Colloquium Biometricum 40
2010, 105–116

NONLINEAR PRINCIPAL COMPONENT ANALYSIS

Karol Deręgowski[1], Mirosław Krzyśko[1,2]

[1]President Stanislaw Wojciechowski Higher Vocational State School in Kalisz
Institute of Management
Nowy Świat 4, 62-800 Kalisz
[2]Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Umultowska 87, 61-614 Poznań
e-mail: kadere@o2.pl; mkrzysko@amu.edu.pl

Summary

In this paper, we propose two approaches to the construction of nonlinear principal components. By the use of kernel functions, one can efficiently compute nonlinear principal components in high-dimensional feature space, related to input space by some nonlinear transformation.

Key words and phrases: nonlinear principal component analysis, kernel functions

Classification AMS 2010: 62H25

1. Introduction

Classical principal component analysis (PCA) (Hotelling, 1933) was introduced as a technique for deriving a reduced set of orthogonal linear projections of a single collection of correlated variables $\mathbf{X} = (X_1, X_2,\ldots, X_p)^T$, where the projections are ordered by decreasing variances. Principal component analysis is used, for example, in lossy data compression, pattern recognition, and image analysis. In addition to reducing dimensionality, principal component analysis can be used to discover important features of the data. Discovery in principal

component analysis takes the form of graphical displays of the principal component scores. The first few principal component scores can reveal whether most of the data actually live on a linear subspace of R^p, and can be used to identify outliers, distributional peculiarities, and clusters of points. The last few principal component scores show those linear projections of $\mathbf{X} = (X_1, X_2,\ldots, X_p)^T$ that have the smallest variance; any principal component with zero or near-zero variance is virtually constant, and hence can be used to detect collinearity, as well as outliers that appear and alter the perceived dimensionality of the data.

The linear projection method can be extremely useful in discovering low-dimensional structure when the data actually lie in a linear (or approximately linear) lower-dimensional subspace (called a manifold) M of input space R^p. But what can we do if we know or suspect that the data actually lie on a low-dimensional nonlinear manifold, whose structure and dimensionality are both assumed unknown? We can then construct the nonlinear principal components. In Section 2, the classical principal components are presented. In Section 3 we show two approaches to the construction of nonlinear principal components. In Section 4 we present an example.

2. Classical principal component analysis

Assume that the random p-vector $\mathbf{X} = (X_1, X_2,\ldots, X_p)^T$ has mean $\boldsymbol{\mu}$ and (p×p) covariance matrix $\boldsymbol{\Sigma}$. PCA seeks to replace the set of p (unordered and correlated) input variables, $X_1, X_2,\ldots, X_p$ by a (potentially smaller) set of t (ordered and uncorrelated) linear projections, $\xi_1,\ldots, \xi_t$ ($t \le p$), of the input variables,

$$\xi_j = \mathbf{b}_j^T \mathbf{X} = b_{j1}X_1 +\ldots+ b_{jp}X_p \text{ , } j = 1, 2,\ldots, t; \tag{2.1}$$

where we minimize the loss of information due to replacement.

In PCA, "information" is interpreted as the "total variation" of the original input variables,

$$\sum_{j=1}^{p} \mathrm{Var}\left(X_j\right) = \mathrm{tr}\,(\boldsymbol{\Sigma}).$$

From the spectral decomposition theorem, we can write

$$\boldsymbol{\Sigma} = \mathbf{U}\,\boldsymbol{\Lambda}\,\mathbf{U}^{\mathrm{T}},\ \ \mathbf{U}^{\mathrm{T}}\,\mathbf{U}\ = \mathbf{I}_{\mathrm{p}},$$

where the diagonal matrix $\boldsymbol{\Lambda}$ has as diagonal elements the eigenvalues $\{\lambda_j\}$ of $\boldsymbol{\Sigma}$, and the columns of $\mathbf{U}$ are the eigenvectors of $\boldsymbol{\Sigma}$. Thus the total variation is $\mathrm{tr}\,(\boldsymbol{\Sigma}) = \mathrm{tr}\,(\boldsymbol{\Lambda}) = \sum_{j=1}^{p} \lambda_j$.

The jth coefficient vector, $\mathbf{b}_j = (b_{j1},\ldots,b_{jp})^{\mathrm{T}}$, is chosen so that:

- The first t linear projections ξ_j, j = 1, 2,…, t, of $\mathbf{X}$ are ranked in importance through their variances $\{\mathrm{Var}(\xi_j)\}$, which are listed in decreasing order of magnitude: $\mathrm{Var}(\xi_1) \geq \mathrm{Var}(\xi_2) \geq \ldots \geq \mathrm{Var}(\xi_t)$.
- ξ_j is uncorrelated with all ξ_k, k < j.

The linear projections (2.1) are then known as the first t principal components of $\mathbf{X}$.

In practice, we estimate the principal components using N independent observations, $\{\mathbf{X}_i, i = 1, 2,\ldots, N\}$, on $\mathbf{X}$. We estimate $\boldsymbol{\mu}$ by

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = \mathrm{N}^{-1}\sum_{i=1}^{N} \mathbf{X}_i.$$

Let $\mathbf{X}_{ic} = \mathbf{X}_i - \overline{\mathbf{X}}$, i = 1, 2,…, N, and set $\mathbf{X}_c = \begin{bmatrix} \mathbf{X}_{1c}^{\mathrm{T}} \\ \cdots \\ \mathbf{X}_{Nc}^{\mathrm{T}} \end{bmatrix}$ to be an (N×p) matrix.

We estimate $\boldsymbol{\Sigma}$ by the sample covariance matrix,

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = (\mathrm{N}-1)^{-1}\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c. \tag{2.2}$$

The ordered eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are denoted by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$, and the eigenvector associated with the jth largest sample eigenvalue $\hat{\lambda}_j$ is the jth sample eigenvector $\hat{\mathbf{v}}_j$, j = 1, 2,…, p.

The jth sample PC score of $\mathbf{X}$ is given by

$$\hat{\boldsymbol{\xi}}_{ij} = \hat{\mathbf{v}}_j^{\mathrm{T}}\mathbf{X}_{ic}, \tag{2.3}$$

where $\mathbf{X}_{ic} = \mathbf{X}_i - \overline{\mathbf{X}}$, i = 1, 2,…, N, j = 1, 2,…, p.

A sample measure of how well the first t principal components represent the p original variables is given by the statistic

$$\frac{\hat{\lambda}_1 + ... + \hat{\lambda}_t}{\hat{\lambda}_1 + ... + \hat{\lambda}_p}$$

which is the proportion of the total sample variation that is explained by the first t sample principal components.

It is hoped that the sample variances of the first few sample PCs will be large, whereas the rest will be small enough for the corresponding set of sample PCs to be omitted. A variable that does not change much (relative to other variables) in independent measurements may be treated approximately as a constant, and so omitting such low-variance sample PCs and focusing exclusively on the high-variance sample PCs is therefore a convenient way of reducing the dimensionality of the data set.

For diagnostic and data analytic purposes, it is usual to plot the first sample PC scores against the second sample PC scores, $(\hat{\xi}_{i1}, \hat{\xi}_{i2})$, where $\hat{\xi}_{ij}$, is given by (2.3), i = 1,…, N, j = 1, 2.

3. Nonlinear principal component analysis

An approach that generalizes linear PCA is given by Kernel PCA (Scholkopf, Smola, and Muller, 1996). This is an application of so-called kernel methods.

Let $\mathbf{X}_{ic} = \mathbf{X}_i - \overline{\mathbf{X}} \in R^p$, i = 1, 2,…, N, be the input centered data points. We can think of kernel PCA as a two-step process:

1. Nonlinearly transform the ith input center data point $\mathbf{X}_{ic} \in R^p$ into point $\Phi(\mathbf{X}_{ic})$ in an N_H-dimensional feature space H (the Hilbert space), where

$$\mathbf{\Phi}(\mathbf{X}_{ic}) = (\Phi_1(\mathbf{X}_{ic}),\ldots,\Phi_{N_H}(\mathbf{X}_{ic}))^T \in H,\ i = 1, 2,\ldots, N.$$

   The transformation $\mathbf{\Phi} : R^p \to H$ is called a feature transformation, and each of the $\{\Phi_j\}$ is a nonlinear transformation.

2. Given $\mathbf{\Phi}(\mathbf{X}_{1c}),\ldots, \mathbf{\Phi}(\mathbf{X}_{Nc}) \in H$, solve a linear PCA problem in feature space H, which has a higher dimensionality than that of the input space (i.e. $N_H > p$).

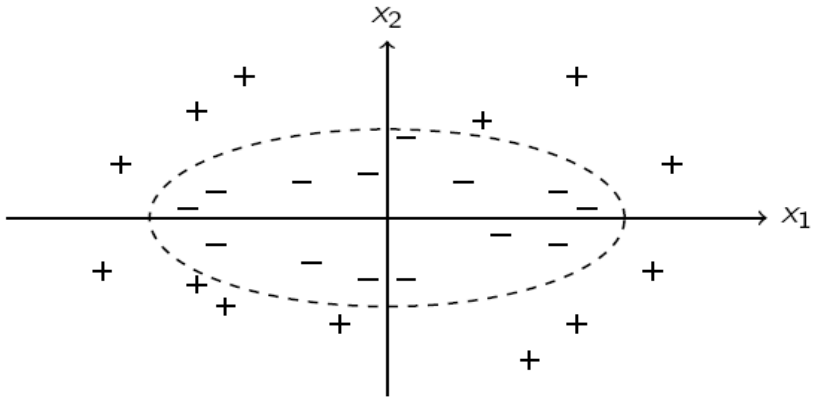

Fig. 1. Original data in the plane. The data cannot be separated linearly.

Consider the data presented in Figure 1. Let $\mathbf{X}_{ic} = (X_{ic1}, X_{ic2})^T$, and define $\mathbf{\Phi} : R^2 \to R^3$ by

$$\mathbf{\Phi}(\mathbf{X}_{ic}) = \mathbf{\Phi}(X_{ic1}, X_{ic2}) = (X_{ic1}^2, \sqrt{2}X_{ic1}X_{ic2}, X_{ic2}^2) = (z_{i1}, z_{i2}, z_{i3})^T.$$

With this $\mathbf{\Phi}$, a difficult nonlinear classification problem in R^2 is converted to a standard linear classification task in R^3 (see Figure 2).

Let $\mathbf{X}_{ic} = (X_{ic1}, X_{ic2})^T$ and $\mathbf{Y}_{ic} = (Y_{ic1}, Y_{ic2})^T$ be two vectors in input space R^2, and consider the transformation to R^3 used earlier. Let $\mathbf{\Phi}(\mathbf{X}_{ic})$ and $\mathbf{\Phi}(\mathbf{Y}_{ic})$ be two feature vectors generated by $\mathbf{X}_{ic}$ and $\mathbf{Y}_{ic}$. Now look at the inner product $\mathbf{\Phi}^T(\mathbf{X}_{ic})\ \mathbf{\Phi}(\mathbf{Y}_{ic})$ in feature space. It is

$$\mathbf{\Phi}^T(\mathbf{X}_{ic})\ \mathbf{\Phi}(\mathbf{Y}_{ic}) = (X_{ic1}^2, \sqrt{2}X_{ic1}X_{ic2}, X_{ic2}^2)\ (Y_{ic1}^2, \sqrt{2}Y_{ic1}Y_{ic2}, Y_{ic2}^2)^T =$$

$$= (X_{ic1}Y_{ic1} + X_{ic2}Y_{ic2})^2 = (\mathbf{X}_{ic}^T\mathbf{Y}_{ic})^2 = k(\mathbf{X}_{ic}, \mathbf{Y}_{ic}). \qquad (3.1)$$

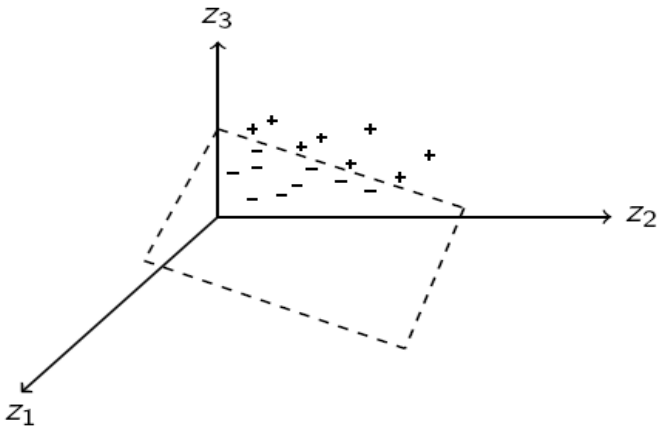

Fig. 2. A three-dimensional representation of the pluses and minuses.

Equation (3.1) shows how an inner product based on $\mathbf{\Phi}$ converts to a function of the two inputs. Since choosing an inner product and computing with it in feature space can quickly become computationally infeasible, it would be desirable to choose a function k, called a kernel, so as to summarize the geometry of feature space vectors and ignore $\mathbf{\Phi}$ entirely.

Now the kernel trick can be applied. Suppose a function $k(\cdot,\cdot) : R^p \times R^p \rightarrow R$ operating on input space can be found such that the feature space inner products are computed directly through k as in (3.1). Then explicit use of $\mathbf{\Phi}$ has been avoided, and yet results can be obtained as if $\mathbf{\Phi}$ were used. This direct computation of feature space inner products without actually explicitly manipulating the feature space vectors themselves is known as the kernel trick.

The existence of the transformation $\mathbf{\Phi} : R^p \rightarrow H$ such that

$$\mathbf{\Phi}^T(\mathbf{X}_{ic})\mathbf{\Phi}(\mathbf{Y}_{ic}) = k(\mathbf{X}_{ic}, \mathbf{Y}_{ic})$$

guarantees the following theorem.

Theorem 1 (Mercer, 1909). Let

$$k : R^p \times R^p \rightarrow R$$

be a bivariate symmetric continuous real-valued function. Then there exists a transformation $\mathbf{\Phi} : R^p \rightarrow H$ such that

$$k(\mathbf{X}_{ic}, \mathbf{Y}_{ic}) = \mathbf{\Phi}^T(\mathbf{X}_{ic})\mathbf{\Phi}(\mathbf{Y}_{ic})$$

if and only if the matrix $\mathbf{K} = (k_{ij})$ is nonnegative definite, where $k_{ij} = k(\mathbf{X}_{ic}, \mathbf{X}_{jc})$, i, j = 1,…,N.

The matrix **K** is known as the kernel Mercer's matrix. For a given bivariate function k, verifying the conditions above might not be easy. In practice, there exist many functions that have been shown to be valid kernels, and fortunately many of them deliver good performance on real-world data.

A short annotated list is presented in Table 1.

Table 1. Kernel functions

Kernel	k(**x**, **y**)
Homogeneous polynomial kernel	$(\mathbf{x}^T \mathbf{y})^d$, d is an integer
Inhomogeneous polynomial kernel	$(\mathbf{x}^T \mathbf{y} + c)^d$, c > 0
Gaussian radial basis function	$\exp\left(-c\|\mathbf{x}-\mathbf{y}\|^2\right)$, c > 0
Laplacian	$\exp\left(-c\|\mathbf{x}-\mathbf{y}\|\right)$, c > 0

In order to carry out linear PCA in feature space so that it mimics the standard treatment of PCA (as carried out in input space), we have to find eigenvalues $\gamma \geq 0$ and nonzero eigenvectors $\mathbf{u} \in H$ of the estimated covariance matrix

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N} \mathbf{\Phi}(\mathbf{X}_{ic})\,\mathbf{\Phi}^T(\mathbf{X}_{ic}) \tag{3.2}$$

of the centered and nonlinearly transformed input vectors. The eigenequation $\mathbf{Cu} = \gamma\mathbf{u}$, where **u** is the eigenvector corresponding to the eigenvalue $\gamma \geq 0$ of **C**, can be written in an equivalent form as

$$\mathbf{\Phi}^T(\mathbf{X}_{ic})\,\mathbf{Cu} = \gamma\,\mathbf{\Phi}^T(\mathbf{X}_{ic})\mathbf{u}\,,\ i = 1, 2,\ldots, N. \tag{3.3}$$

Because

$$\mathbf{Cu} = \frac{1}{N-1}\sum_{i=1}^{N} \mathbf{\Phi}(\mathbf{X}_{ic})\,\mathbf{\Phi}^T(\mathbf{X}_{ic})\mathbf{u}$$

all solutions **u** with nonzero eigenvalue γ are contained in the span of $\mathbf{\Phi}(\mathbf{X}_{1c}),\ldots, \mathbf{\Phi}(\mathbf{X}_{Nc})$. Hence there exist coefficients, α_k, k = 1, 2,…, N, such that

$$\mathbf{u} = \sum_{k=1}^{N} \alpha_k \mathbf{\Phi}(\mathbf{X}_{kc}). \tag{3.4}$$

Substituting (3.4) for **u** in (3.3), we get that

$$\frac{1}{N-1} \sum_{j=1}^{N} \alpha_j \mathbf{\Phi}^T(\mathbf{X}_{ic}) \sum_{k=1}^{N} \mathbf{\Phi}(\mathbf{X}_{kc}) \mathbf{\Phi}^T(\mathbf{X}_{kc}) \mathbf{\Phi}(\mathbf{X}_{jc}) = \gamma \sum_{k=1}^{N} \alpha_k \mathbf{\Phi}^T(\mathbf{X}_{ic}) \mathbf{\Phi}(\mathbf{X}_{kc}), \tag{3.5}$$

for all i = 1, 2,…, N.
The eigenequation (3.5) can be written as $\mathbf{K}^2 \boldsymbol{\alpha} = N \gamma \mathbf{K} \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_1,\ldots, \alpha_N)^T$, or as

$$\mathbf{K}^2 \boldsymbol{\alpha} = \tilde{\gamma} \mathbf{K} \boldsymbol{\alpha}, \tag{3.6}$$

where $\tilde{\gamma} = (N - 1) \gamma$, $\mathbf{K} = (k_{ij})$ and $k_{ij} = k(\mathbf{X}_{ic}, \mathbf{X}_{jc}) = \mathbf{\Phi}^T(\mathbf{X}_{ic}) \mathbf{\Phi}(\mathbf{X}_{jc})$, i, j = 1, 2,…, N.
To find solutions of (3.6), we solve the eigenvalue problem

$$\mathbf{K} \boldsymbol{\alpha} = \tilde{\gamma} \boldsymbol{\alpha} \tag{3.7}$$

for nonzero eigenvalues. Clearly, all solutions of (3.7) do satisfy (3.6). Moreover, it can be shown that any additional solutions of (3.7) do not make a difference in the expansion (3.4) and thus are not of interest to us.
Let us consider the problem of nonlinear principal components from the standpoint of classical principal components.
The classical principal components are determined from the sample covariance matrix **S** of the form (2.2). The ordered eigenvectors of **S**, $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2,\ldots, \hat{\mathbf{v}}_p$, satisfy the equations

$$\mathbf{S}\hat{\mathbf{v}}_j = \hat{\lambda}_j \hat{\mathbf{v}}_j, \; j = 1, 2,\ldots, p. \tag{3.8}$$

Equations (2.2) and (3.8) together lead to

$$(N-1)^{-1} \mathbf{X}_c^T \mathbf{X}_c \hat{\mathbf{v}}_j = \hat{\boldsymbol{\lambda}}_j \hat{\mathbf{v}}_j, \; j = 1, 2,\ldots, p. \tag{3.9}$$

We will put $\hat{\mathbf{v}}_j = \mathbf{X}_c^T \boldsymbol{\alpha}_j$, j = 1, 2,…, p, into (3.9) and reparameterize the eigenvalue problem in terms of $\boldsymbol{\alpha}_j$. For j = 1, 2,…, p, this leads to

$$\mathbf{X}_c^T \mathbf{X}_c \mathbf{X}_c^T \boldsymbol{\alpha}_j = \hat{\beta}_j \mathbf{X}_c^T \boldsymbol{\alpha}_j,$$

where $\hat{\beta}_j = (N-1)\hat{\lambda}_j$, j = 1, 2,…, p.
If we left-multiply both sides by $\mathbf{X}_c$, we get

$$\mathbf{X}_c \mathbf{X}_c^T \mathbf{X}_c \mathbf{X}_c^T \boldsymbol{\alpha}_j = \hat{\beta}_j \mathbf{X}_c \mathbf{X}_c^T \boldsymbol{\alpha}_j, \; j = 1, 2, \ldots, p.$$

Let us observe that

$$\mathbf{X}_c \mathbf{X}_c^T = \begin{bmatrix} \mathbf{X}_{1c}^T \\ \cdots \\ \mathbf{X}_{Nc}^T \end{bmatrix} \left[\mathbf{X}_{1c}, \ldots, \mathbf{X}_{Nc} \right] = (\mathbf{X}_{ic}{}^T \mathbf{X}_{jc})$$

is an N×N matrix of pairwise inner products.

Therefore, if a linear algorithm can be shown to depend on the centered data matrix $\mathbf{X}_c$ only through an N×N matrix $\mathbf{X}_c\mathbf{X}_c{}^T$ of pairwise inner products, then it can be easily "kernelized" – we simply replace $\mathbf{X}_c\mathbf{X}_c{}^T$ by the kernel Mercer's matrix **K**.
Hence, the eigenequations (3.2) can be written as

$$\mathbf{K}^2 \boldsymbol{\alpha} = \hat{\beta}\, \mathbf{K}\, \boldsymbol{\alpha}$$

and are identical to (3.6).

Since the linear algorithm of principal components depends on the centered data matrix $\mathbf{X}_c$ only through the matrix $\mathbf{X}_c\mathbf{X}_c{}^T$ of pairwise inner products, it can be easily transformed to a nonlinear algorithm by replacing $\mathbf{X}_c\mathbf{X}_c{}^T$ by the kernel Mercer's matrix **K**.

Once we obtain the $\hat{\boldsymbol{\alpha}}_i$'s, suppose we would like to project the data $\mathbf{X}_c$ onto a few leading principal components, for example $\mathbf{X}_c \hat{\mathbf{v}}_j$. We immediately find that

$$\mathbf{X}_c \hat{\mathbf{v}}_j = \mathbf{X}_c \mathbf{X}_c^T \hat{\boldsymbol{\alpha}}_j = \mathbf{K} \hat{\boldsymbol{\alpha}}_j.$$

Hence it becomes clear that both finding and projecting onto principal components depends on just the inner products, and PCA can be "kernelized" easily.

4. Example

One hundred points were generated from a uniform distribution in a circle of radius 1, one hundred points from a uniform distribution in a circle of radius 2, and a hundred points from a uniform distribution in a circle of radius 4. In the last two cases, Gaussian noise with standard deviation 0.25 was added to each point. The data are shown in Figure 3.

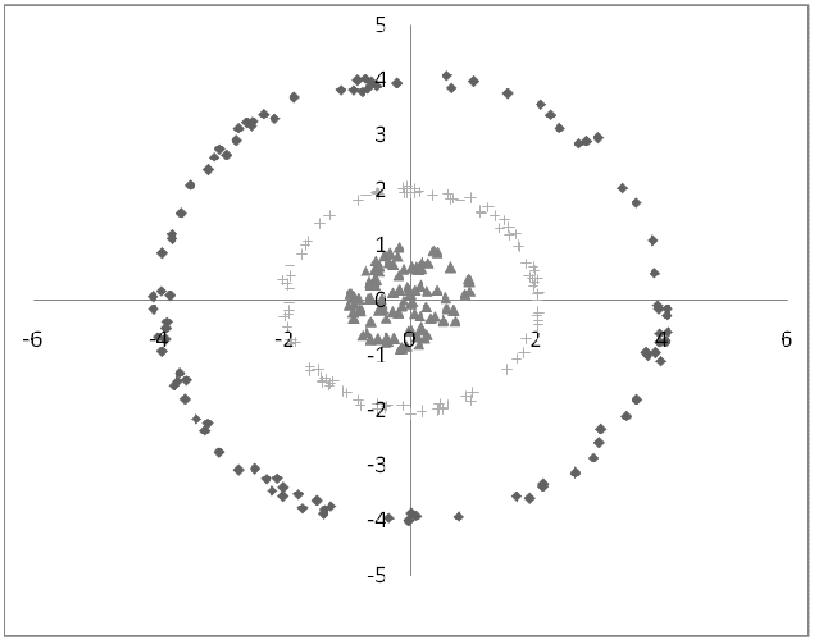

Fig. 3. The original data

The data being spherical, all directions have equal variance and there are no meaningful principal components in the traditional sense. We construct the kernel Mercer's matrix $\mathbf{K} = (k_{ij})$, where

$$k_{ij} = k(\mathbf{X}_{ic}, \mathbf{Y}_{ic}) = \exp\left(-c \left\|\mathbf{X}_{ic} - \mathbf{Y}_{ic}\right\|^2\right),\ c > 0,\ i, j = 1, 2, \ldots, 300.$$

Since the kernel function k depends on the parameter c, all sizes **K**, $\boldsymbol{\alpha}$ and λ depend on this parameter.
The graph of $\lambda_1(c)/\sum_j \lambda_j(c)$ and $\lambda_2(c)/\sum_j \lambda_j(c)$ is shown in Figure 4.

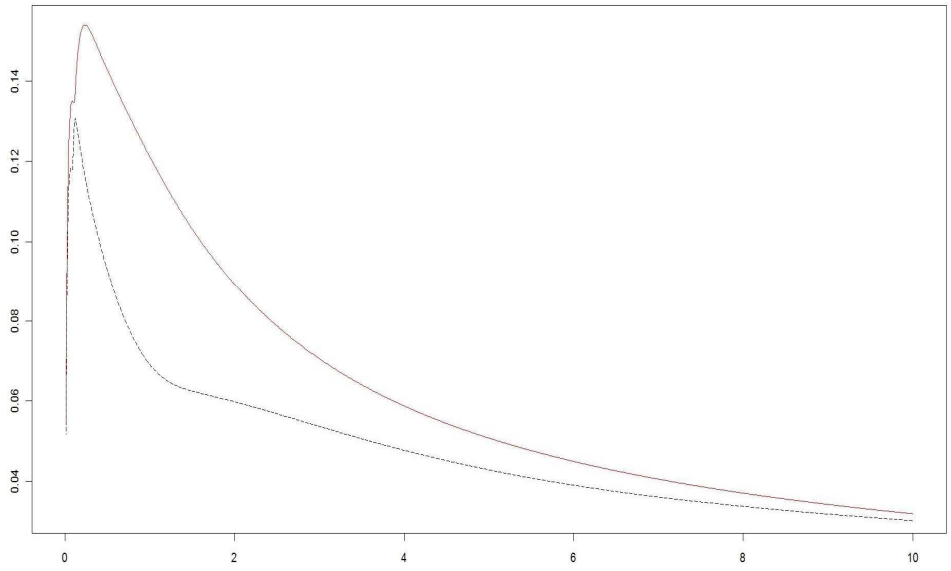

Fig. 4. The graph of $\lambda_1(c)/\sum_j \lambda_j(c)$ and $\lambda_2(c)/\sum_j \lambda_j(c)$

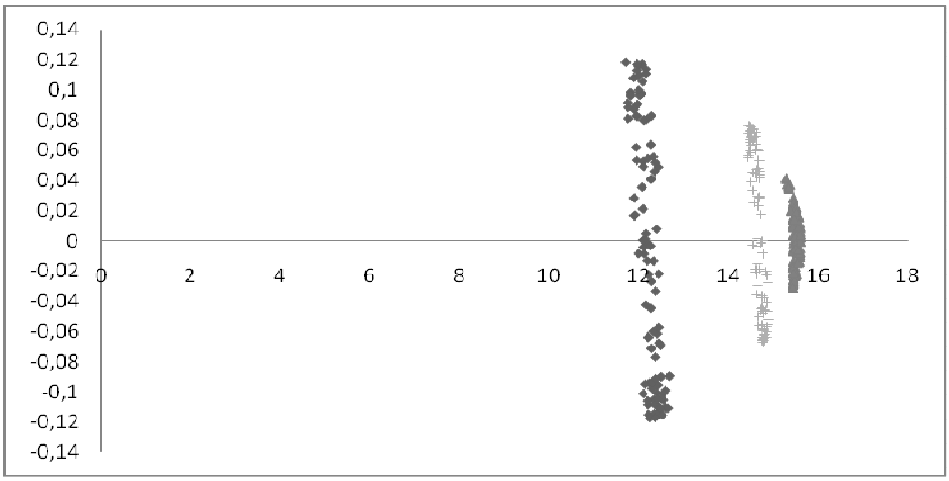

Fig. 5. Projection onto the first two kernel principal components

The function

$$\frac{\lambda_1(c)+\lambda_2(c)}{\sum_j \lambda_j(c)}$$

has a maximum at the point c = 0.017.
Using a Gaussian kernel with c = 0.017 in place of all the inner products, the first kernel principal direction obtained gives a meaningful order of how far each observation is away from the origin (see Figure 5).

In this case, kernel PCA has successfully discovered the (only) underlying pattern in the data, one that is impossible to detect with classical PCA.

Acknowledgments

We would like to thank the anonymous referee for helping us improve this manuscript.

References

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417–441, 498–520.

Mercer J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, Series A, 209, 415–446.

Scholkopf B., Smola A., Muller K.B. (1998). Nonlinear component analysis as a kernel eigenvalues problem. *Neural Computation* 10, 1299–1319.

NIELINIOWE SKŁADOWE GŁÓWNE

Streszczenie

W artykule tym proponujemy dwa podejścia do konstrukcji nieliniowych składowych głównych. Nieliniowe składowe główne można efektywnie skonstruować w nowej przestrzeni cech dużego wymiaru uzyskanej z przestrzeni wyjściowej za pomocą przekształcenia nieliniowego.

Słowa kluczowe: nieliniowe składowe główne, funkcje jądrowe

Klasyfikacja AMS 2010: 62H25