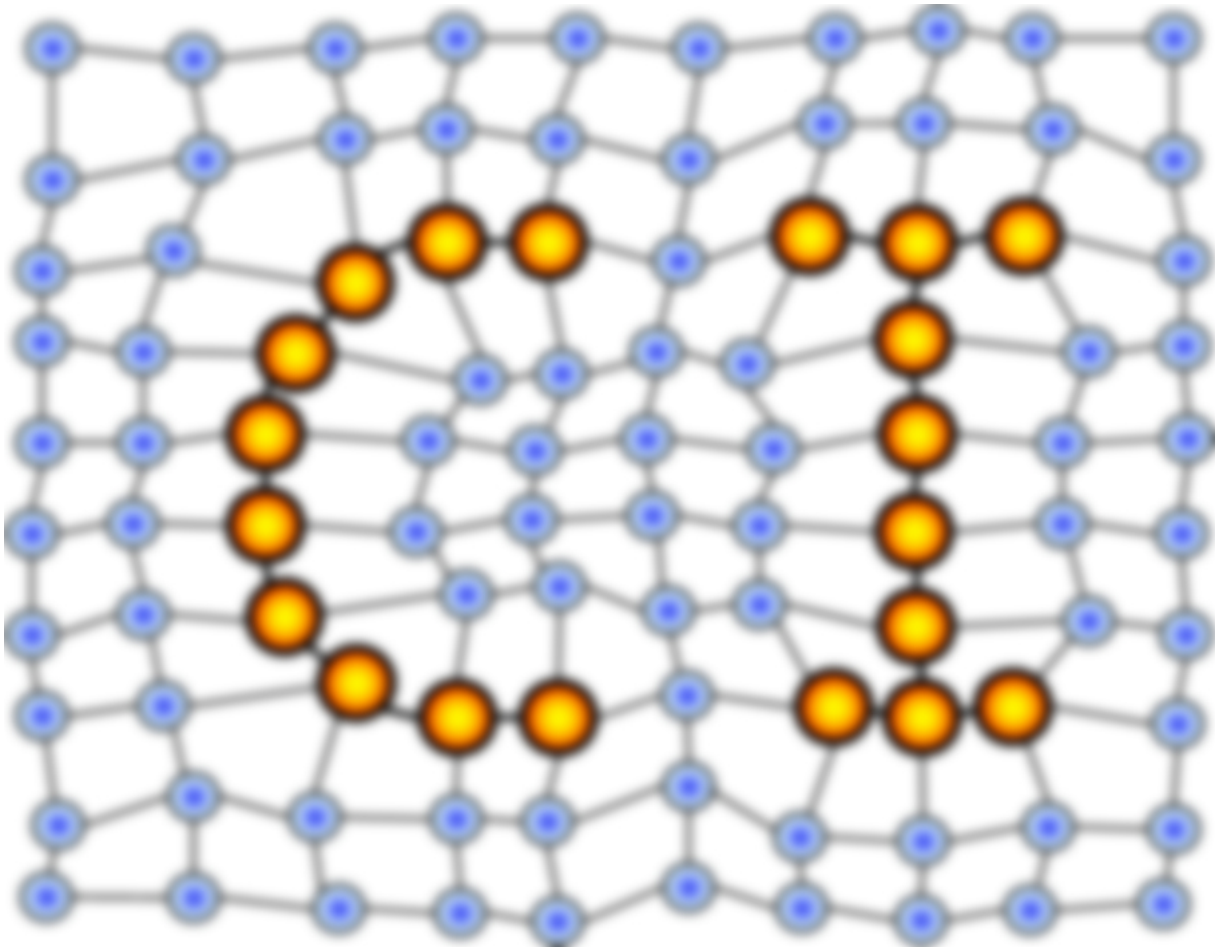
[Tri89] TRIEBEL, H.: *Analysis und mathematische Physik*. 3rd, revised. Leipzig : BSB B.G. Teubner Verlagsgesellschaft, 1989

[VHS+06] VILLMANN, T.; HAMMER, B.; SCHLEIF, F.-M.; GEWENIGER, T.; HERRMANN, W.: Fuzzy Classification by Fuzzy Labeled Neural Gas. In: *Neural Networks* 19 (2006), S. 772–779

[VSMH07] VILLMANN, T.; SCHLEIF, F.-M.; MERÉNYI, E.; HAMMER, B.: Fuzzy labeled self-organizing maps for classification of spectra. In: SANDOVAL, F. (Hrsg.); PRIETO, A. (Hrsg.); CABESTANY, J. (Hrsg.); GRANA, M. (Hrsg.): *Computational and Ambient Intelligence – Proceedings of the 9th Work-conference on Artificial Neural Networks (IWANN), San Sebastian (Spain)*. Berlin : Springer, 2007 (LNCS 4507), S. 556–563

MACHINE LEARNING REPORTS

Report 03/2007



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

PD. Dr. rer. nat. Thomas Villmann & Dr. rer. nat. Frank-Michael Schleif
Medical Department, University of Leipzig
Karl-Tauchnitz 25, D-04107 Leipzig, Germany •
http://www.uni-leipzig.de/compint

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.