Functional data analysis for Statfda

Manuel Escabias and Ana María Aguilera

Functional data analysis

A functional variable is one whose values depend on a continuous magnitude such as time. They are functional in the sense that they are evaluated at any time in the domain, instead of the discrete way, in which they were originally measured or observed (Ramsay and Silverman, 2005). Thus, a functional data set is a set of curves $\{x_1(t), \ldots, x_n(t)\}$, with $t \in T$. Each curve can be observed at different time points of his argument t as $x_i = (x_i(t_0), \ldots, x_i(t_{m_i}))'$ for the set of times $t_0, \ldots, t_{m_i},\ i = 1, \ldots, n$ and these are not necessarily the same for each curve.

Different approaches have been taken to the study of functional data, including the nonparametric methods proposed by Müller (2008) and Ferraty and Vieu (2006) and the basis expansion methods used by Ramsay and Silverman (2005). The latter method is adopted in the present application, in which we seek to reconstruct the functional form of curves in order to evaluate them at any time point t. This method assumes that the curves belong to a finite dimensional space generated a basis of functions $\{\phi_1(t), \ldots, \phi_p(t)\}$ and so they can be expressed as

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t),\ i = 1, \ldots, n. \tag{1}$$

The functional form of the curves is determined when the basis coefficients $a_i = (a_{i1}, \ldots, a_{ip})'$ are known. These can be obtained from the discrete observations either by least squares or by interpolation (see, for example, Escabias et al., 2005 and 2007). In our application the least squares method is considered for functional representation.

Depending on the characteristics of the curves and the observations, various classes of basis can be used (see, for example, Ramsay and Silverman, 2005). In practice, those most commonly used are, on the one hand, the basis of trigonometric functions for regular, periodic, continuous and differentiable curves, and on the other, the basis of B-spline functions, which provides a better local behavior (see De Boor, 2001).

Exploratory analysis: curves representation

Let $x_1(t), x_2(t), \ldots, x_n(t)$ be a set of curves all of them observed at the same time points $t_1, t_2, \ldots, t_m$, then the available information in this situation is the matrix

$$X = \begin{pmatrix} x_1(t_1) & x_1(t_2) & \cdots & x_1(t_m) \\ x_2(t_1) & x_2(t_2) & \cdots & x_2(t_m) \\ \cdots & \cdots & \cdots & \cdots \\ x_n(t_1) & x_n(t_2) & \cdots & x_n(t_m) \end{pmatrix}$$

This matrix is the one of discretized curves in Statfda in .csv format:

```
-4.65,-5.33,-2.53,1.26,5.79,10.79,15.21,15.28,11.62,7.02,2.95,-1.85
-6.16,-6.18,-1.74,3.62,9.44,14.78,18.38,18.20,13.87,8.49,3.24,-2.99
-5.72,-6.80,-2.94,1.85,7.50,13.14,17.49,17.64,13.31,8.27,3.53,-2.03
-3.22,-3.49,-0.15,4.69,9.34,13.40,16.29,16.60,13.59,9.25,4.90,-0.46
```

```
-8.11,-8.26,-3.44,2.32,8.76,14.44,18.29,17.98,13.44,8.03,2.57,-4.15
-15.36,-13.23,-5.82,2.95,10.20,16.00,18.66,17.06,11.95,5.96,-1.16,-11.04
-12.78,-11.28,-4.67,3.29,10.91,16.38,19.05,17.65,12.50,6.46,-0.54,-8.98
-11.82,-10.65,-4.04,3.93,10.79,15.49,17.97,16.69,11.96,6.38,0.00,-8.22
```

And the sampling points $t_1, t_2, \ldots, t_m$,

```
1,2,3,4,5,6,7,8,9,10,11,12
```

The basis coefficients of all curves

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix},$$

are obtained as

$$A^T = \left(\Phi^T \Phi\right)^{-1} \Phi^T X^T$$

where

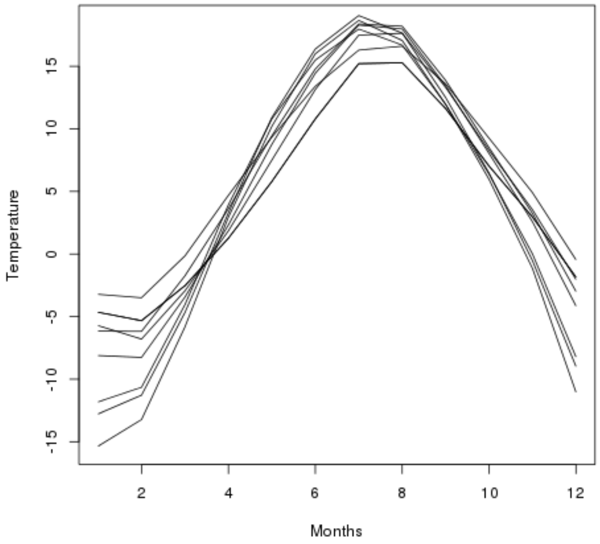
$$\Phi = \begin{pmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_p(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_p(t_2) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_1(t_m) & \phi_2(t_m) & \cdots & \phi_p(t_m) \end{pmatrix}$$

is the matrix of basic functions evaluated at sampling points.

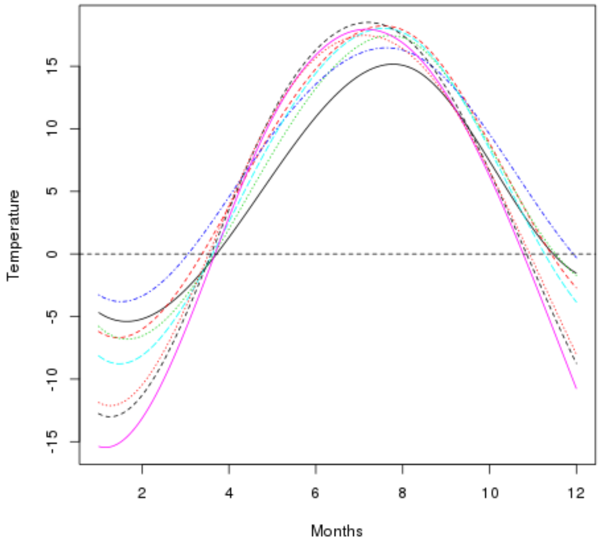
Basis coefficients of all curves are provided by Statfda and can be downloaded in .csv format

```
 ""      ,"bspl4.1","bspl4.2","bspl4.3","bspl4.4","bspl4.5","bspl4.6"
"Curve 1"    ,-4.67,    -7.48,     4.40,    23.45,     0.80,    -1.54
"Curve 2",    -6.19,    -8.96,    10.03,    25.78,     2.00,    -2.71
"Curve 3",    -5.77,    -9.79,     7.12,    25.93,     1.55,    -1.71
"Curve 4",    -3.26,    -5.97,     9.82,    22.32,     4.63,    -0.30
"Curve 5",    -8.13,   -11.51,     9.73,    25.72,     1.41,    -3.85
"Curve 6",   -15.37,   -16.80,    15.25,    22.75,     0.09,   -10.77
"Curve 7",   -12.75,   -15.14,    14.93,    23.96,    -0.16,    -8.76
"Curve 8",   -11.87,   -14.25,    15.26,    21.71,     0.85,    -8.02
```

and the plots of curves we get:



Raw data



Smoothed curves

Exploratory analysis: mean and standard deviation

From a set of curves $x_1(t), x_2(t), \ldots, x_n(t)$ the mean curve is defined as

$$\overline{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$$

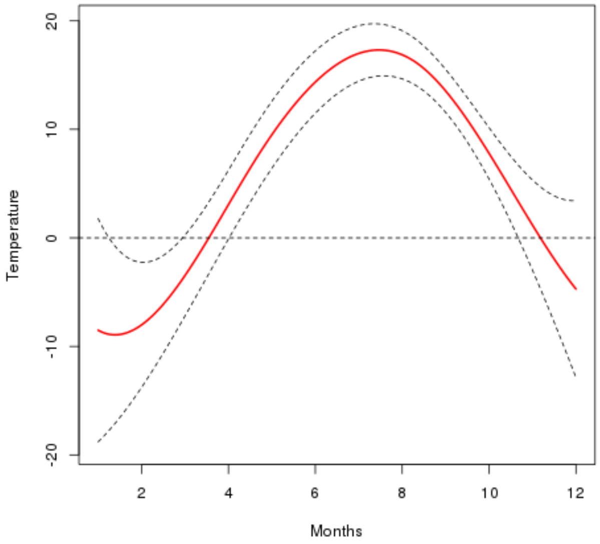
The mean curve can be expressed in terms of basic functions

$$\overline{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}\phi_j(t) = \sum_{j=1}^{p}\overline{a}_j\phi_j(t), \ \overline{a}_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$$

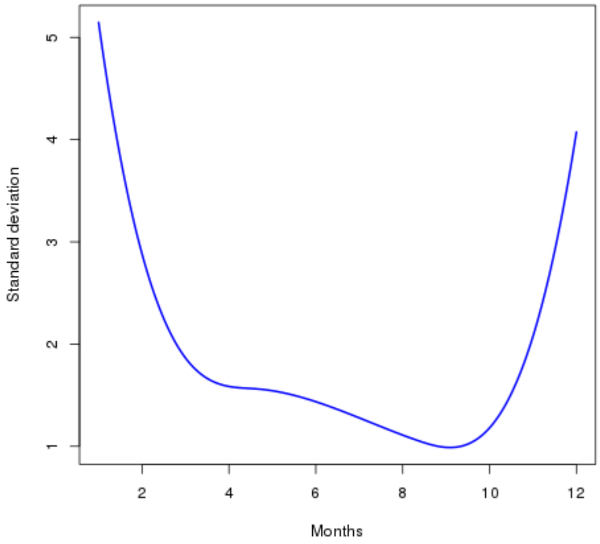
So the mean function is also defined throught its basis coefficients. These basis coeefficients are also provided by Statfda and can be downloaded in .csv format.

```
""    ,"bspl4.1","bspl4.2","bspl4.3","bspl4.4","bspl4.5","bspl4.6"
"mean",    -8.50,   -11.24,    10.82,    23.95,     1.40,   -4.71
```

and the plots of mean curve and standard deviation we get



Mean curve



Standard deviation curve

Exploratory analysis: bivariate correlation function

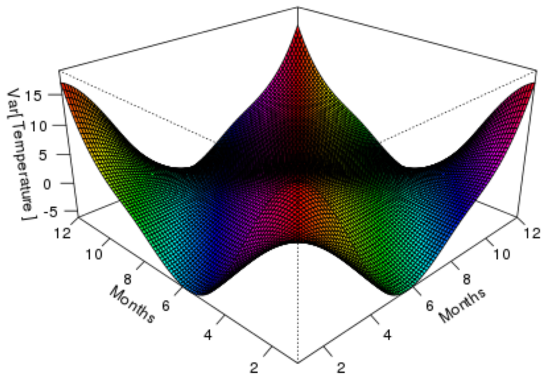
From a set of curves $x_1\left(t\right), x_2\left(t\right), \ldots, x_n\left(t\right)$ with mean curve $\overline{x}(t)$ the covariance surface is defined

$$C\left(s,t\right)=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i\left(s\right)-\overline{x}\left(s\right)\right)\left(x_i\left(t\right)-\overline{x}\left(t\right)\right)$$

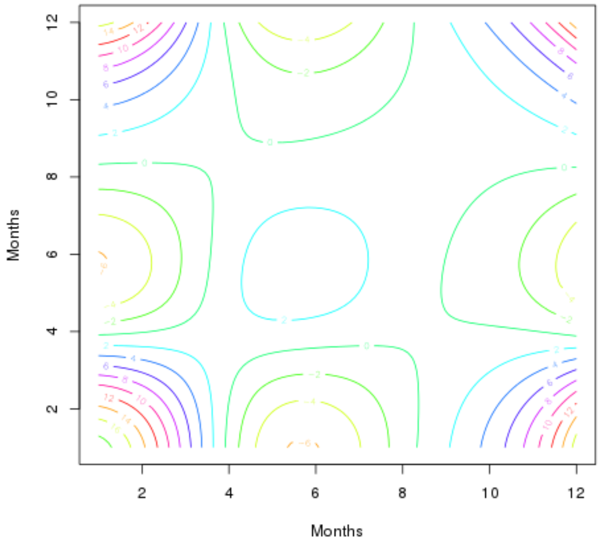
and from it the correlation surface

$$r\left(s,t\right)=\frac{C(s,t)1}{\sqrt{C(s,s)C(t,t)}}$$

the surface and contour plots for bivariate correlation are provided by Statfda



Surface



Contour plot

Functional principal component analysis

Let $x_1(t), \ldots, x_n(t)$ be a set of curves with mean curve

$$\overline{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t)$$

and covariance surface

$$C(s,t) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i(s) - \overline{x}(s))(x_i(t) - \overline{x}(t))$$

Functional principal components are defined as the vectors whose elements are obtained with the linear conbination of the sample curves

$$\xi_i = \int_T (x_i(t) - \overline{x}(t)) f(t)\, dt,\ i = 1, \ldots, n$$

that maximize the variance of $\xi_1, \ldots, \xi_n$ and uncorrelated. By imposing this condition, functional principal components are the solutions of this equation

$$\int_T C(s,t) f(s)\, ds = \lambda f(t)$$

being λ the variance of functional principal components.

When curves are expresed in terms of basic functions as

$$x_i(t) = \sum_{j=1}^{p} a_{ij} \phi_j(t),\ i = 1, \ldots, n.$$

previous equation has p solutions for p values of λ that verify that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Each one of the λ_j is associated to a $f_j(t)$ function that define the funcional principal component and known as eigenfuncions or principal component curves.

$$\xi_{ij} = \int_T (x_i(t) - \overline{x}(t)) f_j(t)\, dt,\ j = 1, \ldots, p,\ i = 1, \ldots, n$$

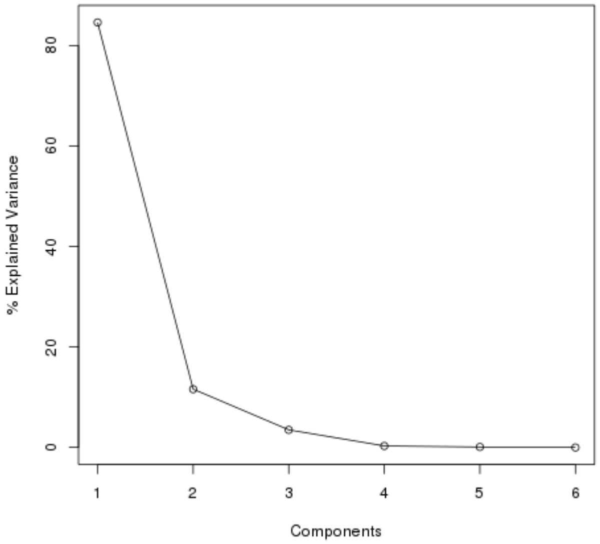
So each functional principal component ξ_j is a vector of dimension n. In matrix form, functional principal components are usually considered in a matrix Γ of dimension $n \times p$. Moreover, each functional principal component cummulates a proportion of the total variability given by

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

The total variability that cummulates the functional principal components is equal to the total variability of curves.

Functional principal component analysis: explained variances

This section of the application provides a graph and a table of the variability cummulation:



```
 [1] "Percentages of explained variances"
    Variance % Exp. Var % Cum. Exp. Var
Comp.1 34.43       84.6            84.6
Comp.2  4.70       11.6            96.2
Comp.3  1.41        3.5            99.7
Comp.4  0.12        0.3           100.0
Comp.5  0.03        0.1           100.1
Comp.6  0.00        0.0           100.1
```

Functional principal component analysis: principal component curves

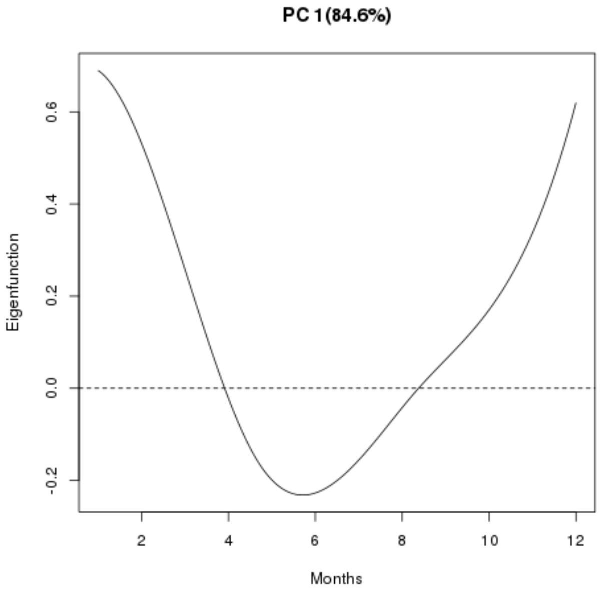
When curves are expresed in terms of basic functions as

$$x_i\left(t\right)=\sum_{j=1}^{p}a_{ij}\phi_j\left(t\right),\ i=1,\ldots,n.$$

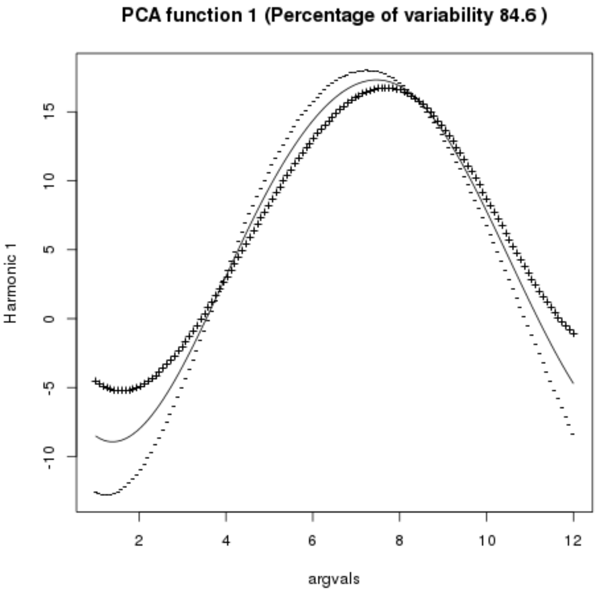
eigenfunctions are also expresed in terms of the same basic functions.

$$f_j\left(t\right)=\sum_{k=1}^{p}F_{jk}\phi_k\left(t\right),\ j=1,\ldots,p.$$

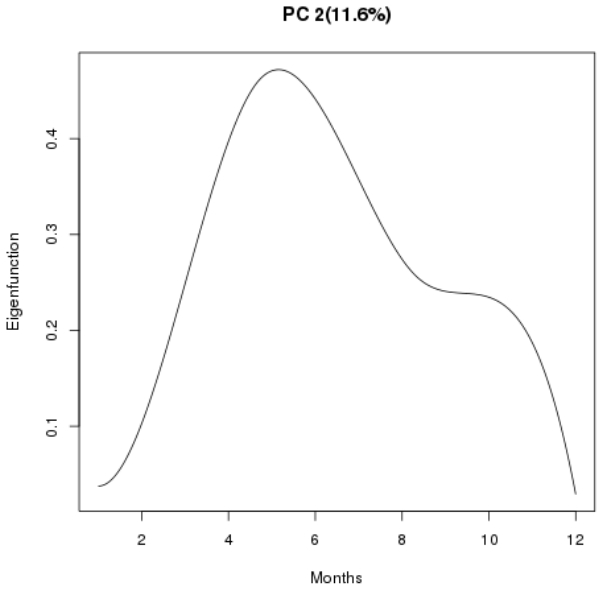
The application provides a plot of principal component curves and how the mean curve is affected by these principal component curves



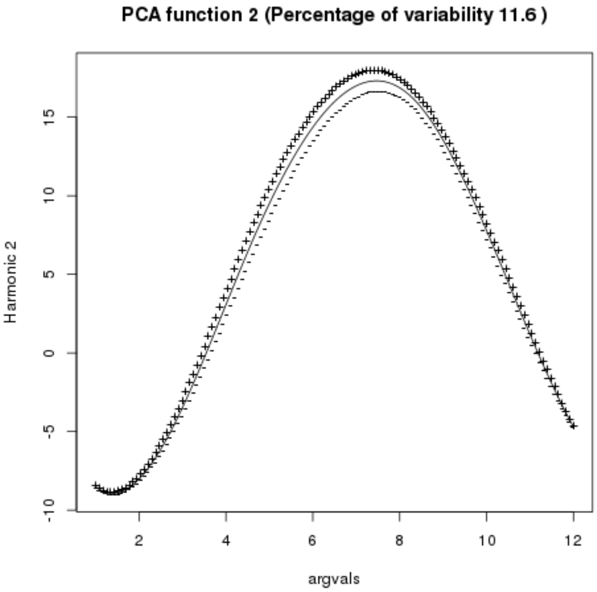
1st PC Curve



Perturbation of mean



2nd PC Curve



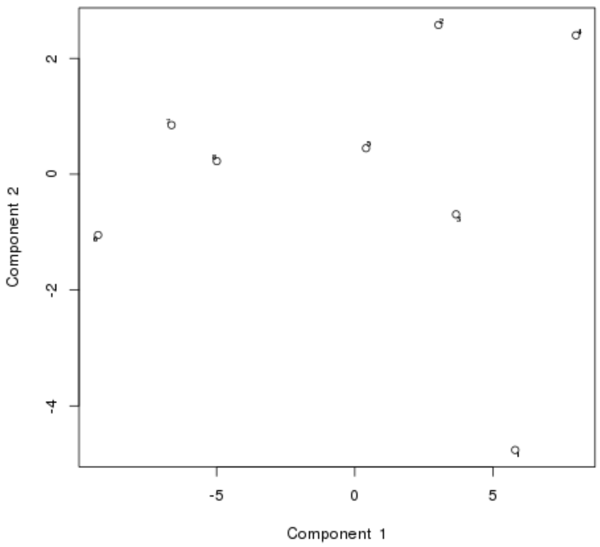
Perturbation of mean

The application also provides the basis coefficients that define all principal component curves

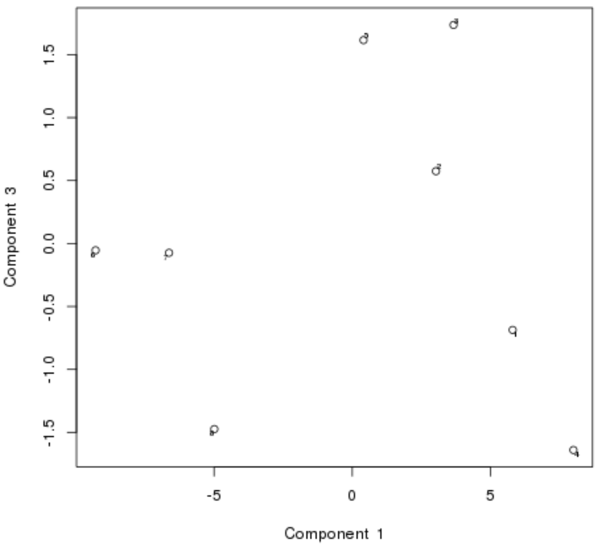
```
""       , "PC1","PC2","PC3","PC4","PC5","PC6"
"bspl4.1",  0.69, 0.04, 0.10,-0.48, 0.85,-1.22
"bspl4.2",  0.61, 0.04,-0.25,-0.45,-1.29, 1.16
"bspl4.3", -0.55, 0.75,-0.86, 0.71, 1.00,-0.95
"bspl4.4",  0.07, 0.09, 1.21,-1.17,-0.45, 0.97
"bspl4.5",  0.18, 0.32,-0.29, 1.36,-0.48,-1.17
"bspl4.6",  0.62, 0.03, 0.34,-0.03, 0.96, 1.22
```

Functional principal component analysis: biplots and scores

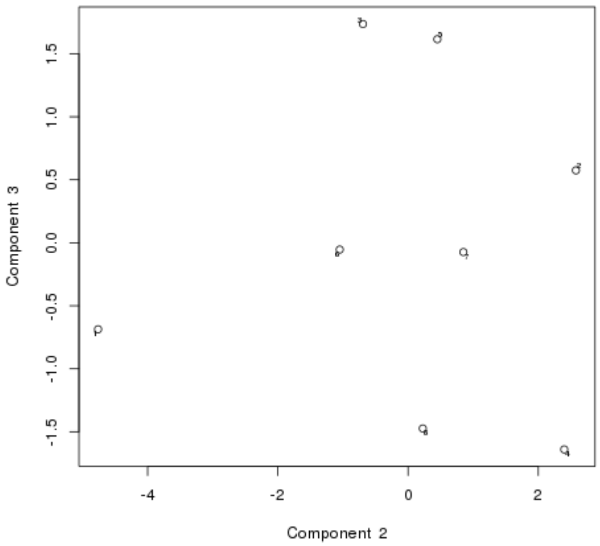
The application provides biplots, that are scatter plots of functional principal components. These plots show the importance of each observation defining the functional principal components.



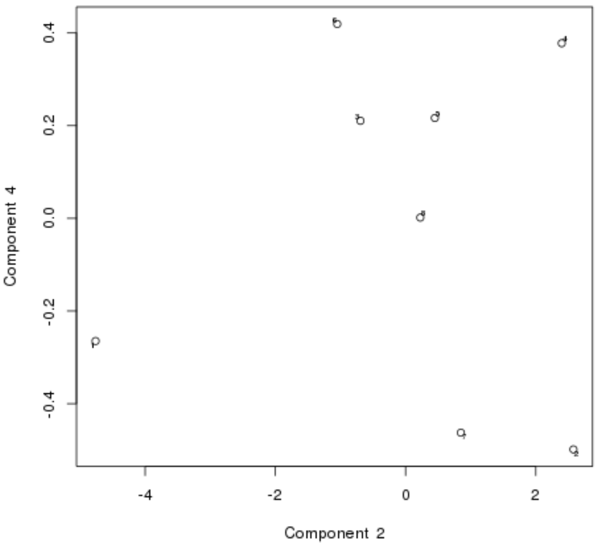
1st PC vs 2nd PC



1st PC vs 3rd PC



2nd PC vs 3rd PC



3rd PC vs 4th PC

The application also provides the scores or functional principal components

```
 ""       ,"Comp. 1","Comp. 2","Comp. 3","Comp. 4","Comp. 5","Comp. 6"
"Curve. 1", 5.80239 ,-4.760870,-0.686252,-0.264771,-0.063697, 0.001146
"Curve. 2", 3.02775 , 2.578402, 0.574890,-0.498647,-0.186860, 0.008495
"Curve. 3", 3.66384 ,-0.692866, 1.736547, 0.210481, 0.178087,-0.008018
"Curve. 4", 8.00051 , 2.398679,-1.639834, 0.377871,-0.044244,-0.006964
```

```
"Curve. 5", 0.40760 , 0.449900, 1.615755, 0.216605, 0.040690, 0.009390
"Curve. 6",-9.28523 ,-1.049343,-0.053715, 0.419396,-0.258698,-0.000796
"Curve. 7",-6.62764 , 0.849808,-0.073716,-0.462598, 0.068240,-0.015430
"Curve. 8",-4.98921 , 0.226289,-1.473676, 0.001663, 0.266482, 0.012178
```

Functional principal component analysis: FPC expansion

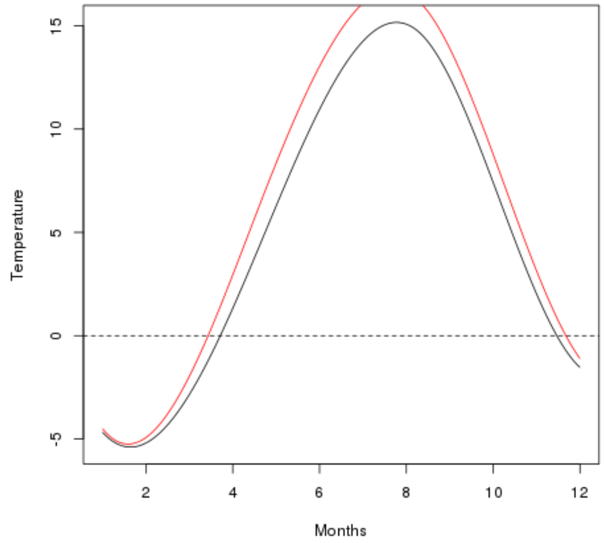
The original curves can be approximated by using a reduced set of functional principal components

$$x_i\left(t\right) = \sum_{j=1}^{q<p} \xi_{ij} f_j\left(t\right),\ i = 1, \ldots, n$$

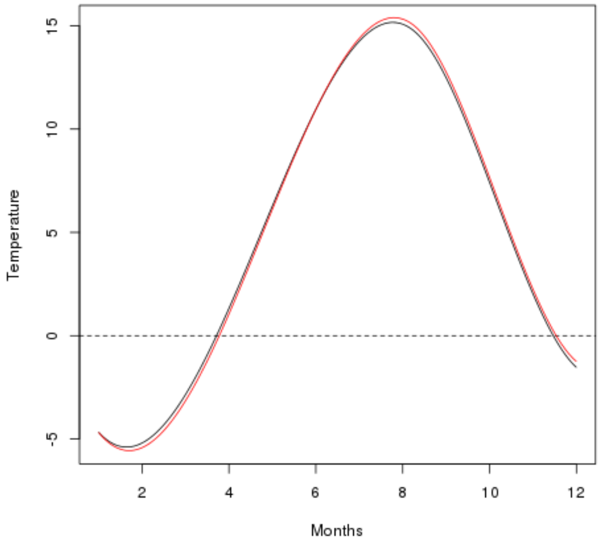
By expressing the principal component curves $f_j(t)$ in terms of basic functions, we have an approximation of the original curves in terms of basis fuctions, that is, by knowing their basis coefficients

$$x_i\left(t\right) = \sum_{j=1}^{q<p} \xi_{ij} \sum_{k=1}^{p} F_{jk} \phi_k\left(t\right) =,\ i = 1, \ldots, n$$

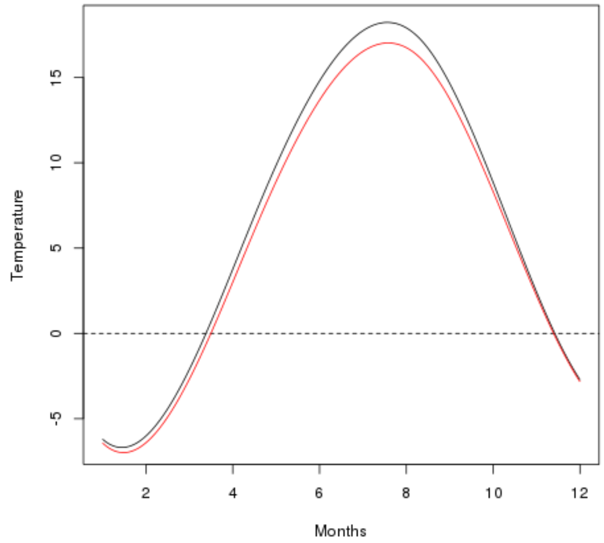
Our application provides a plot of the approximation of different curves in terms of different number q of functional principal components



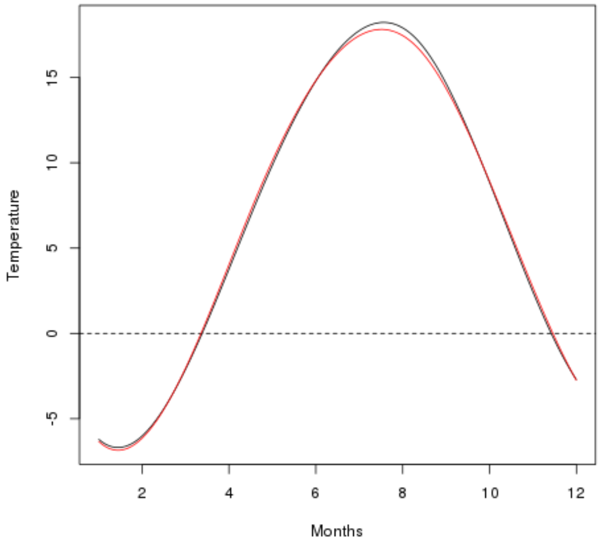
1st curve in terms of 1 FPC



1st curve in terms of 2 FPCs



2nd curve in terms of 1 FPC



2nd curve in terms of 2 FPCs

The application also provides the basis coefficients of the curves approximated by a number q of functional principal components. Here we show the basis coefficients of the approximation in terms of the first two functional principal components.

```
"       ,"bspl4.1","bspl4.2","bspl4.3","bspl4.4","bspl4.5","bspl4.6"
"Curve1", -4.67814, -7.85798, 4.060080,23.943853, 0.936901, -1.24858
"Curve2", -6.31577, -9.29195,11.065877,24.402237, 2.767888, -2.75522
"Curve3", -6.00030, -9.02024, 8.275102,24.154597, 1.842715, -2.45598
"Curve4", -2.89111, -6.25623, 8.184313,24.744840, 3.613687,  0.32198
"Curve5", -8.20413,-10.97127,10.926374,24.022235, 1.614979, -4.44119
"Curve6",-14.94922,-16.95483,15.163884,23.188481,-0.622028,-10.49294
```

```
"Curve7",-13.04369,-15.26084,15.111681,23.550604, 0.464715, -8.79043
"Curve8",-11.93662,-14.28089,13.741470,23.612851, 0.563870, -7.79295
```

Functional principal component linear regression

The functional linear model is a functional method to explain a scalar response variable y in terms of a functional predictor variable $x(t)$. So let us consider a set of observations (curves) of the functional predictor $x_1(t), x_2(t), \ldots, x_n(t)$ and associated to a set of observations of the scalar response $y_1, \ldots, y_n$, then the functional linear regression model is formulated as

$$y_i = \alpha + \int_T x_i(t)\,\beta(t)\,dt + \varepsilon_i,\ i = 1, \ldots, n.$$

This model, has a functional parameter instead of a set of scalar parameters.

In order to fit the functional linear model it is usual to consider that the curves of the predictor variable and the functional parameter are expressed in terms of the same basis of functions

$$\begin{aligned} x_i &= \sum_{j=1}^{p} a_{ij}\phi_j(t),\ i = 1, \ldots, n. \\ \beta(t) &= \sum_{k=1}^{p} \beta_k \phi_k(t) \end{aligned}$$

Under these conditions the functional linear regression model turns to a classical linear regression model

$$\begin{aligned} y_i &= \alpha + \int_T \left(\sum_{j=1}^{p} a_{ij}\phi_j(t)\right)\left(\sum_{k=1}^{p} \beta_k\phi_k(t)\right) dt + \varepsilon_i = \alpha + \sum_{j=1}^{p}\sum_{k=1}^{p} a_{ij}\left(\int_T \phi_j(t)\,\phi_k(t)\,dt\right)\beta_k,\ i = 1, \ldots, n. \\ &= \alpha + \sum_{j=1}^{p}\sum_{k=1}^{p} a_{ij}\psi_{jk}\beta_k,\ i = 1, \ldots, n,\ \psi_{jk} = \int_T \phi_j(t)\,\phi_k(t)\,dt. \end{aligned}$$

This model has as slope parameters the basis coefficients of functional parameter and as design matrix the product of the matrix of basis coefficients of curves multiplied by the matrix of scalar products between basic functions.

The problem with this model is the high multicollinearity that usually has (see Escabias et al. (2004)).

Our application is designed with the solution based on functional principal component analysis, that consist of expressing the curves in terms of a reduced set of functional principal components.

The functional principal component linear regression is formulated as

$$\begin{aligned} y_i &= \alpha + \sum_{l=1}^{q} \xi_{il}\gamma_l + \varepsilon_i = \alpha + \sum_{l=1}^{q}\left(\int_0^T \left(x_i(t) - \overline{x}(t)\right) f_l(t)\,dt\right)\gamma_l + \varepsilon_i \\ &= \alpha - \sum_{l=1}^{q}\gamma_l \int_0^T \overline{x}(t)\,f_l(t)dt + \int_0^T x_i(t)\left(\sum_{l=1}^{q}\gamma_l f_l(t)\right) dt \end{aligned}$$

In this model the intercept parameter is

$$\alpha - \sum_{l=1}^{q}\gamma_l \int_0^T \overline{x}(t)\,f_l(t)dt$$

and the parameter function

$$\beta(t) = \sum_{l=1}^{q}\gamma_l f_l(t)$$

By expressing the functional curves $f_l(t)$ in terms of basic functions

$$\beta(t) = \sum_{l=1}^{q} \gamma_l \left(\sum_{k=1}^{p} F_{lk} \phi_k(t) \right)$$

we obtain the basis coefficients of the functional parameter.

$$\beta(t) = \sum_{k=1}^{p} \left(\sum_{l=1}^{q} F_{lk} \gamma_l \right) \phi_k(t)$$

In Statfda the user can select the number q of functional principal components to use to fit the model.

Functional principal component linear regression: fitted model

The fitted model provides the estimation of the model in terms of the selected functional principal components.

$$y_i \quad = \quad \alpha + \sum_{l=1}^{q} \xi_{il} \gamma_l + \varepsilon_i$$

The user can download the classical fitted model provided by R

```
[1] "Fitted linear model"

Call:
lm(formula = RespuestaLin ~ ., data = DataFrameLineal)

Residuals:
        1         2         3         4         5         6         7         8 
-0.005874 -0.018620  0.016046 -0.003968  0.005309 -0.025248  0.007178  0.025176 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.886250   0.009141 534.537 1.44e-08 ***
‘Comp. 1‘    0.168936   0.001558 108.446 1.73e-06 ***
‘Comp. 2‘    0.252073   0.004215  59.803 1.03e-05 ***
‘Comp. 3‘    0.058054   0.007692   7.547  0.00482 ** 
‘Comp. 4‘   -0.035202   0.026630  -1.322  0.27796    
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.02585 on 3 degrees of freedom
Multiple R-squared:  0.9998,Adjusted R-squared:  0.9995 
F-statistic:  3849 on 4 and 3 DF,  p-value: 6.798e-06
```

Functional principal component linear regression: functional parameter

This section provides the plot of the estimation of the functional parameter of the model

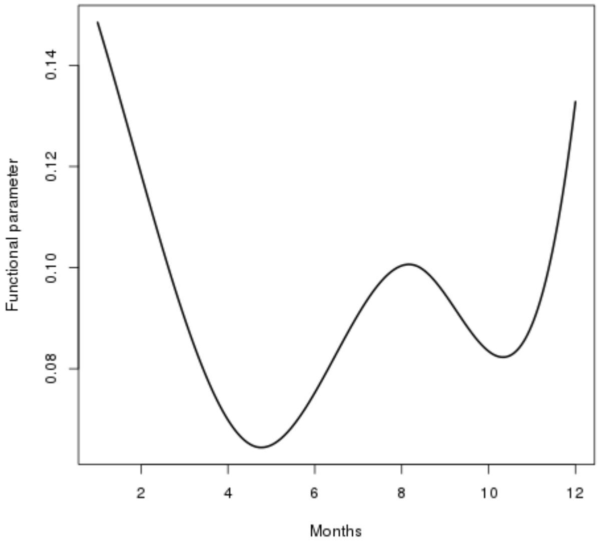
$$y_i = \alpha + \int_T x_i\left(t\right) \beta\left(t\right) dt + \varepsilon_i, \; i = 1, \ldots, n.$$

from its basis coefficients obtained as

$$\beta(t) = \sum_{k=1}^{p} \left(\sum_{l=1}^{q} F_{lk} \gamma_l \right) \phi_k(t)$$

in terms of estimated parameters obtained by using the functional principal components as covariates.

The estimation from the first four functional principal components obtained is



```
""        ,"V1"
"bspl4.1",0.1485
"bspl4.2",0.1138
"bspl4.3",0.0200
"bspl4.4",0.1466
"bspl4.5",0.0462
"bspl4.6",0.1328
```

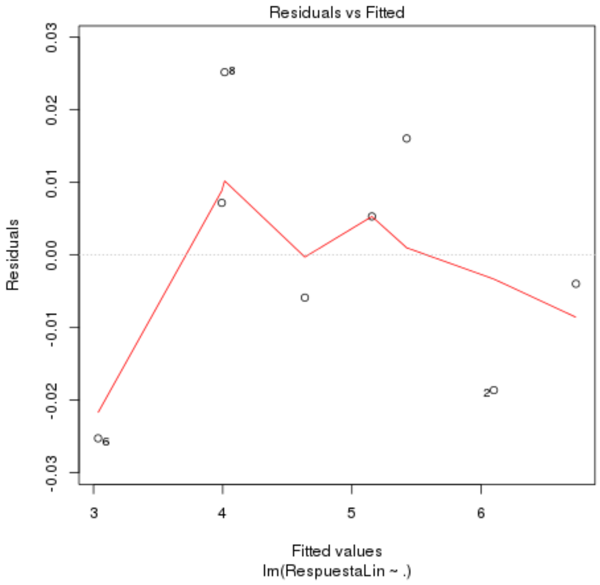
Functional principal component linear regression: residual analysis

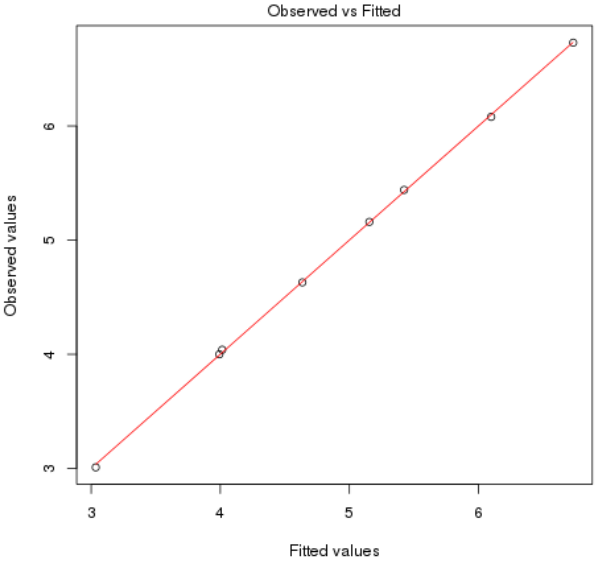
The residual analysis is obtained like any other in a lineal model, by calculating the fited values

$$\widehat{y}_i = \widehat{\alpha} + \int_T x_i\left(t\right)\widehat{\beta}\left(t\right)dt = \widehat{\alpha} + \sum_{j=1}^{p}\sum_{k=1}^{p} a_{ij}\psi_{jk}\widehat{\beta}_k,\ i = 1,\ldots,n,\ \psi_{jk} = \int_T \phi_j\left(t\right)\phi_k\left(t\right)dt.,\ i = 1,\ldots,n.$$

the residual values $e_i = y_i - \widehat{y}_i$ and plotting residuals vs fitted values and observed vs fitted values.

The applicaton provides the classical plots provided by R and the matrix of observed and fitted values and residuals.





```
""       ,"Observed","Fitted","Residuals"
"Curve1",      4.63,  4.6359,   -0.0059
"Curve2",      6.08,  6.0986,   -0.0186
"Curve3",      5.44,  5.4240,    0.0160
"Curve4",      6.73,  6.7340,   -0.0040
"Curve5",      5.16,  5.1547,    0.0053
"Curve6",      3.01,  3.0352,   -0.0252
"Curve7",      4.00,  3.9928,    0.0072
"Curve8",      4.04,  4.0148,    0.0252
```

Functional principal component linear regression: prediction

Once fitted a linear model, it is possible to make predictions of the response for a new curve. Let $x(t)$ be a new curve observed discretely as $x(t_1),\ldots,x(t_m)$, once its basis coefficients are calculated it is possible to plot it and make a prediction of the response as

$$\widehat{y} = \widehat{\alpha} + \int_T x\left(t\right)\widehat{\beta}\left(t\right)dt = \widehat{\alpha} + \sum_{j=1}^{p}\sum_{k=1}^{p} a_j\psi_{jk}\widehat{\beta}_k,\ i=1,\ldots,n,\ \psi_{jk} = \int_T \phi_j\left(t\right)\phi_k\left(t\right)dt.,\ i=1,\ldots,n.$$

The plot of the new curve is shown by Statfda and the response prediction also.

Functional logistic regression model

The functional logistic regression model is defined to model and predict a binary response variable from a functional predictor. To this end, various functional models have been proposed (Cardot and Sarda, 2005; Rossi et al., 2002). Ramsay and Silverman (1997) proposed diverse functional models based on basis expansion, and since then other authors have adopted these methods to predict a binary outcome from functional predictors (Ratcliffe et al., 2002; Escabias et al., 2004; Aguilera et al., 2008).

In order to formulate the functional logit model let Y be a binary response random variable and let $\{X\left(t\right) : t \in T\}$ be a functional covariate related to Y. Given $x_1\left(t\right),\ldots,x_n\left(t\right)$ a sample of curves of the functional predictor and $y_1,\ldots,y_n$ a sample of the response associated with the n curves, the model is expressed as

$$y_i = \pi_i + \varepsilon_i = \pi\left(x_i\left(t\right)\right) + \varepsilon_i,\ i=1,\ldots,n,$$

where

$$\pi_i = \frac{\exp\left\{\alpha + \int_T x_i\left(t\right)\beta\left(t\right)dt\right\}}{1+\exp\left\{\alpha + \int_T x_i\left(t\right)\beta\left(t\right)dt\right\}},\ i=1,\ldots,n, \qquad (2)$$

$\boldsymbol{\varepsilon} = \left(\varepsilon_1,\ldots,\varepsilon_n\right)'$ the vector of centered random errors, with unequal variances and a Bernoulli distribution, and $\beta(\cdot)$ the functional parameter to be estimated.

This model can also be expressed in terms of the *logit* transformations as

$$l_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \alpha + \int_T x_i\left(t\right)\beta\left(t\right)dt,\ i=1,\ldots,n,$$

If we consider the sample paths $x_1\left(t\right),\ldots,x_n\left(t\right)$ and the functional parameter expressed in terms of the basis $\{\phi_1\left(t\right),\ldots,\phi_p\left(t\right)\}$, in the form

$$\begin{aligned} x_i &= \sum_{j=1}^{p} a_{ij}\phi_j\left(t\right),\ i=1,\ldots,n. \\ \beta\left(t\right) &= \sum_{k=1}^{p} \beta_k\phi_k\left(t\right). \end{aligned}$$

The functional logit model in terms of the logit transformations is then expressed as

$$l_i = \alpha + \sum_{j=1}^{p} \sum_{k=1}^{p} a_{ij}\psi_{jk}\beta_k,\ i = 1, \ldots, n$$

with ψ_{jk} being the scalar products between the basis functions

$$\psi_{jk} = \int_T \phi_j(t)\,\varphi_k(t)\,dt,\ j,k = 1, \ldots, p$$

The functional logit model is now a classical logit model which in matrix form and in terms of logit transformations is expressed as

$$L = X\beta$$

where

$L = (l_1, \ldots, l_n)'$ is the vector of logit transformations.

$X = (\mathbf{1} \mid A\Psi)$ is the design matrix, and $\mid$ indicating the separation between the two boxes of the matrix.

$\mathbf{1} = (1, \ldots, 1)'$ is a $n-$length vector of ones.

Ψ is the matrix whose entries (ψ_{jk}) are the scalar products between basic functions setted abobe.

A is the matrix of sample curve basis coefficients as rows.

$\beta = (\beta_0, \beta_1, \ldots, \beta_q)'$ with $\beta_0 = \alpha$ are the basis coefficients of the functional parameter. These coefficients would be the parameters of the multiple model to be estimated.

As in the linear case the problem with this model is the high multicollinearity that usually has (see Escabias et al. (2004)).

Our application is designed with the solution based on functional principal component analysis, that consist of expressing the curves in terms of a reduced set of functional principal components.

The functional principal component logit regression is formulated in terms of the logit transformations as

$$\begin{aligned} l_i &= \alpha + \sum_{l=1}^{q} \xi_{il}\gamma_l + \varepsilon_i = \alpha + \sum_{l=1}^{q} \left(\int_0^T (x_i(t) - \overline{x}(t))\, f_l(t)\,dt \right) \gamma_l + \varepsilon_i \\ &= \alpha - \sum_{l=1}^{q} \gamma_l \int_0^T \overline{x}(t)\, f_l(t)dt + \int_0^T x_i(t) \left(\sum_{l=1}^{q} \gamma_l f_l(t) \right) dt \end{aligned}$$

In this model the intercept parameter is

$$\alpha - \sum_{l=1}^{q} \gamma_l \int_0^T \overline{x}(t)\, f_l(t)dt$$

and the parameter function

$$\beta(t) = \sum_{l=1}^{q} \gamma_l f_l(t)$$

By expressing the functional curves $f_l(t)$ in terms of basic functions

$$\beta(t) = \sum_{l=1}^{q} \gamma_l \left(\sum_{k=1}^{p} F_{lk}\phi_k(t) \right)$$

we obtain the basis coefficients of the functional parameter.

$$\beta(t) = \sum_{k=1}^{p} \left(\sum_{l=1}^{q} F_{lk}\gamma_l \right) \phi_k(t)$$

In Statfda the user can select the number q of functional principal components to use to fit the model.

Functional principal component logit regression: fitted model

The fitted model provides the estimation of the model in terms of the selected functional principal components.

$$l_i \quad = \quad \alpha + \sum_{l=1}^{q} \xi_{il}\gamma_l + \varepsilon_i$$

or equivalently

$$y_i = \pi_i + \varepsilon_i, \; i = 1, \ldots, n,$$

where

$$\pi_i = \frac{\exp\{l_i\}}{1 + \exp\{l_i\}}, \; i = 1, \ldots, n, \tag{3}$$

The user can download the classical fitted model provided by R

```
[1] "Fitted logit model"

Call:
glm(formula = Respuesta ˜ ., family = binomial, data = DataFrameLogit)

Deviance Residuals:
         1           2           3           4           5           6
-5.794e-06   1.208e-05   1.136e-06   1.314e-06   7.294e-06  -7.211e-07
         7           8
-1.308e-05  -5.222e-07

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.822e-01  1.432e+05       0        1
Comp.1       3.021e+00  9.630e+03       0        1
Comp.2       6.391e+00  4.085e+04       0        1
Comp.3       1.061e+01  1.416e+05       0        1
Comp.4       1.605e+01  3.401e+05       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.1090e+01  on 7  degrees of freedom
Residual deviance: 4.0768e-10  on 3  degrees of freedom
AIC: 10

Number of Fisher Scoring iterations: 23
```

Functional principal component logit regression: functional parameter

This section provides the plot of the estimation of the functional parameter of the model

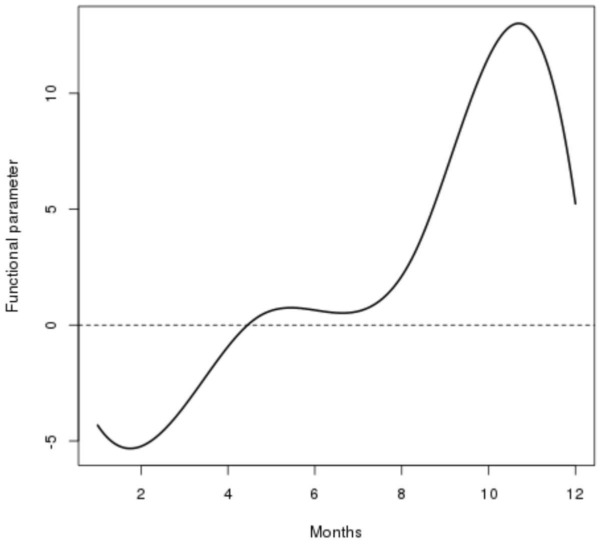
$$l_i = \alpha + \int_T x_i\left(t\right)\beta\left(t\right)dt + \varepsilon_i, \; i = 1, \ldots, n.$$

from its basis coefficients obtained as

$$\beta(t) = \sum_{k=1}^{p}\left(\sum_{l=1}^{q} F_{lk}\gamma_l\right)\phi_k(t)$$

in terms of estimated parameters obtained by using the functional principal components as covariates.

The estimation from the first four functional principal components obtained is



```
""         ,    "V1"
"bspl4.1",-4.3175
"bspl4.2",-7.8030
"bspl4.3", 5.3605
"bspl4.4",-5.1388
"bspl4.5",21.4026
"bspl4.6", 5.2361
```

Functional principal component linear regression: ROC area and CCR

To evaluate the predictive ability of the model, it is usual to calculate the rate of correct classifications and the area under the ROC curve.

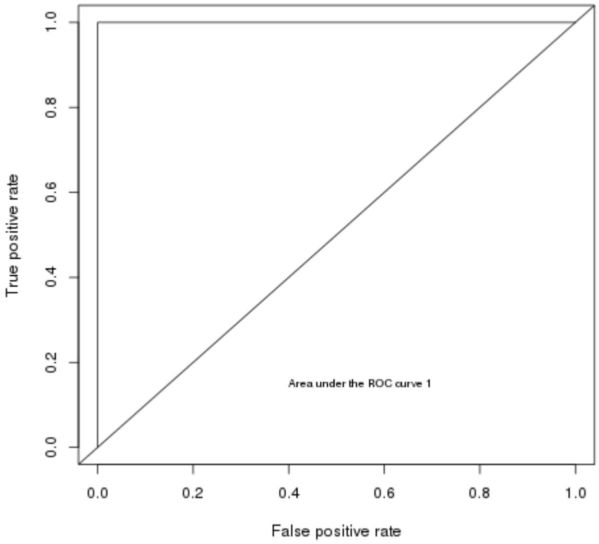
The correct classification rate (CCR) is one of the most commonly used measures in logistic regression to assess the goodness of predictions. To calculate the CCR a cutoff point p_c (usually $p_c = 0.5$) is chosen and an observation is considered to be correctly classified when the estimated probability $\widehat{\pi}_i \geq p_c$ and $y_i = 1$ or $\widehat{\pi}_i < p_c$ and $y_i = 0$, otherwise it is considered classified incorrectly. Thus, the CCR is defined as the ratio between the number of observations correctly classified and the total number of sample observations.

Although a cutoff value of 0.5 is usually used, it would be more appropriate to use the cutoff point that maximises the CCR (Hosmer and Lemeshow, 1989), which is usually very close to the proportion of ones in the sample.

Stafda allows to select any cutoff point p_c that the user wants.

The ROC curve is a graph that evaluates the model's ability to discriminate. The fitted logistic regression model predicts the value of the response depending on whether the predicted probability is greater than or equal to the cut-point chosen to discriminate. The logistic regression model is considered a good predictor if it predicts as a success those individuals actually observed to be successes and predicts as a failure those individuals observed to be failures. The ROC curve plots the true positive rate ($y = 1, \widehat{y} = 1$) against the false positive rate ($y = 0, \widehat{y} = 1$) for different cutoff points. The nearest point to the unit is the best discrimination point and the area under the curve is a measure of the capacity to discriminate. The closer this measure is to one, the better it is, and an acceptable value would be 0.7 or higher.

Statfda provides ROC area, classification table and CCR:



```
[1] "Classification table"
        Predicted
Observed 0 1
       0 4 0
       1 0 4
[1] "Correct Classification Rate"
[1] 100
```

Functional principal component linear regression: prediction

Once fitted a logit model, it is possible to make predictions of the response for a new curve. Let $x(t)$ be a new curve observed discretely as $x(t_1), \ldots, x(t_m)$, once its basis coefficients are calculated it is possible to plot it and make a prediction of the logit transformation and the response as

$$\begin{aligned}
\widehat{l} &= \widehat{\alpha} + \int_T x\left(t\right) \widehat{\beta}\left(t\right) dt = \widehat{\alpha} + \sum_{j=1}^{p} \sum_{k=1}^{p} a_j \psi_{jk} \widehat{\beta}_k, \ i = 1, \ldots, n, \ \psi_{jk} = \int_T \phi_j\left(t\right) \phi_k\left(t\right) dt., \ i = 1, \ldots, n. \\
\widehat{\pi} &= \frac{\exp\{\widehat{l}\}}{1 + \exp\{\widehat{l}\}}
\end{aligned}$$

The plot of the new curve is shown by Statfda and the prediction also.

Bibliography

Aguilera AM, Escabias M and Valderrama MJ (2008) Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus. Computational Statistics and Data Analysis 53:151-163

Aguilera AM, Escabias M and Valderrama MJ (2008) Forecasting binary longitudinal data by a functional PC-ARIMA model. Computational Statistics and Data Analysis 52:3187-3197

Cardot H and Sarda P (2005) Estimation in generalized linear models for functional data via penalized likelihood. J. Multivariate Anal 92:24-41

De Boor, C. (2001). A practical guide to Splines. Springer, New York.

Escabias M, Aguilera AM and Valderrama MJ (2004) Principal component estimation of functional logistic regression. Journal of Nonparametric Statistics 16 (3-4):365-384

Escabias M, Aguilera AM and Valderrama MJ (2005) Modeling environmental data by functional principal component logistic regression. Environmetrics 16:95-107

Escabias M, Aguilera AM and Valderrama MJ (2007) Functional PLS logit regression model. Computational Statistics and Data Analysis 51:4891-4902

Ferraty F and Vieu P (2006) Nonparametric Functional Data Analysis. Springer, New York

Hosmer DW, Hosmer T, Le Cessie S and Lemeshow S (1997) A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine 16:965-980

Hosmer DW and Lemeshow S (1989) Applied Logistic Regression. Wiley, New York.

James JM (2002) Generalized linear models with functional predictors. J. Roy. Statist. Soc. Ser. B 64(3):411-432

Müller HG (2008) Functional modeling of longitudinal data. In: Fitzmaurice G, Davidian M, Verbeke G and Molenberghs G (eds) Longitudinal Data Analysis (Handbooks of Modern Statistical Methods) Chapman & Hall/CRC, New York, pp 223-252.

Ocaña-Peinado FM, Valderrama MJ and Aguilera AM (2008) A dynamic regression model for air pollen concentration. Stoch Environ Res Risk Assess 22(1):59-63.

Ramsay JO and Silverman BW (1997) Functional Data Analysis. Springer-Verlag, New York

Ramsay JO and Silverman BW (2005) Functional Data Analysis. 2nd edition Springer-Verlag, New York

Ratcliffe SJ, Heller GZ and Leader LR (2002) Functional data analysis with application to periodically stimulated foetal heart rate data. II: functional logistic regression. Statist. Med. 21:1115-1127

Rossi N, Wang X and Ramsay JO (2002) Nonparametric item response function estimates with the EM algorithm. J. Behav. Educ. Sci. 27:291-317

Valderrama MJ, Ocaña FA, Aguilera AM and Ocaña-Peinado FM (2010) Forecasting Pollen Concentration by a Two-Step Functional Model. Biometrics 66(2):578-585