STATISTICS IN TRANSITION-new series, December 2012
Vol. 13, No. 3, pp. 559—668

A KERNEL VERSION OF FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

Tomasz Górecki[1], **Mirosław Krzyśko**[2]

ABSTRACT

In this paper a new construction of functional principal components (FPCA) is proposed, based on principal components for vector data. A kernel version of FPCA is also presented. The quality of the two described methods was tested on 20 different data sets.

Key words: PCA, FPCA, kernel version of FPCA.

1. Introduction

Advances in modern technology, including computing environments, have facilitated the collection and analysis of high-dimensional data, or data that consist of repeated measurements of the same subject. If the repeated measurements are taken densely over a period of time, say on an interval I, often by machine, they are typically termed functional or curve data, with one observed curve (or function) per subject. This is often the case even if the data are observed with experimental error, since the operation of smoothing data recorded at closely spaced time points can greatly reduce the effects of noise. In such cases we may regard the entire curve for the i th subject, represented by the graph of the function $X_i(t)$ say, as being observed in the continuum, even though in reality the recording times are discrete. The statistical analysis of a sample of n such graphs is commonly termed functional data analysis (see Ramsay and Dalzell, 1991). Functional data analysis whose main purpose is to provide tools for describing and modelling sets of curves is a topic of growing interest in the statistical community. The books by Ramsay and Silverman (2002, 2005) propose an interesting description of the available procedures dealing with functional observations. These functional approaches have been proved useful in various

[1] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland. E-mail: tomasz.gorecki@amu.edu.pl.

[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland. E-mail: mkrzysko@amu.edu.pl.

domains such as chemometrics, economics, climatology, biology and remote sensing. The statistician generally wants, as a first step, to represent as far as possible a set of random curves in a small space in order to get a description of the functional data that allows interpretation. Functional principal components analysis (FPCA) gives a small-dimension space which captures the main modes of variability of the data. The basic idea in functional principal components analysis is to find functions whose inner products with the data yield the maximum variation in the curves. The first principal component accounts for the most variation, the second principal component accounts for the largest variation orthogonal to the first principal component, and so on. In this way, much of the variation in the modern data can be captured using only a few principal components. A construction method for functional principal components is given in the monograph of Ramsay and Silverman (2005).

In this paper, we present a new way of constructing the functional principal components. This construction is described in Section 3. In Section 4 we present a kernel version of functional principal components analysis. In Section 5 we show how these two methods work on the example of 20 different real data sets. Section 6 contains results and conclusions.

2. Smoothing discrete data

Let y_{ij} denote the observed value of an investigated statistical property on the ith unit at the jth time point, where $j = 1,2,\dots,J_i$, $i = 1,2,\dots,N$. The observation time points t_{ij} of a given statistical property may vary from unit to unit, and the intervals between these points need not be uniform. Our data then consist of pairs (t_{ij}, y_{ij}), where $t_{ij} \in I, j = 1,2,\dots,J_i, i = 1,2,\dots,N$.

The discrete data (t_{ij}, y_{ij}) can be transformed into functional data (see Ramsay and Silverman (2005)):

$$\{x_i(t), i = 1,2,\dots,N, t \in I\}\ .$$

Because the data transformation process is carried out separately for each $i = 1,2,\dots,N$, our further considerations will relate to a single function $x(t), t \in I$.

One of the ways of smoothing discrete data $\{t_j, y_j,\}$, $j \in J$, $t_j \in I$, to a continuous function $x(t)$ on the interval I is to present that function as a linear combination N of orthonormal basis functions φ_k:

$$x(t\ ) = \sum_{k=0}^{N-1} c_k \varphi_k(t)\ , t \in I. \tag{2.1}$$

The coefficients c_k of this linear combination are selected by the least squares method, i.e. so as to minimize the function:

$$S(c_0, c_1, \dots, c_{N-1}) = \sum_{j=1}^{J}\left(y_j - \sum_{k=0}^{N-1} c_k\varphi_k(t_j)\right)^2.$$

In matrix notation, the function S takes the form:

$$S(\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})'(\boldsymbol{y} - \boldsymbol{\Phi c}),$$

where $\boldsymbol{c} = (c_0, c_1, \dots, c_{N-1})'$ and $\boldsymbol{\Phi}$ is a $J \times N$ matrix containing the values $\varphi_k(t_j)$. By differentiating $S(\boldsymbol{x})$ with respect to the vector $\boldsymbol{c}$ we obtain a system of normal equations in the form:

Hence, the estimate of vector $\boldsymbol{c}$ is equal to

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{y}. \tag{2.2}$$

3. Construction of functional principal components

Suppose we observe a sample of the process $X(t) \in L_2(I)$, where $L_2(I)$ is the Hilbert space of square integrable functions on an interval I, equipped with the scalar product $< u, v >= \int u(s)v(s)ds$.

Remark 1. All integrals are taken over the interval I.

Moreover, suppose that $EX(t) = 0$ and

$$\int E||X||^2 = E[< X, X >] = E\int X^2(s)ds < \infty.$$

A principal component as defined for finite-dimensional vectors, $X \in \mathbb{R}^k$, and for a stochastic process $X(t) \in L_2(I)$ is characterized as follows: Let $H = \mathbb{R}^k$ in the vector case, $H = L_2(I)$ in the functional case. Then, the first eigenvalue λ_1 and associated weight function or vector u_1 are defined as

$$\lambda_1 = \sup_{u\in H} Var(< u, X >) = Var(< u_1, X >), \tag{3.1}$$

subject to the constraint that

$$||u|| = 1. \tag{3.2}$$

The condition (3.2) is imposed to ensure the uniqueness (except for sign) of the principal component.

The kth eigenvalue λ_k and weight function or vector u_k, for $k > 1$, are defined as

$$\lambda_k = \sup_{u\in H} Var(< u, X >) = Var(< u_k, X >),$$

where u is subject to (3.2) and the kth principal component U_k is uncorrelated with the first $(k-1)$ principal components $U_i,\ i = 1, \dots, k-1$ where

$$U_k =< u_k, X >. \tag{3.3}$$

We shall call (λ_k, u_k) the kth principal configuration.

Regarding the case where $X(t)$ is a stochastic process, we will assume that the process $X(t)$ can be represented by a finite number of orthonormal basis functions. Let

$$X(t) = \sum_{k=0}^{N-1} c_k \varphi_k(t)\,, t \in I, \tag{3.4}$$

where $\{\varphi_k\}$ are the first N elements of an orthonormal basis of $L_2(I)$, and $\{c_k\}$ are random variables with zero means and finite variances. We adopt the notation

$$\varphi(t) = \big(\varphi_0(t), \varphi_1(t), \dots, \varphi_{N-1}(t)\big)',$$
$$c = (c_0, c_1, \dots, c_{N-1})', 0 < N-1 < \infty,$$

with $E(c) = 0$ and $Var(c) = \Sigma$. The process $X(t)$ can be written in vector form as

$$X(t) = c'\varphi(t), t \in I. \tag{3.5}$$

Theorem 1. *The kth principal configuration of random vector c, defined by (σ_k, u_k), is related to the kth principal configuration of stochastic process $X(t)$, $\big(\lambda_k, u_k(t)\big)$, as follows:*

$$u_k(t) = u_k'\varphi(t), \qquad \lambda_k = \sigma_k. \tag{3.6}$$

Proof.
Each function $u(t) = L_2(I)$ can be written as

$$u(t) = \boldsymbol{u}'\boldsymbol{\varphi}(t), \text{gdzie } \boldsymbol{u} \in \mathrm{R}^N.$$

Then

$$\langle u, X\rangle = \langle \boldsymbol{u}'\boldsymbol{\varphi}, \boldsymbol{c}'\boldsymbol{\varphi}\rangle = \boldsymbol{u}'\langle \boldsymbol{\varphi}, \boldsymbol{\varphi}'\rangle\boldsymbol{c} = \boldsymbol{u}'\boldsymbol{I}_N\boldsymbol{c} = \boldsymbol{u}'\boldsymbol{c},$$
$$E[\langle u, X\rangle] = \boldsymbol{u}'E(\boldsymbol{c}) = 0,$$
$$Var([\langle u, X\rangle] = \boldsymbol{u}^{\uparrow\prime}E\big(\boldsymbol{c}\boldsymbol{c}^{\uparrow\prime}\big)\boldsymbol{u} = \boldsymbol{u}^{\uparrow\prime}\boldsymbol{\Sigma}\boldsymbol{u}.$$

Consider the first principal component of the process $X(t)$:

$$\lambda_1 = \sup_{u \in L(I)} Var(\langle u, X\rangle) = Var(\langle u_1, X\rangle),$$

where $\langle u_1, u_1\rangle = 1$.

This is equivalent to stating that

$$\lambda_1 = \sup_{\boldsymbol{u}\in R^N} \boldsymbol{u}'\boldsymbol{\Sigma u} = \boldsymbol{u}_1'\boldsymbol{\Sigma u_1},$$

where $\boldsymbol{u}_1'\boldsymbol{u_1} = 1$.

This is the definition of the first principal component of the random vector $\boldsymbol{c}$.

On the other hand, if we begin with the first principal component of the random vector $\boldsymbol{c}$ as given by the principal configuration (σ_1, u_1), we will obtain the first principal component of the process $X(t)$ from the equation

$$\left(\lambda_1, u_1(t)\right) = \left(\sigma_1, \boldsymbol{u}_1\boldsymbol{\varphi}(t)\right).$$

Similarly we can extend this reasoning to the second principal component and so on.

Principal components analysis for a random process with finite basis expansion is therefore equivalent to multivariate principal component analysis.

Since Σ is ordinarily unknown, we use the estimator $\hat{\Sigma}$ based on N independent realizations of the random vector c:

$$\hat{C} = \begin{bmatrix} \hat{c}_{10} & \hat{c}_{11} & \dots & \hat{c}_{1,N-1} \\ \hat{c}_{20} & \hat{c}_{21} & \dots & \hat{c}_{2,N-2} \\ \dots & \dots & \dots & \dots \\ \hat{c}_{N0} & \hat{c}_{N1} & \dots & \hat{c}_{N,N-1} \end{bmatrix} = \begin{bmatrix} \hat{c}_1{}' \\ \hat{c}_2{}' \\ \dots \\ \hat{c}_N{}' \end{bmatrix} \tag{3.7}$$

where $\hat{c}_{ik}$ are least squares estimates of the parameters c_{ik} in the representation

$$x_i(t) = \sum_{k=0}^{N-1} c_{ik}\varphi_k(t) \tag{3.8}$$

of the process $X(t)$, $t \in I$, $i = 1,2,\dots,N$.

The unbiased estimator $\hat{\Sigma}$ of the unknown matrix Σ has the following form

$$\hat{\Sigma} = \frac{1}{N-1}\hat{C}'\hat{C}, \tag{3.9}$$

where $\hat{C}$ is given by (3.7).

Then we find the nonzero eigenvalues λ_k and corresponding eigenvectors u_k of matrix $\hat{\Sigma}$. Having determined the eigenvectors u_k we determine its weight functions

$$u_k(t) = u_k'\varphi(t), t \in I. \tag{3.10}$$

Hence, the jth functional principal component $X_i(t)$ is given by

$$U_{ij} = \langle u_j(t), X_i(t)\rangle = \int u_j(t)X_i(t)dt = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} c_{il}u_{jk}\int \varphi_k(t)\varphi_l(t)dt =$$

$$= \sum_{k=0}^{N-1} c_{ik}u_{jk} = c_i'u_j, i = 1,2 \dots, N, j = 1,2, \dots \qquad (3.11)$$

4. A kernel version of functional principal components

The space R^N of values of the random vector $\boldsymbol{c} = (c_0, c_1, \dots, c_{N-1})'$ is transformed by the non-linear function $\boldsymbol{\Psi}$ into a Hilbert space $H(k)$ with a nonnegative definite reproducing kernel k:

$$\boldsymbol{\Psi}: \mathrm{R}^N \to H(k).$$

Then, the Moore–Aronszajn theorem (see Aronszajn (1950)) guarantees a one-to-one correspondence between the functions $\boldsymbol{\Psi}$ and their scalar products.

Let $\hat{\boldsymbol{c}}_1, \hat{\boldsymbol{c}}_2, \dots, \hat{\boldsymbol{c}}_N$ be centred realizations of the random vector $\boldsymbol{c} = (c_0, c_1, \dots, c_{N-1})'$. In order to construct kernel principal components in space $H(k)$ we find the eigenvalues $\lambda > 0$ and the corresponding eigenvectors $\boldsymbol{v} \in H(k)$ of the covariance matrix

$$\widehat{\boldsymbol{\Sigma}}_\Psi = \frac{1}{N}\sum_{k=1}^{N} \Psi(\hat{\boldsymbol{c}}_k)\Psi'(\hat{\boldsymbol{c}}_k) \qquad (4.1)$$

constructed from the N centred and non-linear transformed vectors $\hat{\boldsymbol{c}}_1, \hat{\boldsymbol{c}}_2, \dots, \hat{\boldsymbol{c}}_N$.

If $\boldsymbol{v}$ is an eigenvector of the matrix $\widehat{\boldsymbol{\Sigma}}_\Psi$ corresponding to the eigenvalue λ, then

$$\widehat{\boldsymbol{\Sigma}}_\Psi \boldsymbol{v} = \lambda \boldsymbol{v}$$

or equivalently

$$\boldsymbol{\Psi}'(\hat{\boldsymbol{c}}_i)\widehat{\boldsymbol{\Sigma}}_\Psi \boldsymbol{v} = \lambda \boldsymbol{\Psi}'(\hat{\boldsymbol{c}}_i)\widehat{\boldsymbol{\Sigma}}_\Psi \boldsymbol{v}, i = 1,2, \dots, N. \qquad (4.2)$$

Each eigenvector $\boldsymbol{v}$ must lie in the subspace span$\{\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_1), \boldsymbol{\Psi}(\hat{\boldsymbol{c}}_2), \dots, \boldsymbol{\Psi}(\hat{\boldsymbol{c}}_N)\}$ spanned by the vectors $\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_1), \boldsymbol{\Psi}(\hat{\boldsymbol{c}}_2), \dots, \boldsymbol{\Psi}(\hat{\boldsymbol{c}}_N)$, i.e. the eigenvector $\boldsymbol{v}$ can be written as some linear combination of the vectors $\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_1), \boldsymbol{\Psi}(\hat{\boldsymbol{c}}_2), \dots, \boldsymbol{\Psi}(\hat{\boldsymbol{c}}_N)$. There therefore exist coefficients $\alpha_j, j = 1,2, \dots, N$, such that

$$\boldsymbol{v} = \sum_{j=1}^{N} \alpha_j \boldsymbol{\Psi}'(\hat{\boldsymbol{c}}_j) \qquad (4.3)$$

By substituting (4.3) into (4.2) we obtain:

$$\frac{1}{N}\sum_{j=1}^{N}\alpha_j\,\boldsymbol{\Psi}'(\hat{\boldsymbol{c}}_i)\sum_{k=1}^{N}\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_k)\boldsymbol{\Psi}'(\hat{\boldsymbol{c}}_k)\,\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_j) = \\ = \lambda\sum_{j=1}^{N}\alpha_j\boldsymbol{\Psi}'(\hat{\boldsymbol{c}}_i)\,\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_j), \tag{4.4}$$

for $i = 1,2,\dots,N$. In matrix notation, equation (4.4) takes the form:

$$\boldsymbol{K}^2\boldsymbol{\alpha} = N\lambda\boldsymbol{K}\boldsymbol{\alpha} \tag{4.5}$$

where $\boldsymbol{K} = (k_{ij}), \boldsymbol{\alpha} = (\alpha_1,\alpha_2,\dots,\alpha_N)'$ or

$$\boldsymbol{K}^2\boldsymbol{\alpha} = \tilde{\lambda}\boldsymbol{K}\boldsymbol{\alpha} \tag{4.6}$$

where $\tilde{\lambda} = N\alpha$.

Let us note that every vector $\boldsymbol{\alpha} \neq \boldsymbol{0}$ being a solution to the equation is also a solution to equation (4.6), and that the solutions of (4.6) and (4.7) differ

$$\boldsymbol{K}\boldsymbol{\alpha} = \tilde{\lambda}\boldsymbol{\alpha} \tag{4.7}$$

only by the eigenvectors of matrix $\boldsymbol{K}$ corresponding to zero eigenvalues, which is not significant for the problem of principal components (the principal components correspond only to non-zero eigenvalues).

We assumed earlier that the vectors $\{\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_i)\}, i = 1,2,\dots,N$ are centred. In the general case we cannot centre the vectors $\{\boldsymbol{\Psi}(\hat{\boldsymbol{c}}_i)\}$, because we do not know the form of the function $\boldsymbol{\Psi}$. Let

$$\widetilde{\boldsymbol{\Psi}}_i = \boldsymbol{\Psi}_i - \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{\Psi}_k$$

and

$$\widetilde{\boldsymbol{K}} = (\tilde{k}_{ij}) = (\langle\widetilde{\boldsymbol{\Psi}}_i, \widetilde{\boldsymbol{\Psi}}_j\rangle),$$

where $\boldsymbol{\Psi}_i = \boldsymbol{\Psi}(c_i), i = 1,2,\dots,N$. We cannot compute the matrix $\widetilde{\boldsymbol{K}}$ directly, but we can express it in terms of the matrix $\boldsymbol{K}$:

$$\widetilde{\boldsymbol{K}} = \boldsymbol{PKP},$$

where $\boldsymbol{P} = \left(\delta_{ij} - \frac{1}{N}\right)$ and δ_{ij} is the Kronecker delta. Hence, in the general case the construction of kernel principal components must be based on matrix $\widetilde{\boldsymbol{K}}$**.** Because the kernel matrix $\boldsymbol{K}$ is nonnegative definite, matrix $\widetilde{\boldsymbol{K}}$ is nonnegative definite also. This results from the fact (Seber (1984), p. 521) that if $\boldsymbol{A} \geq 0$, then $\boldsymbol{C}'\boldsymbol{AC} \geq 0$. In our case $\boldsymbol{C} = \boldsymbol{P}$, where $\boldsymbol{P}$ is a symmetric matrix, i.e. $\boldsymbol{P}' = \boldsymbol{P}$. Hence, all eigenvalues of $\widetilde{\boldsymbol{K}}$ are nonnegative.

Having determined the eigenvectors $\boldsymbol{\alpha}_k$ we determine its weight functions

$$u_k(t) = \boldsymbol{\alpha}_k'\varphi(t), t \in I. \tag{4.8}$$

5. Example

The quality of the two described methods (functional principal components analysis and the kernel version of functional principal components analysis) was tested on the 20 different data sets listed in Table 1. The data sets originate from the UCR Time Series Classification/Clustering Homepage (Keogh et al. (2006)).

Table 1. Data sets

Data set	Time series length	Number of classes	Number of observations
50Words	270	50	450
Adiac	176	37	390
Beef	470	5	30
CBF	128	3	30
Coffee	286	2	28
ECG200	96	2	100
Face (all)	131	14	560
Face (four)	350	4	24
Fish	463	7	175
Gun-Point	150	2	50
Lightning-2	637	2	60
Lightning-7	319	7	70
OliveOil	570	4	30
OSU Leaf	427	6	200
Swedish Leaf	128	15	500
Synthetic Control	60	6	300
Trace	275	4	100
Two Patterns	128	4	1000
Wafer	152	2	1000
Yoga	426	2	300

Elements from different classes were combined into one data set. For each data set separately, the discrete time series were centred, and then transformed into continuous functions of the form (2.1). As an orthonormal basis of $L_2(I)$ we took the orthonormal system of Legendre polynomials in the space $L_2([-1,1])$:

$$\tilde{P}_k(x) = \sqrt{\frac{2k+1}{2}} P_k(x),$$

where

$$P_{k+1}(x) = \frac{1}{k+1}[(2k+1)xP_k(x) - kP_{k-1}(x)], k \geq 1,$$

$$P_0(x) = 1, P_1(x) = x.$$

Each finite interval $[a, b]$ can be transformed into the interval $[-1, 1]$ by the substitution

$$x = \frac{2}{b-1} t - \frac{b+a}{b-a}, t \in [a, b], x \in [-1,1].$$

In the case of the kernel version of functional principal components analysis we chose the kernel polynomial function of the form

$$k(\boldsymbol{x}, \boldsymbol{y}) = (1 + \boldsymbol{x}'\boldsymbol{y})^2.$$

The most frequently considered objects are presented on a plot of the first two functional principal components. In this case the criterion of goodness of the constructed functional principal components is the expression:

$$\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} 100\%, \tag{5.1}$$

where $\lambda_1 \geq \lambda_2 \geq$ are the nonzero eigenvalues of the matrix (3.9) or the matrix (4.1). The greater the value of expression (5.1), the greater is the variability shown by the first two functional principal components. The percentage of variability accounted for by the first two functional principal components is given in Table 2. Table 2 shows that better results are obtained in the case of the kernel version of functional principal components analysis.

Table 2. Values of the criterion (5.1)

Data set	FPCA	Kernel version of the FPCA
50Words	83.18	99.48
Adiac	95.85	96.04
Beef	94.89	99.39
CBF	71.92	94.48
Coffee	99.67	100.00
ECG200	99.99	100.00
Face (all)	70.58	97.74
Face (four)	75.14	96.59
Fish	98.66	100.00
Gun-Point	87.37	98.94
Lighting-2	66.79	95.90
Lighting-7	53.68	88.29
OliveOil	100.00	100.00
OSU Leaf	99.99	100.00
Swedish Leaf	91.22	99.86

Table 3. Values of the criterion (5.1) (cont.)

Data set	FPCA	Kernel version of the FPCA
Synthetic Control	79.77	92.16
Trace	100.00	98.98
Two Patterns	76.58	98.80
Wafer	92.76	99.95
Yoga	99.78	99.99
Mean	**85.93**	**97.78**

6. Conclusions

The effectiveness of functional principal components and their new kernel version was compared for 20 different data sets. Efficiency criterion was the ratio of the sum of the first two eigenvalues to the sum of all eigenvalues of the matrix (3.9) or the matrix (4.1). Average efficiency of functional principal components is 85.93%, while the average efficiency of the kernel version of functional principal components is 97.78%. This is a significant increase in efficiency, which supports the use of kernel version of functional principal components.

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels, Trans. Amer. Math. Soc. 68, 337–404.

KEOGH, E., XI, X., WEI, L. & RATANAMAHATANA, C. A. (2006). The UCR Time Series Classification/Clustering Homepage, http://www.cs.ucr.edu/~eamonn/time_series_data/.

RAMSAY, J. O., DALZELL, C. J. (1991). Some tools for functional data analysis, J. Royal Statist. Soc. B 53, 539–572.

RAMSAY, J. O., SILVERMAN, B. W. (2002). Applied Functional Data Analysis, Springer, New York.

RAMSAY, J. O., SILVERMAN, B. W. (2005). Functional Data Analysis, Springer, New York.

SEBER, G. A. F. (1984). Multivariate Observations, Wiley, New York.