Kernel Principal Component Analysis

Laura Pavone

Università degli Studi di Milano

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

Abstract

A new method for performing a nonlinear form of principal component analysis is proposed. By the use of integral operator kernel functions, one can efficiently compute principal components in high-dimensional feature spaces, related to input space by some nonlinear map. First we study some kernels properties and theorems and then we give the derivation of the method and also present experimental results on polynomial feature extraction for pattern recognition.

1 Introduction

Principal component analysis (PCA) is a powerful technique for extracting structure from possibly high-dimensional data sets. It is readily performed by solving an eigenvalue problem. PCA is an orthogonal transformation of the coordinate system in which we describe our data. The new coordinate values by which we represent the data are called principal components. It is often the case that a small number of principal components is sufficient to account for most of the structure in the data. We are interested not in principal components in input space but in principal components of variables, or features, which are nonlinearly related to the input variables. Among these there are variables obtained by taking arbitrary higher-order correlations between input variables. In the case of image analysis, this amounts to finding principal components in the space of products of input pixels. To this end, we are computing dot products in feature space by means of kernel functions in input space. So we need some technical results about kernels of RKHS (Reproducing Kernel Hilbert Spaces).

2 Kernels: something useful

Definition 2.1.
- An inner product on an $\mathbb{R}$-vector space H is a mapping $(f,g) \mapsto \langle f,g\rangle_H$ from H^2 to $\mathbb{R}$ that is bilinear, symmetric and such that $\langle f,f\rangle_H > 0 \ \forall f \in H\backslash\{0\}$.
- A vector space endowed with an inner product is called pre-Hilbert. It is endowed with a norm defined as $\|f\|_H = \langle f,f\rangle_H^{\frac{1}{2}}$.
- A Hilbert space is a pre-Hilbert space complete for the norm $\|\cdot\|_H$. That is, any Cauchy sequence in H converges in H.

Definition 2.2. A positive definite (p.d.) kernel on a set X is a function $K : X \times X \to \mathbb{R}$ that is symmetric:

$$\forall (x,x') \in X^2,\ K(x,x') = K(x',x),$$

and which satisfies, for all $N \in \mathbb{N}, (x_1,x_2,...,x_N) \in X^N$ and $(a_1,a_2,...,a_N) \in R^N$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j K(x_i,x_j) \geq 0$$

Remark 2.3. Equivalently, a kernel K is p.d. if and only if, for any $N \in \mathbb{N}$ and any set of points $(x_1,x_2,...,x_N) \in X^N$, the matrix K such that $[K]_{ij} := K(x_i,x_j)$ is positive semidefinite.

Kernel methods are algorithms that take such matrices as input.

Lemma 2.1. *Let $X = \mathbb{R}^d$. The function $K : X^2 \to \mathbb{R}$ defined by:*

$$\forall (x,x') \in X^2,\ K(x,x') = \langle x,x'\rangle_{\mathbb{R}^d} \tag{1}$$

is p.d. (it is often called the linear kernel).

A more ambitious p.d. kernel is the following:

Lemma 2.2. *Let X be any set and $\Phi : X \to \mathbb{R}^d$. Then the function $K : X^2 \to \mathbb{R}$ defined as follows is p.d.:*

$$\forall (x,x') \in X^2,\ K(x,x') = \langle \Phi(x),\Phi(x')\rangle_{\mathbb{R}^d} \tag{2}$$

But we are mostly interested in the converse statement.

2.1 Aronszajn's theorem

Theorem 2.3 (Aronszajn, 1950). *K is a p.d. kernel on the set X if and only if there exist a Hilbert space H and a mapping $\Phi : X \to H$ such that*

$$\forall x,x' \in X,\ K(x,x') = \langle \Phi(x),\Phi(x')\rangle_H \tag{3}$$

The proof is split up into 2 parts: the finite and the infinite case. We now give the proof of the first one. For the infinite case we need some other informations and results (in the next paragraph).

Proof. **Finite case**
Assume $X = \{x_1,x_2,...,x_N\}$ is finite of size N.
Any p.d. kernel $K : X \times X \to \mathbb{R}$ is entirely defined by the $N \times N$ symmetric positive semidefinite matrix $[K]_{ij} := K(x_i,x_j)$. It can therefore be diagonalized on an orthonormal basis of eigenvectors $(u_1,u_2,...,u_N)$, with non-negative eigenvalues $0 \leq \lambda_1 \leq ... \leq \lambda_N$, i.e.

$$K(x_i,x_j) = \left[\sum_{l=1}^{N} \lambda_l u_l u_l^T\right]_{ij} = \sum_{l=1}^{N} \lambda_l [u_l]_i [u_l]_j = \langle \Phi(x_i),\Phi(x_j)\rangle_{\mathbb{R}^N}$$

with

$$\Phi(x_i) = (\sqrt{\lambda_1}[u_1]_i,...,\sqrt{\lambda_N}[u_N]_i)^T.$$

□

For the infinite case, we have to introduce some tools.
Among the Hilbert spaces H mentioned in Aronszjan's theorem, one of them, called RKHS, is of interest to us.

Definition 2.4. Let X be a set and $H \subset \mathbb{R}^X$ be a class of functions forming a (real) Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$. The function $K : X^2 \to \mathbb{R}$ is called a reproducing kernel (r.k.) of H if

- H contains all functions of the form: $\forall x \in X$, $K_x : t \mapsto K(x,t)$;
- For every $x \in X$ and $f \in H$ the reproducing property holds: $f(x) = \langle f, K_x \rangle_H$.

If a r.k. exists, then H is called a reproducing kernel Hilbert space (RKHS).

The principle of RKHS gives us a simple recipe to do machine learning:

1. Map data x in X to a high-dimensional Hilbert space H (the RKHS) through a kernel mapping $\Phi : X \to H$, with $\Phi(x) = K_x$.
2. In H, consider simple linear models $f(x) = \langle f, \Phi(x) \rangle_H$;
3. If $X = \mathbb{R}^p$, a linear function in $\Phi(x)$ may be nonlinear in x.

Theorem 2.4. *A Hilbert space of functions $H \subset \mathbb{R}^X$ is a RKHS if and only if for any $x \in X$, the mapping $f \mapsto f(x)$ (from H to $\mathbb{R}$) is continuous.*

Corollary 2.4.1. *Convergence in a RKHS implies pointwise convergence on any point, i.e., if $(f_n)_{n\in\mathbb{N}}$ converges to $f \in H$, then $(f_n(x))_{n\in\mathbb{N}}$ converges to $f(x)$ for any $x \in X$.*

Theorem 2.5. *If H is a RKHS, then it has a unique r.k.*
Conversely, a function K can be the r.k. of at most one RKHS.

As a consequence we can talk of "the" kernel of a RKHS, or "the" RKHS of a kernel.

Proof. • **If a r.k. exists then it is unique.**
Infact let K and K' be two r.k. of a RKHS H. Then for any $x \in X$:

$$\begin{aligned} \|K_x - K'_x\|_H^2 &= \langle K_x - K'_x, K_x - K'_x \rangle_H \\ &= \langle K_x - K'_x, K_x \rangle_H - \langle K_x - K'_x, K'_x \rangle_H \\ &= K_x(x) - K'_x(x) - K_x(x) + K'_x(x) = 0 \end{aligned}$$

H being a Hilbert space, only the zero function has a norm equal to 0. This shows that $K_x = K'_x$ as functions, i.e., $K_x(y) = K'_x(y)$ for any $y \in X$ or, equivalently, $K(x,y) = K'(x,y)$. In other words, $K = K'$.

- **The RKHS of a r.k. K is unique.**
To prove the converse, first consider a RKHS H_1 with r.k. K. By definition of the r.k., we know that all the functions K_x for $x \in X$ are in H_1, therefore their linear span

$$H_0 = \left\{ \sum_{i=1}^{n} \alpha_i K_{x_i} : n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{R}, x_1, ..., x_n \in X \right\}$$

is a subspace of H_1. Now we observe that if $f \in H_1$ is orthogonal to H_0, then in particular it is orthogonal to K_x for any x which implies $f(x) = \langle f, K_x \rangle_{H_1} = 0$, i.e., $f = 0$. In other words, H_0 is dense in H_1.
Moreover the H_1 norm for functions in H_0 only depends on the r.k. K, because it is given for a function $f = \sum_{i=1}^{n} \alpha_i K_{x_i} \in H_0$ by

$$\|f\|_{H_1}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle K_{x_i}, K_{x_j} \rangle_{H_1} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j). \tag{4}$$

Suppose now that H_2 is also a RKHS that admits K as r.k. Then by the same argument, the space H_0 is dense in H_2, and the H_2 norm in H_0 is given by (4). In particular, for any $f \in H_0, \|f\|_{H_1} = \|f\|_{H_2}$.
Now let $f \in H_1$. By density of H_0 in H_1, there is a sequence $(f_n)_n$ in H_0 such that $\|f_n - f\|_{H_1} \to 0$. The converging sequence $(f_n)_n$ is in particular a Cauchy sequence for the H_1 norm, and since this norm coincides with the H_2 norm on H_0, $(f_n)_n$ is also a Cauchy sequence for the H_2 norm and converges in H_2 to a function $g \in H_2$. By the previous Corollary applied to both H_1 and H_2, we see that, for any $x \in X$, $\lim_{n\to+\infty} f_n(x) = f(x) = g(x)$. In other words, $f = g$ and therefore $f \in H_2$. By the arbitrariness of f this shows that $H_1 \subset H_2$ and, by symmetry of the argument, that $H_1 = H_2$. We now need to check that the norms in H_1 and H_2 coincide, which results from:

$$\|f\|_{H_1} = \lim_{n\to+\infty} \|f_n\|_{H_1} = \lim_{n\to+\infty} \|f_n\|_{H_2} = \|f\|_{H_2}.$$

□

Theorem 2.6. *A function $K : X \times X \to \mathbb{R}$ is p.d. if and only if it is a r.k.*

Proof. $\Leftarrow$) A r.k. is symmetric thanks to the reproducing property. Infact for any $(x, y) \in X^2$:

$$K(x, y) = \langle K_x, K_y\rangle_H = \langle K_y, K_x\rangle_H = K(y, x).$$

It is also p.d. because for any $N \in \mathbb{N}$, $(x_1, x_2, ..., x_N) \in X^N$ and $(a_1, a_2, ..., a_N) \in \mathbb{R}^N$:

$$\sum_{i,j=1}^{N} a_i a_j K(x_i, x_j) = \sum_{i,j=1}^{N} a_i a_j \langle K_{x_i}, K_{x_j}\rangle_H = \|\sum_{i=1}^{N} a_i K_{x_i}\|_H^2 \geq 0.$$

$\Rightarrow$) Conversely, let H_0 be the vector subspace of $\mathbb{R}^X$ spanned by the functions $\{K_x\}_{x\in X}$. For any $f, g \in H_0$, given by:

$$f = \sum_{i=1}^{m} a_i K_{x_i},\ g = \sum_{j=1}^{n} b_j K_{y_j}$$

let:

$$\langle f, g\rangle_{H_0} := \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j K(x_i, y_j) = \sum_{i=1}^{m} a_i g(x_i) = \sum_{j=1}^{n} b_j f(y_j).$$

This shows that $\langle f, g\rangle_{H_0}$ does not depend on the expansion of f and g and also shows that $\langle\cdot,\cdot\rangle_{H_0}$ is a symmetric bilinear form and $\forall x \in X$ and $f \in H_0$, $\langle f, K_x\rangle_{H_0} = f(x)$.
K is assumed to be p.d., therefore:

$$\|f\|_{H_0}^2 = \sum_{i,j=1}^{m} a_i a_j K(x_i, x_j) \geq 0.$$

By Cauchy-Schwarz, $\forall x \in X$:

$$\mid f(x) \mid = \langle f, K_X\rangle_{H_0} \leq \|f\|_{H_0} K(x, x)^{\frac{1}{2}}$$

therefore $\|f\|_{H_0} = 0 \Rightarrow f = 0$. H_0 is therefore a pre-Hilbert space endowed with the inner product $\langle ., .\rangle_{H_0}$. For any Cauchy sequence $(f_n)_{n\geq 0}$ in $(H_0, \langle ., .\rangle_{H_0})$, we note that:

$$\forall (x, m, n) \in X \times \mathbb{N}^2,\ \mid f_m(x) - f_n(x) \mid \leq \|f_m - f_n\|_{H_0} K(x, x)^{\frac{1}{2}}$$

Therefore for any x the sequence $(f_n(x))_{n\geq 0}$ is Cauchy in $\mathbb{R}$ and has therefore a limit. If we add to H_0 the functions defined as the pointwise limits of Cauchy sequences, the space becomes complete after some technical effort (completion) and is therefore a Hilbert space, with K as r.k. □

Now the proof of Aronszajn's theorem with X infinite set is an easy application of the previous results.

Proof. **Infinite case**
If K is p.d. over a set X then it is the r.k. of a Hilbert space $H \subset \mathbb{R}^X$. Let the mapping $\Phi : X \to H$ defined by: $\forall x \in X, \Phi(x) = K_x$. By the reproducing property:

$$\forall (x, y) \in X^2, \ \langle \Phi(x), \Phi(y) \rangle_H = \langle K_x, K_y \rangle_H = K(x, y).$$

□

When X is a compact set and K is continuous, another important result holds, the Mercer's theorem (we will use it in kernel PCA).

2.2 Mercer's Theorem

Definition 2.5. A kernel K on a set X is called a Mercer kernel if:

- X is a compact metric space;
- K: $X \times X \to \mathbb{R}$ is a continuous p.d. kernel (w.r.t. the Borel topology).

Definition 2.6. Let H be a Hilbert space

- a linear operator is a continuos linear mapping from H to itself;
- a linear operator L is called compact if, for any bounded sequence $\{f_n\}_n$, the sequence $\{Lf_n\}_n$ has a subsequence that converges;
- L is called self-adjoint if, for any $f, g \in H$:

$$\langle f, Lg \rangle = \langle Lf, g \rangle$$

- L is called positive if it is self-adjoint and for any $f \in H$:

$$\langle f, Lf \rangle \geq 0.$$

Now let ν be any Borel measure on X and $L^2_\nu(X)$ the Hilbert space of (equivalence classes of) square integrable functions on X. For any function $K : X^2 \to \mathbb{R}$ let the transform:

$$\forall f \in L^2_\nu(X), \ (L_K f)(x) = \int K(x, t) f(t) d\nu(t). \tag{5}$$

Lemma 2.7. *If K is a Mercer kernel, then L_K is a compact and bounded linear operator over $L^2_\nu(X)$, self-adjoint and positive.*

Proof. The proof is divided into 5 parts:

1. **L_K is a mapping from $L^2_\nu(X)$ to $L^2_\nu(X)$.**
   $\forall f \in L^2_\nu(X)$ and $\forall (x_1, x_2) \in X^2$:

$$\begin{aligned} \mid (L_K f)(x_1) - (L_K f)(x_2) \mid &= \left| \int (K(x_1, t) - K(x_2, t)) f(t) d\nu(t) \right| \\ &= \langle K_{x_1} - K_{x_2}, f \rangle \leq \| K_{x_1} - K_{x_2} \|_{L^2_\nu(X)} \| f \|_{L^2_\nu(X)} \\ &\leq \sqrt{\nu(X)} \max_{t \in X} \mid K(x_1, t) - K(x_2, t) \mid \| f \|_{L^2_\nu(X)}. \end{aligned}$$

   where the first inequality holds thanks to Cauchy-Schwarz.
   K being continuous and X compact, K is uniformly continuous, therefore $L_K f$ is continuous. In particular, $L_K f \in L^2_\nu(X)$.

2. L_K **is linear and continuous (that is equivalent to bounded).**

   Linearity is obvious. Instead for continuity we notice that for all $f \in L^2_\nu(X)$ and $x \in X$:

$$\begin{aligned} \mid (L_K f)(x) \mid &= \left| \int K(x,t) f(t) d\nu(t) \right| \\ &\leq \sqrt{\nu(X)} \max_{t \in X} \mid K(x,t) \mid \|f\|_{L^2_\nu(X)} \\ &\leq \sqrt{\nu(X)} C_K \|f\|_{L^2_\nu(X)}. \end{aligned}$$

   where $C_K = \max_{x,t \in X} \mid K(x,t) \mid < +\infty$. Therefore:

$$\|L_K f\|_{L^2_\nu(X)} = \left(\int (L_K f)(t)^2 d\nu(t) \right)^{\frac{1}{2}} \leq \nu(X) C_K \|f\|_{L^2_\nu(X)}.$$

3. L_K **is compact.**
   In order to prove the compactness of L_K recall a definition and a criterion (Ascoli Theorem).

   Definition 2.7. Let $C(X)$ denote the set of continuous functions on X endowed with infinite norm $\|f\|_\infty = \max_{x \in X} \mid f(x) \mid$.
   A set of functions $G \subset C(X)$ is called equicontinuous if:

$$\forall \epsilon > 0, \exists \delta > 0, \forall (x,y) \in X^2 \text{ s.t. } \|x - y\| < \delta \Rightarrow \forall g \in G, \mid g(x) - g(y) \mid < \epsilon. \qquad (6)$$

   Theorem 2.8 (Ascoli). *A set $H \subset C(X)$ is relatively compact (i.e. its closure is compact) if and only if it is uniformly bounded and equicontinuous.*

   We now complete the proof of compactness.
   So let $(f_n)_{n \geq 0}$ be a bounded sequence of $L^2_\nu(X)$, i.e. $(\|f_n\|_{L^2_\nu(X)} \leq M)$. The sequence $(L_K f_n)_{n \geq 0}$ is a sequence of continuous functions, uniformly bounded because:

$$\|L_K f_n\|_\infty \leq \sqrt{\nu(X)} C_K \|f_n\|_{L^2_\nu(X)} \leq \sqrt{\nu(X)} C_K M.$$

   It is equicontinuous because (recall K uniformly continuous):

$$\mid L_K f_n(x_1) - L_K f_n(x_2) \mid \leq \sqrt{\nu(X)} \max_{t \in X} \mid K(x_1,t) - K(x_2,t) \mid M.$$

   By Ascoli theorem, we can extract a sequence uniformly convergent in $C(X)$, and therefore in $L^2_\nu(X)$ since X is compact.

4. L_K **is self-adjoint.**
   K being symmetric, for all $f, g \in L$

$$\langle f, Lg \rangle_{L^2_\nu(X)} = \int f(x)(Lg)(x) d\nu(x) = \int \int f(x) g(t) K(x,t) d\nu(x) d\nu(t) = \langle Lf, g \rangle_{L^2_\nu(X)}$$

   thanks to Fubini.

5. L_K **is positive.**
   It is possible to approximate the integral by finite sums:

$$\langle f, Lf \rangle_{L^2_\nu(X)} = \int \int f(x) f(t) K(x,t) d\nu(x) d\nu(t) = \lim_{k \to +\infty} \frac{\nu(X)}{k^2} \sum_{i,j=1}^{k} K(x_i, x_j) f(x_i) f(x_j) \geq 0$$

   because K is positive definite.

□

In order to give the proof of the main result of this section, we need to recall the following:

Theorem 2.9. *Let L be a compact, self-adjoint, linear operator on a Hilbert space H. Then there exists in H a complete orthonormal system $(\psi_1, \psi_2, ...)$ of eigenvectors of L, with real eigenvalues $(\lambda_1, \lambda_2, ...)$ which are non-negative if L is positive.*

Thanks to the previous results, this theorem can be applied to L_K. In that case the eigenfunctions ψ_k associated to the eigenvalues $\lambda_k \neq 0$ can be considered as continuous functions, because:

$$\psi_k = \frac{1}{\lambda_k} L_K \psi_k. \tag{7}$$

We are finally ready to present and give the proof of Mercer's theorem:

Theorem 2.10 (**Mercer**). *Let X be a compact metric space, ν a nondegenerate Borel measure on X (i.e. $\nu(U) > 0$ for any nonempty open set $U \subset X$) , and K a continuous p.d. kernel. Let $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ denote the nonnegative eigenvalues of L_K and $(\psi_1, \psi_2, ...,)$ the corresponding eigenfunctions. Then all functions ψ_k are continuous, and for any $x, t \in X$:*

$$K(x,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x) \psi_k(t) \tag{8}$$

where the convergence is absolute for each $x, t \in X$, and uniform on $X \times X$.

Proof. For the sake of clarity the proof is split up into 5 parts.

1. $\forall k \geq 1$ **such that** $\lambda_k > 0$**,** $\psi_k \in H$ **(RKHS of** K**).**
   If $\lambda_k > 0$, we have

   $$\forall x \in X, \psi_k(x) = \frac{1}{\lambda_k} L_k \psi_k(x) = \frac{1}{\lambda_k} \int K(x,t) \psi_k(t) d\nu(t) = \lim_{n \to +\infty} \frac{\nu(X)}{\lambda_k n} \sum_{i=1}^{n} K(x, t_i) \psi_k(t_i)$$

   for a set $t_1, t_2, ...$ conveniently chosen. Besides, set $h_n := \frac{\nu(X)}{\lambda_k n} \sum_{i=1}^{n} K(\cdot, t_i) \psi_k(t_i) \in H$ for any $n \in N$ and, for any $n, m \in N$,

   $$\langle h_n, h_m \rangle_H = \frac{\nu(X)^2}{\lambda_k^2 nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \psi_k(t_i) \psi_k(t_j) K(t_i, t_j).$$

   Therefore,

   $$\lim_{n,m \to +\infty} \langle h_n, h_m \rangle_H = \frac{1}{\lambda_k^2} \int \int K(t,t') \psi_k(t) \psi_k(t') d\nu(t) d\nu(t') =: R,$$

   and

   $$\lim_{n,m \to +\infty} \|h_n - h_m\|^2 = R + R - 2R = 0$$

   $(h_n)_n$ is therefore a Cauchy sequence in H, which converges to a function $h \in H$. In particular, for any $x \in X$,

   $$h(x) = \lim_{n \to +\infty} h_n(x) = \psi_k(x),$$

   and finally $\psi_k = h \Rightarrow \psi_k \in H$.

2. $\{\sqrt{\lambda_k}\psi_k : \lambda_k > 0\}$ **is an orthonormal system (ONS) of** H**.**
Let $i, j \geq 1$ such that $\lambda_i, \lambda_j > 0$. Then $\sqrt{\lambda_i}\psi_i, \sqrt{\lambda_j}\psi_j \in H$ and

$$\begin{aligned}\langle\sqrt{\lambda_i}\psi_i, \sqrt{\lambda_j}\psi_j\rangle_H &= \langle\frac{1}{\sqrt{\lambda_i}}\int K_t\psi_i(t)d\nu(t), \sqrt{\lambda_j}\psi_j\rangle_H \\ &= \sqrt{\frac{\lambda_j}{\lambda_i}}\int \langle K_t, \psi_j\rangle_H\psi_i(t)d\nu(t) = \sqrt{\frac{\lambda_j}{\lambda_i}}\int \langle K_t, \psi_j\rangle_H\psi_j(t)\psi_i(t)d\nu(t) \\ &= \sqrt{\frac{\lambda_j}{\lambda_i}}\langle\psi_i, \psi_j\rangle_{L^2_\nu(X)} = \delta_{ij}.\end{aligned}$$

3. **For any** $x \in X$ $\sum_{k:\lambda_k>0}\lambda_k\psi_k(x)^2 \leq C_K$**.**
For any $x \in X$, $K_x \in H$ and $\|K_x\|^2_H = K(x, x) \leq C_K$.
Therefore, since $\{\sqrt{\lambda_k}\psi_k : \lambda_k > 0\}$ is an ONS of H:

$$C_K \geq \|K_x\|^2_H \geq \sum_{k:\lambda_k>0}\langle K_x, \sqrt{\lambda_k}\psi_k\rangle^2_H = \sum_{k:\lambda_k>0}\lambda_k\psi_k(x)^2.$$

4. **For any** $x \in X$**, the series of functions** $t \mapsto \sum_i \lambda_i\psi_i(x)\psi_i(t)$ **converges uniformly to a continuous function** g_x**.**
By Cauchy- Schwarz, for any fixed $x \in X$ and for any $t \in X$ (restricting the sum to the indices $i \geq 1$ such that $\lambda_i > 0$):

$$\begin{aligned}\left|\sum_{i=m}^{m+l}\lambda_i\psi_i(x)\psi_i(t)\right| &\leq \sum_{i=m}^{m+l}|\lambda_i\psi_i(x)\psi_i(t)| \\ &\leq \left(\sum_{i=m}^{m+l}\lambda_i\psi_i(x)^2\right)^{\frac{1}{2}}\left(\sum_{i=m}^{m+l}\lambda_i\psi_i(t)^2\right)^{\frac{1}{2}} \\ &\leq C_K\left(\sum_{i=m}^{m+l}\lambda_i\psi_i(t)^2\right)^{\frac{1}{2}},\end{aligned}$$

which tends to 0 uniformly in $t \in X$. Therefore $\sum_i \lambda_i\psi_i(x)\psi_i(t)$ converges uniformly in t for fixed x. Thus the series of the function $t \mapsto \sum_i \lambda_i\psi_i(x)\psi_i(t)$ is continuous and convergences uniformly to a continuous function g_x (because it's an uniform limit of continuous objects). The inequalities above also give us the absolute convergence.

5. $K_x = g_x \in L^2_\nu$**.**
On the other hand, we can expand K_x over the ONB $\{\psi_k, k \geq 1\}$ of $L^2_\nu(X)$:

$$K_x = \sum_{k\geq 1}\langle K_x, \psi_k\rangle_{L^2_\nu(X)}\psi_k = \sum_{k\geq 1}(L\psi_k)(x)\psi_k = \sum_{k\geq 1}\lambda_k\psi_k(x)\psi_k = \sum_{k\geq 1:\lambda_k>0}\lambda_k\psi_k(x)\psi_k,$$

therefore $K_x = g_x$ in L^2_ν, i.e. $\|K_x - g_x\|_{L^2_\nu} = 0$.
Since ν in nondegenerate, and both K_x and g_x are continuous, this implies

$$\forall t \in X,\ K_x(t) = g_x(t) = \sum_i \lambda_i\psi_i(x)\psi_i(t).$$

Since step (4) can be repeated for any fixed t and any x, we obtain:

$$\left|\sum_{i=m}^{m+l}\lambda_i\psi_i(x)\psi_i(t)\right| \leq C^2_K$$

so the convergence of the series (8) is uniform and absolute in $X \times X$.

□

Remark 2.8. The eigensystem (λ_k and ψ_k) depends on the choice of the measure ν: different measures lead to different feature spaces for a given kernel and a given space X.
Now let l^2 denote the Hilbert space of real-value sequences $u = (u_k)_{k\in\mathbb{N}}$ such that $\sum_{k\in\mathbb{N}} u_k^2 < +\infty$, endowed with the inner product $\langle u, v\rangle = \sum_{k\in\mathbb{N}} u_k v_k$.

Finally we obtain:

Theorem 2.11. *Let $L_K \in L^2_\nu(X)$, $(\lambda_1, \lambda_2, ...)$ and $(\psi_1, \psi_2, ...)$ as in Mercer's theorem. Then it holds that for any $x, y \in X$:*

$$K(x, y) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x)\psi_k(y) = \langle \Phi(x), \Phi(y)\rangle_{l^2}$$

with $\Phi : X \to l^2$ defined by $\Phi(x) = (\sqrt{\lambda_k}\psi_k(x))_{k\in\mathbb{N}}$.

In the next section we will introduce PCA and see how kernels are involved in this setting.

3 PCA

3.1 A brief introduction to standard PCA

Let $S = \{x_1, ..., x_M\}$ be a set of vectors ($x_i \in \mathbb{R}^N$). PCA is a classical algorithm in multivariate statistics to define a set of orthogonal directions that capture the maximum variance. One of its most common applications is low-dimensional representation of high-dimensional points.
It consists of an orthogonal transformation of the coordinate system in which we describe our data. The new coordinate values by which we represent the data are called principal components. It is readily performed by solving an eigenvalue problem.
Assume that the data are centered, i.e. $\frac{1}{M}\sum_{i=1}^{M} x_i = 0$. The orthogonal projection onto a direction $w \in \mathbb{R}$ is the function $h_w : \mathbb{R}^N \to \mathbb{R}$ defined by $h_w(x) = x^T \frac{w}{\|w\|}$.

In order to be able to generalize Standard PCA algorithm to the nonlinear case, we formulate it in a way that uses exclusively dot products.
We now describe standard PCA algorithm. Given a set of centered observations x_k, PCA diagonalizes the covariance matrix

$$C = \frac{1}{M}\sum_{j=1}^{M} x_j x_j^T. \tag{9}$$

Remark 3.1. More precisely, the covariance matrix is defined as the expectation of xx^T; for convenience, we shall use the same term to refer to the estimate in equation (9) of the covariance matrix from a finite sample.

To do this, we have to solve the eigenvalue equation

$$\lambda v = Cv \tag{10}$$

for eigenvalues $\lambda \geq 0$ and $v \in R^N\backslash\{0\}$. As $Cv = \frac{1}{M}\sum_{j=1}^{M}(x_j \cdot v)x_j$, all solutions v with $\lambda \neq 0$ must lie in the span of $x_1, ..., x_M$; hence, equation (10) in that case is equivalent to

$$\lambda(x_k \cdot v) = (x_k \cdot Cv) \text{ for all } k = 1, ..., M. \tag{11}$$

In the next paragraph we will describe the same computation in an other dot product space F, which is related to the input space by a possibly nonlinear map,

$$\Phi : R^N \to F, \qquad x \mapsto X. \tag{12}$$

Note that F, which we will refer to as the feature space, could have an arbitrarily large, possibly infinite, dimensionality. Here and in the following, uppercase characters are used for elements of F, and lowercase characters denote elements of R^N.

3.2 PCA in Feature Spaces

Again, we assume that we are dealing with centered data, that is $\sum_{k=1}^{M} \Phi(x_k) = 0$. Using the covariance matrix in F,

$$\bar{C} = \frac{1}{M} \sum_{j=1}^{M} \Phi(x_j)\Phi(x_j)^T \tag{13}$$

(if F is infinite dimensional, we think $\Phi(x_j)\Phi(x_j)^T$ as the linear operator that maps $X \in F$ to $\Phi(xj)(\Phi(xj)^T \cdot X)$. We now have to find eigenvalues $\lambda \geq 0$ and eigenvectors $V \in F\backslash\{0\}$ satisfying

$$\lambda V = \bar{C}V. \tag{14}$$

All solutions V with $\lambda \neq 0$ must lie in the span of $\Phi(x_1), ..., \Phi(x_M)$. For us, this has two useful consequences. First, we may instead consider the set of equations,

$$\lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \bar{C}V) \text{ for all } k = 1, ..., M, \tag{15}$$

and, second, there exist coefficients α_i , $i = 1, ..., M$, such that,

$$V = \sum_{i=1}^{M} \alpha_i \Phi(x_i). \tag{16}$$

Combining equations (15) and (16), we get

$$\lambda \sum_{i=1}^{M} \alpha_i(\Phi(x_k) \cdot \Phi(x_i)) = \frac{1}{M} \sum_{i=1}^{M} \alpha_i(\Phi(x_k) \cdot \sum_{j=1}^{M}(\Phi(x_j))(\Phi(x_j) \cdot (\Phi(x_i)) \text{ for all } k = 1, ..., M,$$

Defining an $M \times M$ matrix K by

$$K_{ij} := (\Phi(x_i) \cdot (\Phi(x_j)), \tag{17}$$

this reads

$$M\lambda K\alpha = K^2\alpha, \tag{18}$$

where α denotes the column vector with entries $\alpha_1, ..., \alpha_M$. To find solutions of the last equation we can solve, instead,

$$M\lambda\alpha = K\alpha. \tag{19}$$

Since K is symmetric, it has an orthonormal basis of eigenvectors β^i with corresponding eigenvalues μ_i; thus, for all i, we have $K\beta^i = \mu_i\beta^i$, $i = 1, ..., M$. First suppose λ, α satisfy equation (18). Then expand α in K's eigenvector basis as $\alpha = \sum_{i=1}^{M} a_i\beta^i$. By substituing in (18), we obtain

$$M\lambda = \mu_i \text{ or } a_i = 0 \text{ or } \mu_i = 0 \tag{20}$$

Note that the above are not exclusive ors. We next assume that λ, α satisfy equation (19), to carry out a similar derivation. In that case, we find

$$M\lambda = \mu_i \text{ or } a_i = 0. \tag{21}$$

Comparing equations (20) and (21), we notice that all solutions of the latter satisfy the former. However, they do not give its full set of solutions: given a solution of equation (19), it can be always added multiples of eigenvectors of K with eigenvalue 0 and still satisfy equation (18), with the same eigenvalue.

This means that there exist solutions of equation (18) that belong to different eigenvalues yet are not orthogonal in the space of the α_k. It does not mean, however, that the eigenvectors of $\bar{C}$ in F are not orthogonal. Indeed, if α is an eigenvector of K with eigenvalue 0, then the corresponding vector $\sum_{i=1}^{M} \alpha_i\Phi(x_i)$ is orthogonal to all vectors in the span of the $\Phi(x_j)$ in F, since $(\Phi(x_j) \cdot \sum_{i=1}^{M} \alpha_i\Phi(x_i)) = (K\alpha)_j = 0$ for all $j = 1, ...M$, which means that $\sum_{i=1}^{M} \alpha_i\Phi(x_i) = 0$. Thus, the above difference between the solutions of equations (18) and (19) is irrelevant, since

we are interested in vectors in F rather than vectors in the space of the expansion coefficients of equation (16). Thus it is sufficient to diagonalize K to find all relevant solutions of equation (18).

Let $\lambda_1 \geq \lambda_2 \geq \lambda_M$ denote the eigenvalues of K and $\alpha_1, ..., \alpha_M$ the corresponding complete set of eigenvectors, with λ_p being the first nonzero eigenvalue (assuming $\Phi \neq 0$). We normalize $\alpha_p, ..., \alpha_M$ by requiring that the corresponding vectors in F be normalized, that is,

$$(V^k \cdot V^k) = 1 \text{ for all } k = p, ...M.$$

By virtue of equations (16) and (19), this translates into a normalization condition for $\alpha_p, ..., \alpha_M$:

$$1 = \sum_{i,j=1}^{M} \alpha_i^k \alpha_j^k (\Phi(x_i) \cdot \Phi(x_j)) = \sum_{i,j=1}^{M} \alpha_i^k \alpha_j^k K_{ij} = (\alpha^k \cdot K\alpha^k) = \lambda_k (\alpha^k \cdot \alpha^k).$$

For the purpose of principal component extraction, we need to compute projections onto the eigenvectors V^k in F, $k = p, ..., M$. Let x be a test point, with an image $\Phi(x)$ in F; then

$$(V^k \cdot \Phi(x)) = \sum_{i=1}^{M} \alpha_i^k (\Phi(x_i) \cdot \Phi(x)) \tag{22}$$

may be called its nonlinear principal components corresponding to Φ. Note that neither (17) or (22) requires the $\Phi(x_i)$ in explicit form- they are only needed in dot products. Therefore, we are able to use kernel functions for computing these dot products without actually performing the map Φ.

4 Kernel PCA

In F, we can thus assert that PCA is the orthogonal basis transformation with the following properties (assuming that the eigenvectors are sorted in descending order of the eigenvalue size):

- the first q ($q \in \{1, ..., M\}$) principal components, that is, projections on eigenvectors, carry more variance than any other q orthogonal directions;
- the principal components are uncorrelated;
- the first q principal components have maximal mutual information with respect to the inputs.

In order to compute dot products of the form $K(x, y) = (\Phi(x) \cdot \Phi(y))$, we use kernel representations, which allow us to compute the value of the dot product in F without having to carry out the map Φ. The general question that function K does correspond to a dot product in some space F has been discussed a lot.
In section **(2)** we tried to offer a solution, in particular we pointed out (and discussed) Aronszajn **(2.3)** and Mercer's **(2.10)** theorems and, finally, theorem **(2.11)** .

Infact we recall that it gives the conditions under which we can construct the mapping Φ from the Eigenfunction decomposition of K. We fix the space L^2 and obtain:

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y)$$

with positive coefficients λ_i, $(\psi_i \cdot \psi_j) = \delta_{ij}$ and $\Phi(x) := (\sqrt{\lambda_i}\psi_i(x))_{i \in \mathbb{N}}$ is a map into a space where K acts as the Euclidean dot product, i.e. $(\Phi(x) \cdot \Phi(y)) = K(x, y)$.

The application of equation $K(x, y) = (\Phi(x) \cdot \Phi(y))$ to our problem is straightforward. We simply substitute an a priori chosen kernel function $K(x, y)$ for all occurrences of $(\Phi(x) \cdot \Phi(y))$.

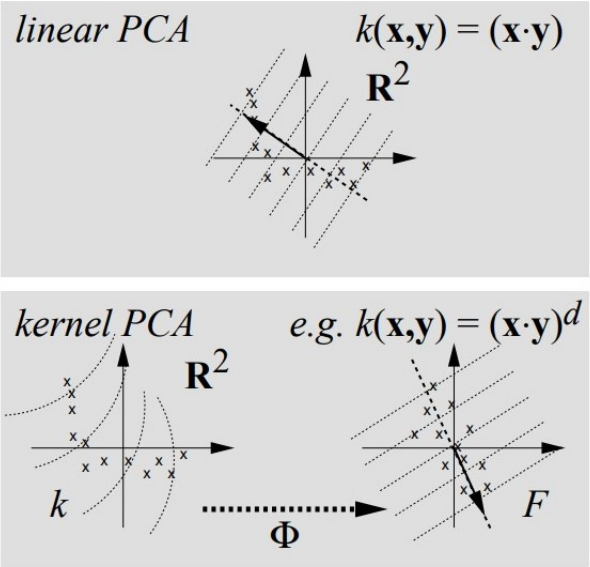

Figure 1: The basic idea of kernel PCA. In some high-dimensional feature space F (bottom right), we are performing linear PCA, just like a PCA in input space (top). Since F is nonlinearly related to input space (via Φ), the contour lines of constant projections onto the principal eigenvector (drawn as an arrow) become nonlinear in input space. Note that we cannot draw a preimage of the eigenvector in input space, because it may not even exist. Crucial to kernel PCA is the fact that there is no need to carry out the map into F. All necessary computations are carried out by the use of a kernel function K in input space (here: $\mathbb{R}^2$).

The choice of K then implicitly determines the mapping Φ and the feature space F.
A common example is the polynomial kernel:

$$K(x,y) = (x \cdot y)^d = \left(\sum_{j=1}^{N} x_j y_j\right)^d = \sum_{j_1,j_d=1}^{N} x_{j_1}...x_{j_d} y_{j_1}...y_{j_d} = (C_d(x) \cdot C_d(y)) \tag{23}$$

where C_d maps x to the vector $C_d(x)$ whose entries are all possible d-th degree ordered products of the entries of x.

4.1 The algorithm

To perform kernel-based PCA (see Figure 1), the following steps have to be carried out:

1. we compute the matrix with $K_{ij} = K(x_i, x_j)$;
2. we solve equation (17) by diagonalizing K and normalize the eigenvector expansion coefficients α_n by requiring $\lambda_n(\alpha^n \cdot \alpha^n) = 1$;
3. To extract the principal components (corresponding to the kernel K) of a test point x, we then compute projections onto the eigenvectors by $(V^n \cdot \Phi(x)) = \sum_{i=1}^{M} \alpha_i^n K(x_i, x)$.

4.2 Centering in High-Dimensional Space

In the previous sections we supposed centered observations. But in general they aren't so we shall drop this assumption. Given any Φ and any set of observations $x_1, ..., x_M$, the points

$$\tilde{\Phi}(x_i) = \Phi(x_i) - \frac{1}{M}\sum_{i=1}^{M} \Phi(x_i) \tag{24}$$

are centered. Thus the assumption of section **(4)** now hold and we go on defining covariance matrix through $\tilde{K}_{ij} = (\tilde{\Phi}(x_i) \cdot \tilde{\Phi}(x_j))$ in F. We arrive at the already known eigenvalue problem $\tilde{\lambda}\tilde{\alpha} = \tilde{K}\tilde{\alpha}$, with $\tilde{\alpha}$ being the expansion coefficients of an eigenvector (in F) in terms of the points in equation (24), $\tilde{V} = \sum_{i=1}^{M} \tilde{\alpha}_i \tilde{\phi}(x_i)$. But we do not have the centered data, so we cannot compute $\tilde{K}$ directly; however, we can express it in terms of its noncentered counterpart K.

Lemma 4.1. *Covariance matrix $\tilde{K}$ can be expressed in terms of Covariance matrix K.*

Proof. Some notations:

$$\tilde{K}_{ij} := (\tilde{\Phi}(x_i) \cdot \tilde{\Phi}(x_j)); \; K_{ij} := (\Phi(x_i) \cdot \Phi(x_j));$$
$$\text{matrices defined by } 1_{ij} := 1 \; \forall i,j \text{ and } (1_M)_{ij} := \frac{1}{M} \; \forall i,j.$$

Making $\tilde{K}_{ij} = (\tilde{\Phi}(x_i) \cdot \tilde{\Phi}(x_j))$ explicit:

$$\begin{aligned}
\tilde{K}_{ij} &= (\tilde{\Phi}(x_i) \cdot \tilde{\Phi}(x_j)) \\
&= (\Phi(x_i) - \frac{1}{M}\sum_{m=1}^{M} \Phi(x_m)) \cdot (\Phi(x_j) - \frac{1}{M}\sum_{n=1}^{M} \Phi(x_n)) \\
&= \Phi(x_i)^T\Phi(x_j) - \frac{1}{M}\sum_{m=1}^{M} \Phi(x_m)^T\Phi(x_j)) \\
&- \frac{1}{M}\sum_{n=1}^{M} \Phi(x_i)^T\Phi(x_n)) + \frac{1}{M^2}\sum_{m,n=1}^{M} \Phi(x_m)^T\Phi(x_n)) \\
&= K_{ij} - \frac{1}{M}\sum_{m=1}^{M} 1_{im}K_{mj} - \frac{1}{M}\sum_{n=1}^{M} K_{in}1_{nj} + \frac{1}{M^2}\sum_{m,n=1}^{M} 1_{im}K_{mn}1_{nj} \\
&= (K - 1_M K - K1_M + 1_M K 1_M)_{ij}
\end{aligned}$$

□

As before, the solutions $\tilde{\alpha}^k$ are normalized by normalizing the corresponding vectors $\tilde{V}^k$ in F which translates into $\tilde{\lambda}_k(\tilde{\alpha}^k \cdot \tilde{\alpha}^k) = 1$.
For feature extraction, we compute projections of centered Φ-images of test patterns t onto the eigenvectors of the covariance matrix of the centered points,

$$(\tilde{V}^k \cdot \tilde{\Phi}(t)) = \sum_{i=1} M\tilde{\alpha}_i^k(\tilde{\Phi}(x_i) \cdot \tilde{\Phi}(t)). \tag{25}$$

Consider a set of test points $t_1, ..., t_L$, and define two $L \times M$ matrices by $K_{ij}^{test} = (\Phi(t_i) \cdot \Phi(x_j))$ and $\tilde{K}_{ij}^{test} = (K^{test} - 1'_M K - K^{test}1_M + 1'_M K 1_M)_{ij}$ where $1'_M$ is the $L \times M$ matrix with all entries equal to $\frac{1}{M}$.

4.3 Computational complexity

A fifth-order polynomial kernel on a 256-dimensional input space yields a 1010-dimensional feature space. For two reasons kernel PCA can deal with this huge dimensionality. First, we do not need to look for eigenvectors in the full space F, but just in the subspace spanned by the images of our observations x_k in F. Second, we do not need to compute dot products explicitly between vectors in F because we are using kernel functions. If K is easy to compute, as for polynomial kernels, for example, the computational complexity is hardly changed by the fact that we need to evaluate kernel functions rather than just dot products.

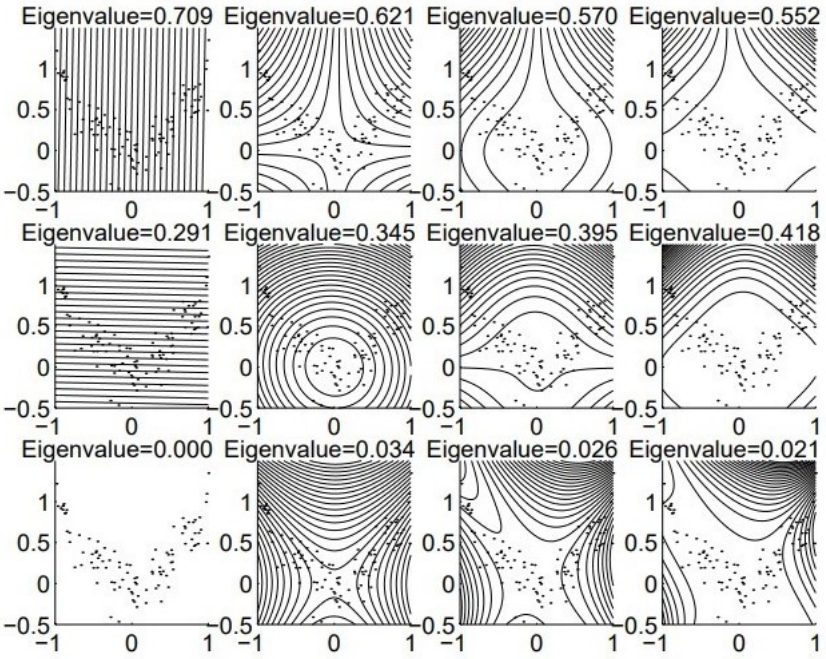

Figure 2: 2-d toy example with data generated in the following way: x-values have uniform distribution in $[-1, 1]$, y-values are generated from $y_i = x_i^2 + \xi$, where ξ is normal noise with standard deviation 0.2.

5 Experiments

To provide some insight into how PCA behaves, we show a set of experiments with an artificial two dimensional data set (Figure 2). Polynomial kernels of degree 1 through 4 are used. Besides, from top to bottom, the first 3 eigenvectors are shown (in order of decreasing Eigenvalue size). Linear PCA (on the left) leads to only two nonzero eigenvalues, as the input dimensionality is 2. In contrast, nonlinear PCA allows the extraction of further components. In the figure, note that nonlinear PCA produces contour lines (of constant feature value), which reflect the structure in the data better than in linear PCA. In all cases, the first principal component varies monotonically along the parabola underlying the data.
In the nonlinear cases, the second and the third components show behaviour that is similar for different polynomial degrees. The third component, which comes with small eigenvalues (rescaled to sum to 1), seems to pick up the variance caused by the noise, as can be nicely seen in the case of degree 2. Dropping this component would thus amount to noise reduction.

For an investigation of the utility of kernel PCA features for a realistic pattern recognition problem, Vapnik & Chervonenkis [5] and then Cortes & Vapnik [6] trained a separating hyperplane classifier on nonlinear features extracted from the US postal service (USPS) handwritten digits database by kernel PCA. This database contains 9300 examples of dimensionality 256; 2000 of them make up the test set. For computational reasons, they decided to use a subset of 3000 training examples for the dot product matrix.
Using polynomial kernels (23) of degrees $d = 1, .., 7$ and extracting the first 2^n ($n = 5, ...11$) principal components, they found the following results. In the case of linear PCA ($d = 1$) the best classification performance (8.6% error) is attained for 128 components. Extracting the same number of nonlinear components ($d = 2, ..., 7$) in all cases lead to superior performance (around 6% error). Moreover in the nonlinear case, the performance can be further improved by using a larger number of components (Figure 3).
Using $d > 2$ and 2048 components, they obtained around 4% error which coincides with the best result reported for standard nonlinear Support vector machines (Scholkopf, Burges & Vapnik,

# of components	Test Error Rate for degree						
	1	2	3	4	5	6	7
32	9.6	8.8	8.1	8.5	9.1	9.3	10.8
64	8.8	7.3	6.8	6.7	6.7	7.2	7.5
128	8.6	5.8	5.9	6.1	5.8	6.0	6.8
256	8.7	5.5	5.3	5.2	5.2	5.4	5.4
512	n.a.	4.9	4.6	4.4	5.1	4.6	4.9
1024	n.a.	4.9	4.3	4.4	4.6	4.8	4.6
2048	n.a.	4.9	4.2	4.1	4.0	4.3	4.4

Figure 3: Test error rate on the USPS handwritten digits database for linear Support Vector machines trained on nonlinear principal components extracted by PCA with polynomial kernel from degrees 1 through 7.

[8]). This result is much better than linear classifiers operating directly on the image data. We also believe that choosing a suitable kernel with respect to the problem could further improve the results.

6 Conclusion

Kernel PCA is a nonlinear generalization of PCA in the sense that if we use kernel $K(x, y) = (x \cdot y)$, we recover original PCA.
To get nonlinear forms of PCA, we simply choose a nonlinear kernel. Moreover, kernel PCA is a generalization of PCA in the respect that it is performing PCA in feature spaces of arbitrarly large (possibly infinite) dimension.
We now want to collect the advantages (and some disadvantages) of kernel PCA.
First of all kernel PCA can deal with huge dimensionality because we do not need to look for eigenvectors in the full space F, but just in the subspace spanned by the images of our observations x_k in F and we do not need to compute dot products explicitly between vectors in F.
Besides, in experiments, we found two advantages of nonlinear kernels. First, nonlinear principal components afforded better recognition rates than corresponding numbers of linear principal components; and, second, the performance for nonlinear components can be improved by using more components than is possible in the linear case.
Lastly, compared to other techniques for nonlinear feature extraction, kernel PCA doesn't require nonlinear optimization but just the solution of an Eigenvalue problem and different kernels led to fine classification performances. The general question of how to select the ideal kernel for a given task, however, is an open problem.
The main drawback of kernel PCA compared to linear PCA is that up to date we do not have a simple method for reconstructing patterns from their principal components.
Possible applications: Linear PCA is being used in numerous technical and scientific application. As some further examples not discussed in the present report so far, we mention noise reduction, density estimation and the analysis of natural image statistics. Kernel PCA can be applied to all domains where traditional PCA has been used for feature extraction before, with little extra computational effort.

Acknowledgement

To write this paper I used mainly the lecture notes "Machine Learning with Kernel Methods" [1] with regard to the Kernel theory but I also consult [2] and [3] for integral operators.
For Kernel PCA I mainly based my work on [4].

References

[1] J. Mairal, J.-P. Vert, *Machine Learning with Kernel Methods*, Lecture notes. Last version, 2020.

[2] H. Hochstadt, *Integral equations*, John Wiley & Sons Inc, 1989.

[3] R. Courant & D. Hilbert, *Methods of mathematical physics* (Vol. 1), Interscience publisher, Inc, New York, 1953.

[4] B. Schölkopf, A. J. Smola & K.-R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem.*, Technical Report 44, Max-Planck-Institut für biologische, 1996.

[5] V. Vapnik & A. Chervonenkis, *Theory of pattern recognition*, 1974.

[6] C. Cortes & V. Vapnik, Support vector networks. *Machine Learning*, 1995.

[7] B. Schölkopf, C. Burges & V. Vapnik, *Extracting support data for a given task*, 1995.