*Discussiones Mathematicae*
*Probability and Statistics 34 (2014) 127–141*
*doi:10.7151/dmps.1163*

# A LEARNING ALGORITHM COMBINING FUNCTIONAL DISCRIMINANT COORDINATES AND FUNCTIONAL PRINCIPAL COMPONENTS

Tomasz Górecki

*Adam Mickiewicz University*
*Faculty of Mathematics and Computer Science*
*Umultowska 87, 61–614 Poznań, Poland*

**e-mail:** tomasz.gorecki@amu.edu.pl

AND

Mirosław Krzyśko

*Adam Mickiewicz University*
*Faculty of Mathematics and Computer Science*
*Umultowska 87, 61–614 Poznań, Poland*
*President Stanisław Wojciechowski Higher Vocational State School in Kalisz*
*Faculty of Management*
*Nowy Świat 4, 62–800 Kalisz, Poland*

**e-mail:** mkrzysko@amu.edu.pl

**Abstract**

A new type of discriminant space for functional data is presented, combining the advantages of a functional discriminant coordinate space and a functional principal component space. In order to provide a comprehensive comparison, we conducted a set of experiments, testing effectiveness on 35 functional data sets (time series). Experiments show that constructed combined space provides a higher quality of classification of LDA method compared with component spaces.

**Keywords:** functional principal components, functional discriminant coordinates.

**2010 Mathematics Subject Classification:** 62H25, 62H30, 68T10.

## 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 $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$, 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 find 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 $\mathbb{R}^p$, and can be used to identify outliers, distributional peculiarities, and clusters of points. The last few principal component scores show those linear projections of $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ 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 affect the perceived dimensionality of the data.

When we have samples originating from $L$ groups, we would often like to present them graphically, to see their configuration or to eliminate outlying observations. However it may be difficult to produce such a presentation even if only three features are observed, and with a higher number of features it becomes impossible. A different method must therefore be sought for presenting multidimensional data originating from multiple groups. To make the task easier, in the first step every $p$-dimensional observation $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ can be transformed into a one-dimensional observation $u_1 = a_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$, and the resulting one-dimensional observations can be presented graphically as points on a line. In the second step we can define a second linear combination $u_2 = a_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$ not correlated with the first, and present the observations graphically as points on a plane. Generally, the aim is to construct new uncorrelated variables $u_1, u_2, \ldots, u_s$, $s = \min(L-1, p)$, which will be linear combinations of the original observations $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ and which will discriminate the $L$ groups to a maximum degree; that is to say, in the new system the centers of the L groups will be maximally spaced, and the observations from a given group will be maximally concentrated around its center. These new variables are called discriminant coordinates (see Seber (1984), p. 270). They are also sometimes called canonical variates, but this name is misleading, because canonical variables with completely different properties occur in canonical correlation analysis. Another name used is "discriminant functions" – this is inappropriate because discriminant functions are surfaces that separate $L$ groups from one another. The space of discriminant coordinates is a space which is convenient for the application of various classification methods (methods of

discriminant analysis). In the case $L = 2$ we obtain only one discriminant coordinate, coinciding with the well-known linear discriminant function of Fisher (1936).

Functional principal components (FPC) and functional discriminant coordinates (FDC), which are the extensions of principal components and discriminant coordinates, respectively, from a vector domain to a functional domain, are two very popular feature extraction methods. The functional discriminant coordinates space has proven to be a very powerful space for pattern recognition. However, further study shows that there are still drawbacks in this method. One of the major drawbacks of the functional discriminant coordinate method is that it will lose the within-class scatter information for so-called "small sample size" problems because all the optimal discriminant vectors in this case are limited to the null space of the within-class scatter matrix and this information is also important for pattern recognition. To improve the performance of pattern recognition, we propose another learning algorithm combining the advantages of FPC and FDC. Our proposed algorithm can be divided into three steps:

1. compute the optimal discriminant vectors of FDC;
2. compute the optimal vectors of FPC;
3. use the two kinds of features for recognition.

The paper is organized as follows. In section 2, transformation of discrete data to functional data is presented. Functional principal components are presented in section 3. In section 4, functional discriminant coordinates are presented. The discriminant algorithm based on a combination of features from the kernel discriminant coordinates space and kernel principal components space is described in section 5. Finally, section 6 examines the quality of the new discriminant algorithm presented in this paper on real data sets, as well as statistical analysis of the results. We conclude in section 7.

## 2. Transformation of discrete data to functional data

Many financial, meteorological and other data are recorded at discrete moments in time. Let $x_j$ denote an observed value of feature $X$ at the $j$th time point $t_j$, where $j = 1, 2, \ldots, J$. Then our data consist of $J$ pairs $(t_j, x_j)$. This discrete data can be smoothed by continuous functions $x(t)$, where $t \in I$ (Ramsay and Silverman, 2005). Le $I$ be a compact set such that $t_j \in I$, for $j = 1, \ldots, J$. Let us assume that the function $x(t)$ has the following representation:

$$x(t) = \sum_{k=0}^{K} c_k \varphi_k(t), \qquad t \in I, \tag{1}$$

where $\{\varphi_k\}$ are orthonormal basis functions, and $c_0, c_1, \ldots, c_K$ are the coefficients.

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_J)'$, $\boldsymbol{c} = (c_0, c_1, \ldots, c_K)'$ and $\boldsymbol{\Phi}(t)$ be a matrix of dimension $J \times (K+1)$ containing the values $\varphi_k(t_j)$, $k = 0, 1, \ldots, K, j = 1, 2, \ldots, J$. The coefficient $\boldsymbol{c}$ in (1) is estimated by the least squares method, that is, so as to minimize the function:

$$S(\boldsymbol{c}) = (\boldsymbol{x} - \boldsymbol{\Phi}(t)\boldsymbol{c})' (\boldsymbol{x} - \boldsymbol{\Phi}(t)\boldsymbol{c}) .$$

Differentiating $S(\boldsymbol{c})$ with respect to the vector $\boldsymbol{c}$, we obtain the least squares method estimator

$$\hat{\boldsymbol{c}} = \left(\boldsymbol{\Phi}'(t)\boldsymbol{\Phi}(t)\right)^{-1} \boldsymbol{\Phi}'(t)\boldsymbol{x}.$$

Smoothing of the function $x(t)$ depends on the value $K$ (a small value of $K$ causes more smoothing of the data). The optimum value for $K$ is selected using the Bayesian information criterion BIC (see Shmueli, 2010):

$$\mathrm{BIC}(x(t)) = \ln\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{2}\right) + (K+1)\left(\frac{\ln J}{J}\right),$$

where $\boldsymbol{e} = (e_1, \ldots, e_J)'$, $e_j = x_j - \sum_{b=0}^{K} \hat{c}_k \varphi_k(t_j)$,$j = 1, 2, \ldots, J$.

Let us assume that there are $N$ independent pairs of values $(t_{ij}, x_{ij})$, $j = 1, \ldots, J$, $i = 1, \ldots, N$. These discrete data are smoothed to continuous functions in the following form:

$$x_i(t) = \sum_{k=0}^{K_i} \hat{c}_{ik}\varphi_k(t), i = 1, \ldots, N, t \in I.$$

Among all the $K_1, K_2, \ldots, K_N$ one common value of $K$ is chosen, as the modal value of the numbers $K_1, K_2, \ldots, K_N$, and we assume that each function $x_i(t)$ has the form

$$x_i(t) = \sum_{k=0}^{K} \hat{c}_{ik}\varphi_k(t), i = 1, \ldots, N, t \in I.$$

The set of functions $\{x_1(t), \ldots, x_N(t)\}$ obtained in this way is called functional data (see Ramsay and Silverman, 2005).

## 3. Construction of functional principal components

Let $x_1(t), x_2(t), \ldots, x_N(t)$ be $N$ independent realizations of a stochastic process $X(t)$ with continuous parameter $t \in I$. We will further assume that $E(X(t)) = 0$

and $X(t) \in L_2(I)$, where $L_2(I)$ is a Hilbert space of square integrable functions on the interval $I$ equipped with the following inner product:

$$< u(t), v(t) >= \int_I u(t)v(t)dt.$$

We consider the case, where the process $X(t)$ can be represented by a finite number of orthonormal basis functions $\{\varphi_k(t)\}$

$$X(t) = \sum_{k=0}^{K} c_k \varphi_k(t),\ t \in I, \tag{2}$$

where $\{c_k\}$ are random variables, that $E(c_k) = 0$, $\mathrm{Var}(c_k) < \infty$, $k = 0, 1, \ldots, K$.

Let

$$\boldsymbol{c} = (c_0, c_1, \ldots, c_K)'$$

and

$$\boldsymbol{\varphi}(t) = (\varphi_0(t), \varphi_1(t), \ldots, \varphi_K(t))',\ 0 < K < \infty.$$

Then

$$X(t) = \boldsymbol{c}'\boldsymbol{\varphi}(t),\ t \in I, \tag{3}$$

with $E(\boldsymbol{c}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{c}) = \boldsymbol{\Sigma}$.

In functional principal component analysis, we are interested in finding the inner product

$$U =< u(t), X(t) >= \int_I u(t)X(t)dt$$

having maximal variance for all $u(t) \in L_2(I)$ such that $< u(t), u(t) >= 1$.

Let

$$\lambda_1 = \sup_{u(t) \in L_2(I)} \mathrm{Var}(< u(t), X(t) >) = \mathrm{Var}(< u_1(t), X(t) >),$$

where $< u_1(t), u_1(t) >= 1$. The inner product $U_1 =< u_1(t), X(t) >$ will be called the first functional principal component, and the function $u_1(t)$ will be called the first weight function. Subsequently we look for the second principal component $U_2 =< u_2(t), X(t) >$, which maximizes $\mathrm{Var}(< u(t), X(t) >)$, is such that $< u_2(t), u_2(t) >= 1$, and is not correlated with the first functional principal component $U_1$, i.e., is subject to the restriction $< u_1(t), u_2(t) >= 0$.

In general, the $k$th functional principal component $U_k =< u_k(t), X(t) >$ satisfies the conditions:

$$\lambda_k = \sup_{u(t)\in L_2(I)} \mathrm{Var}(< u(t), X(t) >) = \mathrm{Var}(< u_k(t), X(t) >),$$
$$< u_i(t), u_j(t) >= \delta_{ij},\ \ i,j = 1,2,\ldots,k.$$

The expression $(\lambda_k, u_k(t))$ will be called the $k$th principal system of the process $X(t)$.

Let us consider the principal component of the random vector $\boldsymbol{c}$. The $k$th principal component $U_k^* =< \boldsymbol{u}_k, \boldsymbol{c} >$ of this vector satisfies conditions:

$$\gamma_k = \sup_{\boldsymbol{u}\in\mathbb{R}^{K+1}} \mathrm{Var}(< \boldsymbol{u}, \boldsymbol{c} >) = \sup_{\boldsymbol{u}\in\mathbb{R}^{K+1}} \boldsymbol{u}'\,\mathrm{Var}(\boldsymbol{c})\boldsymbol{u} = \sup_{\boldsymbol{u}\in\mathbb{R}^{K+1}} \boldsymbol{u}'\boldsymbol{\Sigma}\boldsymbol{u} = \boldsymbol{u}_k'\boldsymbol{\Sigma}\boldsymbol{u}_k,$$

where

$$\boldsymbol{u}_i'\boldsymbol{u}_j = \delta_{ij},\ \ i,j = 1,2,\ldots,k.$$

The expression $(\gamma_k, \boldsymbol{u}_k)$ will be called the $k$th principal system of vector $\boldsymbol{c}$.

**Theorem 1** [Górecki, Krzyśko, 2012]**.** *The $k$th principal system $(\lambda_k, u_k(t))$ of the stochastic process $X(t)$ is related to the $k$th principal system $(\gamma_k, \boldsymbol{u}_k)$ of the random vector $\boldsymbol{c}$ by the equations:*

$$\lambda_k = \gamma_k,$$
$$u_k(t) = \boldsymbol{u}_k'\boldsymbol{\varphi}(t),$$

*where $t \in I$ and $k = 1,2,\ldots,K+1$.*

Principal components analysis for random vectors $\boldsymbol{c}$ is based on the matrix $\boldsymbol{\Sigma}$. I practice this matrix is unknown. We estimate it on the basis of $N$ independent realizations $x_1(t), x_2(t), \ldots, x_N(t)$ of the form $x_i(t) = \hat{\boldsymbol{c}}_i'\boldsymbol{\varphi}(t)$ of the random process $X(t)$, where the vectors $\hat{\boldsymbol{c}}_i'$ are centered, $i = 1,2,\ldots,N$. Let $\hat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_1, \hat{\boldsymbol{c}}_2 \ldots, \hat{\boldsymbol{c}}_N)'$. Then

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\hat{\boldsymbol{C}}'\hat{\boldsymbol{C}}.$$

If $N > K+1$, then the matrix $\hat{\boldsymbol{\Sigma}}$ is positive definite with probability 1.

Let $\hat{\gamma}_1 \geq \hat{\gamma}_2 \geq \cdots \geq \hat{\gamma}_s$ be non-zero eigenvalues of the matrix $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{u}}_2, \ldots, \hat{\boldsymbol{u}}_s$ the corresponding eigenvectors, where $s = \mathrm{rank}(\hat{\boldsymbol{\Sigma}})$.

Moreover the $k$th principal system of the random process $X(t)$ determined from a sample has the following form:

$$(\hat{\lambda}_k = \hat{\gamma}_k, \hat{u}_k(t) = \hat{\boldsymbol{u}}_k'\boldsymbol{\varphi}(t)),$$

where $t \in I$ and $k = 1, 2, \ldots, s$. Hence the coordinates of the projection of the $i$th realization $x_i(t)$ of the process $X(t)$ on the $k$th functional principal component are equal to:

$$U_{ik} =< \hat{u}_k(t), x_i(t) >=< \hat{\boldsymbol{u}}_k'\boldsymbol{\varphi}(t), \boldsymbol{c}_i'\boldsymbol{\varphi}(t) >= \hat{\boldsymbol{u}}_k' < \boldsymbol{\varphi}(t), \boldsymbol{\varphi}(t) > \boldsymbol{c}_i = \hat{\boldsymbol{u}}_k'\boldsymbol{c}_i,$$

where $i = 1, 2 \ldots, N,\ k = 1, 2, \ldots, s$.

## 4. Construction of functional discriminant coordinates

Suppose that we observe a sample of the stochastic process $X(t) \in L_2(I)$. Moreover, suppose that

$$E(X(t)) = 0$$

and

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

Let us consider the case where the stochastic process $X(t)$ can be represented by a finite number of orthonormal basis functions, i.e., $X(t)$ can be represented as:

$$X(t) = \sum_{k=0}^{K} c_k\varphi_k(t),\ t \in I, \tag{4}$$

where $\{\varphi_k\}$ are the first $K + 1$ 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

$$\boldsymbol{\varphi}(t) = (\varphi_0(t), \varphi_1(t), , \ldots, \varphi_K(t))'$$
$$\boldsymbol{c} = (c_0, c_1, \ldots, c_K)',$$

where $E(\boldsymbol{c}) = \boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{c}) = \boldsymbol{\Sigma} > 0$. The process $X(t)$ can be written in vector form as

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

The aim of linear discriminant analysis for functional data is to find linear functionals (discriminant coordinates):

$$U(X) =< u, X >= \int u(t)X(t)dt,\ u \in L_2(I)$$

such that the between-group variance is maximized with respect to the total variance. The function $u(t)$ is called the weighting function.

We construct the first functional discriminant coordinate $U_1 =< u_1, X >$, where the weight function $u_1(t)$ is defined by:

$$\lambda_1 = \sup_{u \in L_2(I)} \frac{\mathrm{Var}_B(< u, X >)}{\mathrm{Var}_T(< u, X >)} = \frac{\mathrm{Var}_B(< u_1, X >)}{\mathrm{Var}_T(< u_1, X >)}$$

subject to the constraint that

$$\mathrm{Var}_T(< u_1, X >) = 1, \tag{5}$$

where $\mathrm{Var}_B(< u, X >)$ and $\mathrm{Var}_T(< u, X >)$ are respectively the between-group variance and the total variance of the inner product $< u, X >$. The condition (5) is imposed to ensure the uniqueness of the discriminant coordinate (without sign).

Proceeding analogously, we can construct the $k$th functional discriminant coordinate $U_k =< u_k, X >$, where the weight function $u_k(t)$ is defined by

$$\lambda_k = \sup_{u \in L_2(I)} \frac{\mathrm{Var}_B(< u, X >)}{\mathrm{Var}_T(< u, X >)} = \frac{\mathrm{Var}_B(< u_k, X >)}{\mathrm{Var}_T(< u_k, X >)}$$

subject to the constraint that

$$\mathrm{Var}_T(< u_k, X >) = 1,$$

and the $k$th functional discriminant coordinate $U_k =< u_k, X >$ is not correlated with the first $k-1$ functional discriminant coordinates $\{U_i =< u_i, X >,\ i = 1, 2, \ldots, k-1\}$. We shall call $(\lambda_k, u_k)$ the $k$th discriminant configuration of X.

Let

$$\rho_k = \sup_{u \in \mathbb{R}^{K+1}} \frac{\boldsymbol{u}' \,\mathrm{Var}_B(\boldsymbol{c})\boldsymbol{u}}{\boldsymbol{u}' \,\mathrm{Var}_T(\boldsymbol{c})\boldsymbol{u}} = \frac{\boldsymbol{u}_k' \,\mathrm{Var}_B(\boldsymbol{c})\boldsymbol{u}_k}{\boldsymbol{u}_k' \,\mathrm{Var}_T(\boldsymbol{c})\boldsymbol{u}_k}$$

subject to the constraint $\boldsymbol{u}_i' \,\mathrm{Var}_T(\boldsymbol{c})\boldsymbol{u}_j = \delta_{ij}$ (Kronecker delta function), $i, j = 1, 2, \ldots, k$. We shall call $(\rho_k, \boldsymbol{u}_k)$ the $k$th discriminant configuration of a random vector $\boldsymbol{c} = (c_0, c_1, \ldots, c_k)'$.

**Theorem 2** [Górecki, Krzyśko, Waszak, 2014]**.** *The kth discriminant configuration of a random vector* $\boldsymbol{c}$ *defined by* $(\rho_k, \boldsymbol{u}_k)$ *is related to the kth discriminant configuration of the stochastic process* $X(t)$*,* $(\lambda_k, u_k(t))$*, as follows:*

$$\lambda_k = \rho_k,$$
$$u_k(t) = \boldsymbol{u}_k' \boldsymbol{\varphi}(t).$$

Discriminant coordinate analysis for a random process $X(t)$ with finite basis expansion (4) is therefore equivalent to multivariate discriminant coordinate analysis of a random vector $c = (c_0, c_1, \ldots, c_K)'$. The $N = N_1 + N_2 + \cdots + N_L$ independent realizations of a random vector $\boldsymbol{c}$ can be compiled in a matrix $L$ with size $N_l \times N$, of the form:

$$\hat{\boldsymbol{C}}_l = \begin{bmatrix} \hat{c}_{l10} & \hat{c}_{l11} & \ldots & \hat{c}_{l1K} \\ \hat{c}_{l20} & \hat{c}_{l21} & \ldots & \hat{c}_{l2K} \\ \ldots & \ldots & \ldots & \ldots \\ \hat{c}_{lN_l0} & \hat{c}_{lN_l1} & \ldots & \hat{c}_{lN_lK} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{c}}'_{l1} \\ \hat{\boldsymbol{c}}'_{l2} \\ \ldots \\ \hat{\boldsymbol{c}}'_{lN_l} \end{bmatrix},$$

where the $\hat{c}_{lrs}$ are the coefficients estimated from the data by the least squares method.

The vector of means

$$\bar{\boldsymbol{c}}_l = \frac{1}{N_l} \sum_{i=1}^{N_l} \hat{\boldsymbol{c}}_{li}, l = 1, 2, \ldots, L,$$

the between-group sum of squares matrix

$$\widehat{\mathrm{Var}}_B(\boldsymbol{c}) = \hat{\boldsymbol{B}} = \frac{1}{L} \sum_{l=1}^{L} N_l \bar{\boldsymbol{c}}_l \bar{\boldsymbol{c}}'_l,$$

and the total sum of squares matrix

$$\widehat{\mathrm{Var}}_T(\boldsymbol{c}) = \hat{\boldsymbol{T}} = \sum_{l=1}^{L} \sum_{i=1}^{N_l} \hat{\boldsymbol{c}}_{li} \hat{\boldsymbol{c}}'_{li},$$

are calculated based on the whole sample of $N = N_1 + N_2 + \cdots + N_L$ elements.

Next we find the nonzero eigenvalues $\hat{\lambda}_k$ and corresponding eigenvectors $\hat{\boldsymbol{u}}_k$ of the matrix $\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{B}}$. The number of nonzero eigenvalues of this matrix is equal to $\min(K + 1, L - 1)$. Having determined the eigenvectors $\hat{\boldsymbol{u}}_k$, we determine its eigenfunctions (weight functions):

$$\hat{u}_k(t) = \hat{\boldsymbol{u}}'_k \boldsymbol{\varphi}(t), \ t \in I.$$

Hence the coordinate projection of the $i$th realization $x_{li}(t)$ coming from the $l$th group of the process on the $k$th functional discriminant coordinate is equal to

$$\hat{U}_{lik} = < \hat{u}_k, x_{li} > = \int \hat{u}_k(t) x_{li}(t) dt = \sum_{j=0}^{K} \hat{u}_{kj} \hat{c}_{lij} = \hat{\boldsymbol{c}}'_{li} \hat{\boldsymbol{u}}_k,$$

where

$$\hat{\boldsymbol{c}}_{li} = (\hat{c}_{li0}, \hat{c}_{li1}, \ldots, \hat{c}_{liK})',$$
$$\hat{\boldsymbol{u}}_k = (\hat{u}_{k0}, \hat{u}_{k1}, , \ldots, \hat{u}_{kK}, )',$$

$$i = 1, 2, \ldots, N_l,\ l = 1, 2, \ldots, L,\ k = 1, 2, \ldots, \min(K+1, L-1).$$

## 5. A new discriminant space

Consider the $(L-1)$-dimensional space of the first functional discriminant coordinates. The training sample $\{x_1(t), x_2(t), , x_N(t)\}$ transformed into this space will be denoted by $\{Y_1, Y_2, \ldots, Y_N\}$, where $Y_i \in \mathbb{R}^{L-1}$. We then further consider the $(L-1)$-dimensional space of the first functional principal components. The training sample transformed into this space will be denoted by $\{Z_1, Z_2, \ldots, Z_N\}$, where $Z_i \in \mathbb{R}^{L-1}$.

We will create a new $2(L-1)$-dimensional space determined by the first $(L-1)$ functional discriminant coordinates and the first $(L-1)$ functional principal components. The directional vectors determining this new space are normed so that their length is equal to 1. The new space combines the advantages of the space of functional discriminant coordinates and the space of functional principal components. In this space it is possible to apply a variety of classification algorithms obtaining an improvement in classification quality.

## 6. Example

### 6.1. Experimental setup

The quality of performance of the described method was tested on the 35 different data sets described in Table 1. The data sets originate from the UCR Time Series Classification/Clustering Homepage (Keogh et al., 2011). The data sets originate from a plethora of different domains, including medicine, robotics, astronomy, biology, face recognition, handwriting recognition, etc.

The proposed method is used with the LDA classifier in the classification process. In this case the decision was motivated by the simplicity and efficiency of the LDA method. LDA has been proved to be effective (Lim et al., 2000) although relying on strong assumptions. Additionally LDA is still widely used in practice, e.g. in face recognition (Song et al., 2007), medicine (Kwak et al., 2002), chemometrics (Cozzolino et al., 2002) and many other areas.

Table 1. Summary of data sets.

| Data set | Number of classes | Size of training set | Size of testing set | Time series length |
|---|---|---|---|---|
| 50 Words | 50 | 450 | 455 | 270 |
| Beef | 5 | 30 | 30 | 470 |
| Car | 4 | 60 | 60 | 577 |
| CBF | 3 | 30 | 900 | 128 |
| CinC ECG Torso | 4 | 40 | 1380 | 1639 |
| Coffee | 2 | 28 | 28 | 286 |
| Cricket X | 12 | 390 | 390 | 300 |
| Cricket Y | 12 | 390 | 390 | 300 |
| Cricket Z | 12 | 390 | 390 | 300 |
| ECG Five Days | 2 | 23 | 861 | 136 |
| Face All | 14 | 560 | 1690 | 131 |
| Face Four | 4 | 24 | 88 | 350 |
| Faces UCR | 14 | 200 | 2050 | 131 |
| Gun Point | 2 | 50 | 150 | 150 |
| Haptics | 5 | 155 | 308 | 1092 |
| Inline Skate | 7 | 100 | 550 | 1882 |
| Italy Power Demand | 2 | 67 | 1029 | 24 |
| Lightning 2 | 2 | 60 | 61 | 637 |
| Lightning 7 | 7 | 70 | 73 | 319 |
| Mote Strain | 2 | 20 | 1252 | 84 |
| Non Invasive Thorax 1 | 42 | 1800 | 1965 | 750 |
| OSU Leaf | 6 | 200 | 242 | 427 |
| Sony AIBO Robot Surface | 2 | 20 | 601 | 70 |
| Star Light Curves | 3 | 1000 | 8236 | 1024 |
| Swedish Leaf | 15 | 500 | 625 | 128 |
| Symbols | 6 | 25 | 995 | 398 |
| Synthetic Control | 6 | 300 | 300 | 60 |
| Two Patterns | 4 | 1000 | 4000 | 128 |
| Two Lead ECG | 2 | 23 | 1139 | 82 |
| u Wave Gesture Library X | 8 | 896 | 3582 | 315 |
| u Wave Gesture Library Y | 8 | 896 | 3582 | 315 |
| u Wave Gesture Library Z | 8 | 896 | 3582 | 315 |
| Wafer | 2 | 1000 | 6174 | 152 |
| Words Synonyms | 25 | 267 | 638 | 270 |
| Yoga | 2 | 300 | 3000 | 426 |

For each data set we calculated the classification error rate on a test subset. The optimum value for $K$ and LDA parameters we found using only the training subset. An appropriate distribution of the training and test sets was proposed by the authors of the repository (each data set is divided into a training and testing subset). For each set separately, the discrete time series were centered, and then transformed into functions. As a base we used the Fourier orthonormal system in the space $L_2([0,T])$:

$$\varphi_0(x) = \frac{1}{\sqrt{T}}, \varphi_{2k-1}(x) = \frac{\sqrt{2}}{\sqrt{T}} \sin \frac{2\pi kx}{T}, \varphi_{2k}(x) = \frac{\sqrt{2}}{\sqrt{T}} \cos \frac{2\pi kx}{T}, \ k = 1, 2 \dots$$

### 6.2. Main results

The results are presented in Table 2. In the column $K$ we have the optimum value for $K$ parameter (the size of the base). In the next three columns we have absolute testing error rates for all spaces for LDA classifier.

Table 2. Testing error rates (in %).

| Data set | $K$ | FPC | FDC | FPC+FDC |
|---|---|---|---|---|
| 50 Words | 99 | 74.73 | 69.45 | 69.45 |
| Beef | 73 | 56.67 | 50.00 | 23.33 |
| Car | 99 | 68.33 | 68.33 | 68.33 |
| CBF | 11 | 57.11 | 35.44 | 35.44 |
| CinC ECG Torso | 99 | 71.59 | 58.62 | 56.74 |
| Coffee | 53 | 46.63 | 7.14 | 0.00 |
| Cricket X | 51 | 63.33 | 58.21 | 58.21 |
| Cricket Y | 55 | 80.51 | 60.00 | 60.00 |
| Cricket Z | 51 | 77.95 | 61.54 | 61.54 |
| ECG Five Days | 85 | 50.29 | 49.71 | 48.78 |
| Face All | 61 | 53.02 | 44.26 | 44.26 |
| Face Four | 67 | 40.91 | 23.86 | 20.45 |
| Faces UCR | 59 | 45.46 | 25.56 | 25.56 |
| Gun Point | 47 | 54.67 | 49.33 | 49.33 |
| Haptics | 17 | 78.90 | 75.00 | 75.00 |
| Inline Skate | 99 | 86.36 | 85.64 | 82.36 |
| Italy Power Demand | 9 | 45.97 | 42.76 | 42.76 |
| Lighting 2 | 99 | 45.90 | 39.34 | 34.43 |
| Lighting 7 | 25 | 57.53 | 54.79 | 54.79 |
| Mote Strain | 83 | 25.00 | 25.96 | 20.37 |
| Non Invasive Thorax 1 | 99 | 94.76 | 93.49 | 93.49 |
| Osu Leaf | 99 | 78.93 | 73.14 | 73.14 |
| Sony AIBO Robot Surface | 69 | 47.92 | 45.42 | 35.11 |
| Star Light Curves | 99 | 21.87 | 17.97 | 17.97 |
| Swedish Leaf | 45 | 88.80 | 87.20 | 87.20 |
| Symbols | 47 | 81.71 | 29.75 | 29.55 |
| Synthetic Control | 59 | 27.00 | 8.67 | 8.67 |
| Two Patterns | 17 | 60.40 | 12.38 | 12.38 |
| Two Lead ECG | 81 | 46.44 | 36.17 | 36.17 |
| u Wave Gesture Library X | 41 | 49.94 | 37.13 | 37.13 |
| u Wave Gesture Library Y | 39 | 50.50 | 42.96 | 42.96 |
| u Wave Gesture Library Z | 39 | 55.33 | 45.78 | 45.78 |
| Wafer | 41 | 10.79 | 5.34 | 5.34 |
| Words Synonyms | 99 | 77.27 | 76.02 | 76.02 |
| Yoga | 99 | 46.43 | 38.33 | 38.33 |

FPC+FDC method was the best on 10 data sets. On 25 data sets no method was clearly better than the others, but FPC+FDC has always been among the best methods. We can see that FPC method is by far the worst, only on one data set (*Car*) it achieves the same result as the other methods, and on one data set (*Mote Strain*) it achieves better result than FDC method. A graphical comparison of methods is presented in Figure 1.

We see that the method FPC+FDC is clearly superior to FDC method.

### 6.3. Statistical comparison of examined methods

To find differences between the methods we present a detailed statistical comparison. We test the null hypothesis that all classifiers perform the same and the observed differences are merely random. We used the Iman, Davenport (1980) test, which is a nonparametric, based on ranks, equivalent of ANOVA. Due to the fact that in this test the $p$-value is equal to 0, we can proceed with the post hoc tests in order to detect significant pairwise differences among all the classifiers. When comparing multiple algorithms, to retain an overall significance level $\alpha$, one has to adjust the value of $\alpha$ for each post hoc comparison. There are

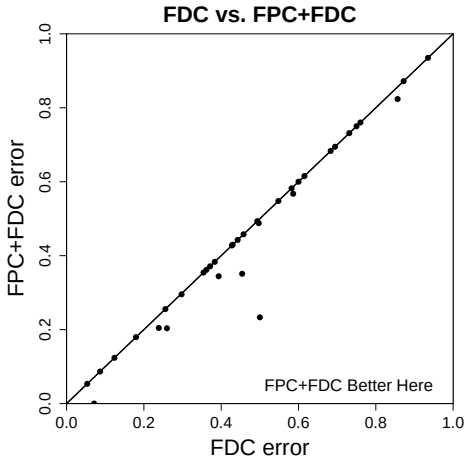


Figure 1. Comparison of test errors.

various methods for this. Garcia, Herrera (2008) explain and compare the use of various correction algorithms. They showed that although it requires intensive computation, the Bergmann, Hommel (1988) dynamic procedure has the highest power. The results of multiple comparisons are given in Table 3 and Table 4. Those methods that are connected by a sequence of stars have average ranks that are not significantly different from one another. Finally we obtained two homogeneous disjoint groups of methods: {FPC+FDC, FDC} and {FPC}. The best methods are in the first group.

Table 3. $p$-values in the Bergmann-Hommel post hoc test.

| i | Hypothesis | $p$-value |
|---|---|---|
| 1 | FPC vs. FPC+FDC | $1.5 \times 10^{-10}$ |
| 2 | FPC vs. FDC | $1.4 \times 10^{-7}$ |
| 3 | FDC vs. FPC+FDC | 0.189 |

Table 4. Results of the Bergmann-Hommel post hoc test.

| Procedure | Ranks mean | | |
|---|---|---|---|
| FPC+FDC | 1.37 | * | |
| FDC | 1.69 | * | |
| FPC | 2.94 | | * |

## 7. Conclusions

In this paper we have introduced and studied a new functional spaces convenient for classification. We used these spaces to classify data with the LDA method. Due to the high degree of nonlinearity, the method does not easily admit a rigorous theoretical analysis. However, the experiments that we have conducted provide evidence of the power and usefulness of the proposed spaces. The new methods adapt well to different data sets without showing signs of an overfitting. The experiments that we have conducted justify the power and usefulness of our methods, especially FPC+FDC. We recommend for use in the classification process the FPC+FDC space.

## References

[1] G. Bergmann and G. Hommel, *Improvements of general multiple test procedures for redundant systems of hypotheses*, in: Multiple Hypotheses Testing, P. Bauer, G. Hommel, E. Sonnemann (Ed.), Springer (1988) 110–115. doi:10.1007/978-3-642-52307-6_8

[2] D. Cozzolino, E. Restaino and A. Fassio, *Discrimination of yerba mate (ilex paraguayensis st. hil.) samples according to their geographical origin by means of near infrared spectroscopy and multivariate analysis*, Sensing and Instrumentation for Food Quality and Safety **4** (2002) 67–72. doi:10.1007/s11694-010-9096-y

[3] R.A. Fisher, *The use of multiple measurements in taxonomic problem*, Annals of Eugenics **7** (1936) 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

[4] S. Garcia and F. Herrera, *An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons*, Journal of Machine Learning Research **9** (2008) 2677–2694.

[5] T. Górecki and M. Krzyśko, Functional Principal Components Analysis, in: Data analysis methods and its applications, J. Pociecha, R. Decker (Ed.), C.H. Beck (2012) 71–87.

[6] T. Górecki, M. Krzyśko and Ł. Waszak, Functional discriminant coordinates, Communication in Statistics - Theory and Methods **43** (5) (2014) 1013–1025. doi:10.1080/03610926.2013.828074

[7] H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology **24** (1933) 417–441, 498–520. doi:10.1037/h0071325

[8] R. Iman and J. Davenport, *Approximations of the critical region of the freidman statistic*, Communications in Statistics - Theory and Methods **9** (6) (1980) 571–595. doi:10.1080/03610928008827904

[9] E. Keogh, Q. Zhu, B. Hu, Y. Hao, X. Xi, L. Wei and C.A. Ratanamahatana, The UCR Time Series Classification/Clustering, Homepage, 2011. http://www.cs.ucr.edu/~eamonn/time_series_data/.

[10] N.K. Kwak, S.H. Kim, C.W. Lee and T.S. Choi, *An application of linear programming discriminant analysis to classifying and predicting the symptomatic status of hiv/aids patients*, Journal of Medical Systems **26** (5) (2002) 427–438.

[11] T.S. Lim, W.Y. Loh and Y.S. Shih, *A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms*, Machine Learning **40** (3) (2000) 203–228. doi:10.1023/A:1007608224229

[12] J.O. Ramsay and B.W. Silverman, Functional Data Analysis, Second Edition (Springer, 2005). doi:10.1007/b98888

[13] G.A.F. Seber, Multivariate Observations (Wiley, 1984). doi:10.1002/9780470316641

[14] G. Shmueli, *To explain or to predict?* Statistical Science **25** (3) (2010) 289–310. doi:10.1214/10-STS330

[15] F. Song, D. Zhang, Q. Chen and J. Wang, *Face recognition based on a novel linear discriminant criterion*, Pattern Analysis and Applications **10** (2007) 165–174. doi:10.1007/s10044-006-0057-3

Received 10 September 2014
Revised 16 September 2014