

# Statfda: a web APP for functional data analysis

Manuel Escabias and Ana María Aguilera

## Introduction

This web application allows to use functional data analysis with the minimum knowledge of theoretical aspects of functional data analysis.

The application is designed to get a functional data analysis in three different situations depending of the amount of information the user has:

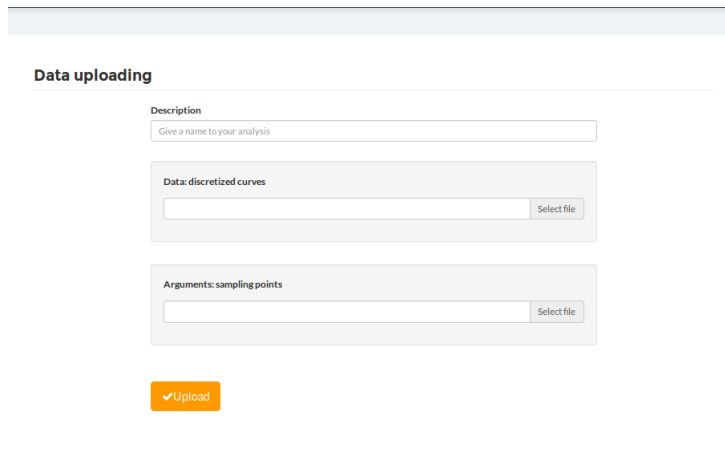
**Case 1.** User has only discrete observations of curves  $x_{ij}$ . The number of discrete observations must be the same for all curves. Time points where curves are observed is not important for the analysis.

**Case 2.** User has discrete observations of curves  $x_i(t_j)$ . The number of discrete observations must be the same for all curves. Time points where curves are observed are important for the analysis and the same for each curve  $t_1, t_2, \dots, t_m$

**Case 3.** User has discrete observations of curves  $x_i(t_{ij})$ . The number of discrete observations cannot be the same for all curves. Time points where curves are observed are important for the analysis and can be different for each curve  $t_{i1}, t_{i2}, \dots, t_{im_i}$

## Step 1: Data uploading

The first step consist of uploading data for our analysis. The uploading-data page allows selecting a name for our analysis and has available two places for data uploading.



The screenshot shows a web form titled "Data uploading". It contains a "Description" section with a text input field labeled "Give a name to your analysis". Below this are two sections for file uploads: "Data: discretized curves" and "Arguments: sampling points". Each section has a text input field and a "Select file" button. At the bottom of the form is an orange "Upload" button with a checkmark icon.

**Description.** Write here the description of your analysis

**Data: discretized curves.** Browse to the folder with the file of discretized curves and select it.

**Arguments: sampling points.** Browse to the folder with the file of sampling points and select it.

Depending on the case, the file format may be different.

**Case 1.** The user has only to upload the file with the discretized curves leaving the sampling points free.

**It is important the file format to be correct.** The file must be a .csv file separated by commas with point as decimal character and with no names in rows or columns. It is considered that each row corresponds with the observations of one curve. **It is important the number of columns in each row to be the same.** The system returns a warning if it detect a row with less or more columns than the rest.

```
-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
```

**Case 2.** The user has to upload the file with the discretized curves and the file with the sampling points. **It is important the file format to be correct.** The discretized curves file must have the same format explained for Case 1. The sampling points file must be a .csv file separated by commas with point as decimal character with no names in rows or columns and with only one row. **It is important the number of columns of both files to be the same.** The system returns a warning if it detect a row with less or more columns than the rest and if the number of columns of both files is different.

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

**Case 3.** The user has to upload the file with the discretized curves and the file with the sampling points. **It is important the files format to be correct.** Both files must be .csv files separated by commas with point as decimal character with no names in rows or columns. **It is not necessary the number of columns of all rows to be the same, but the format of discretized curves file and sampling points file to be the same. The same row in both files with the same length.** The system returns a warning if it detect differences in both files.

Discretized curves file

```
-4.65,-5.33,-2.53,1.26,5.79,10.79,15.21,15.28,11.62,7.02
-6.16,-6.18,-1.74,3.62,9.44,14.78,18.38,18.20,13.87,8.49,3.24
-5.72,-6.80,-2.94,1.85,7.50,13.14,17.49,17.64,13.31
-3.22,-3.49,-0.15,4.69,9.34,13.40,16.29,16.60,13.59,9.25
-8.11,-8.26,-3.44,2.32,8.76,14.44,18.29,17.98,13.44,8.03,2.57
-15.36,-13.23,-5.82,2.95,10.20,16.00,18.66,17.06,11.95
-12.78,-11.28,-4.67,3.29,10.91,16.38,19.05,17.65
-11.82,-10.65,-4.04,3.93,10.79,15.49,17.97,16.69,11.96,6.38
```

Sampling points file

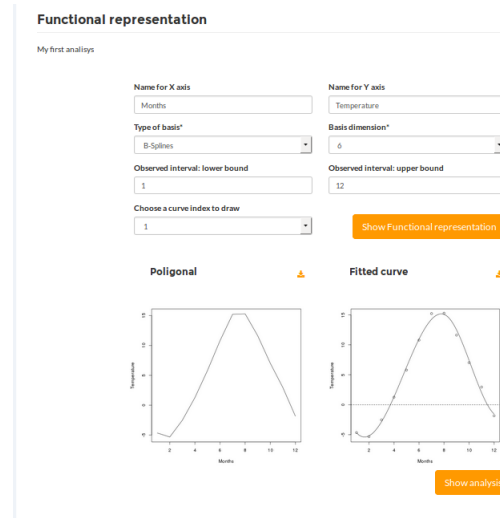
```
1,2,3,4,5,6,7,8,9,10
2,3,4,5,6,7,8,9,10,11,12
2,3,4,5,6,7,8,9,10
1,2,3,4,6,7,9,10,11,12
1,2,3,4,5,6,7,8,9,10,11,12
1,2,4,5,6,7,9,10,12
2,3,5,6,7,8,10,11
2,3,4,5,6,7,8,9,10,11
```

## Step2: functional representation

In the second step, the user can select the names that will appear in X-axis and Y-axis of the different curves plots that the application shows. This step is very important because user must select the type of basis and

the dimension to use for the functional data analysis. Two type of basis have been considered: cubic B-spline and Fourier basis. The user can also select the interval bounds of the curves domain in case the user had only uploaded the file of discretized curves. Finally the user can select the curve to show in the functional representation.

Once all the fields are completed, the application shows the functional representation for the selected curve.



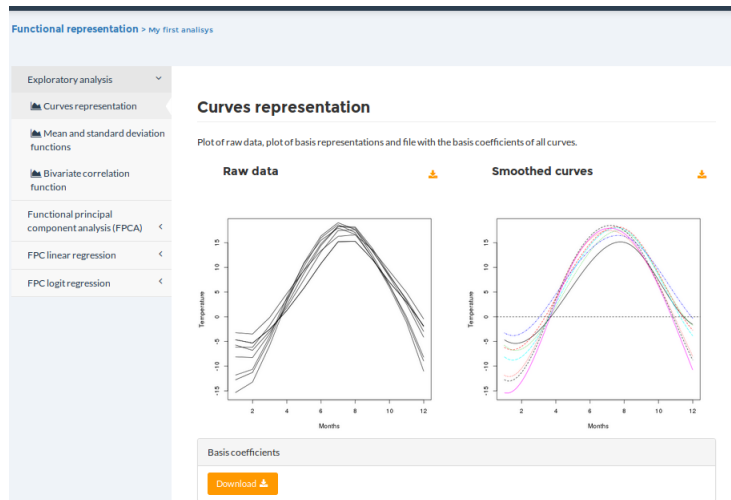
The user can change the type of basis, the dimension and the curve to show to see different functional representations. Once the user decides the optimum functional representation he/she must click the "Show analysis" button to get the different functional analysis. All these analysis will be driven with the type of basis and the dimension selected here.

## Analysis provided by the application

The application provide four analysis blocks: functional exploratory analysis, functional principal component analysis, functional principal component linear regression analysis and functional principal component logit regression analysis.

### Functional exploratory analysis: curves representation

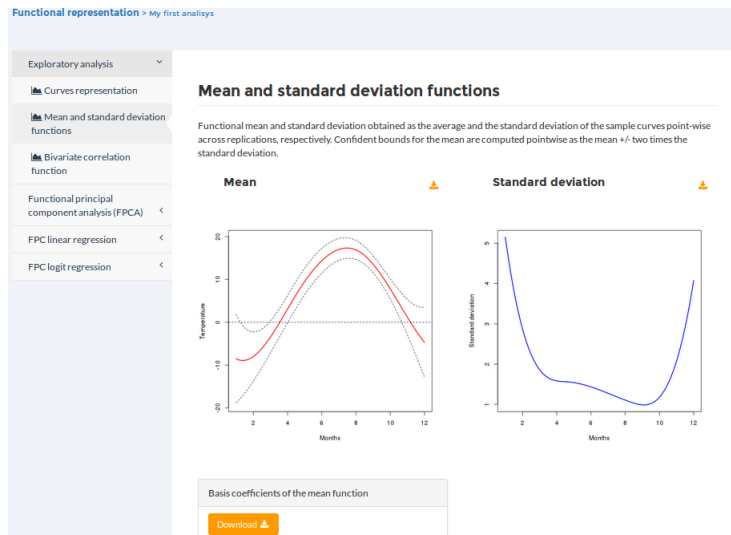
In this section, the application provide a plot of raw data and functional representation, and the matrix of basis coefficients of all curves.



The user can download both plots and the matrix of basis coefficients in .csv format with columns separated by comma and with period as decimal point.

## Functional exploratory analysis: Mean and standard deviation functions

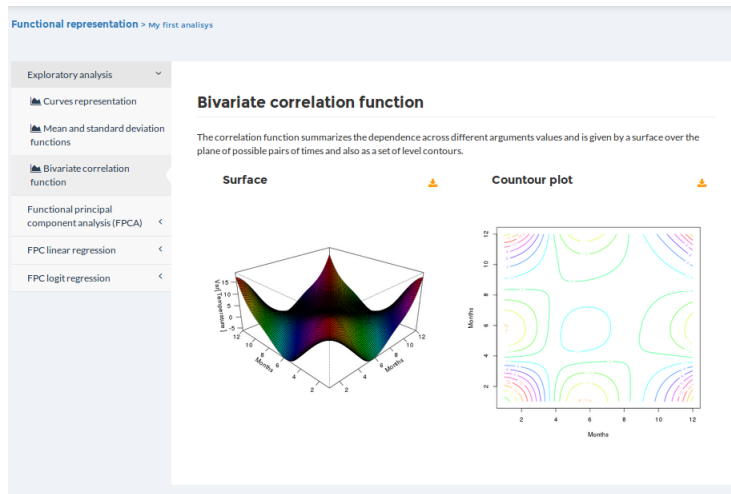
In this section, the application provide a plot of mean and standard deviation functions, and the vector of basis coefficients of mean curve.



The user can download both plots and the vector of basis coefficients in .csv format with columns separated by comma and with period as decimal point.

## Functional exploratory analysis: Bivariate correlation function

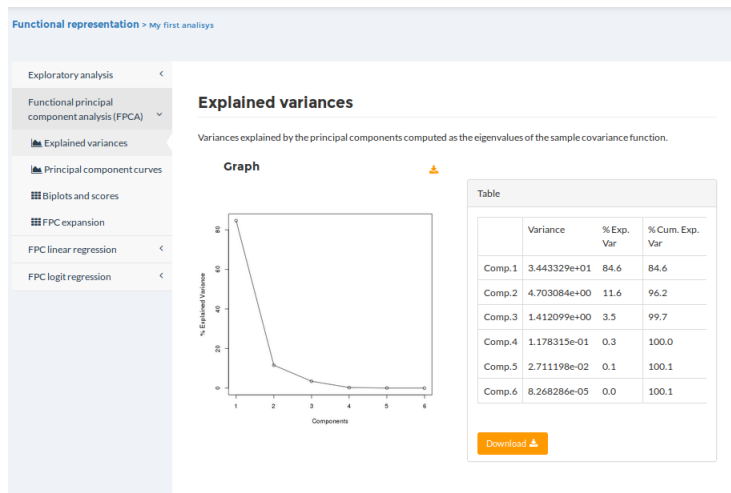
In this section, the application provide the surface and contour plots of bivariate correlation function.



The user can download both plots.

## Functional principal component analysis: Explained variances

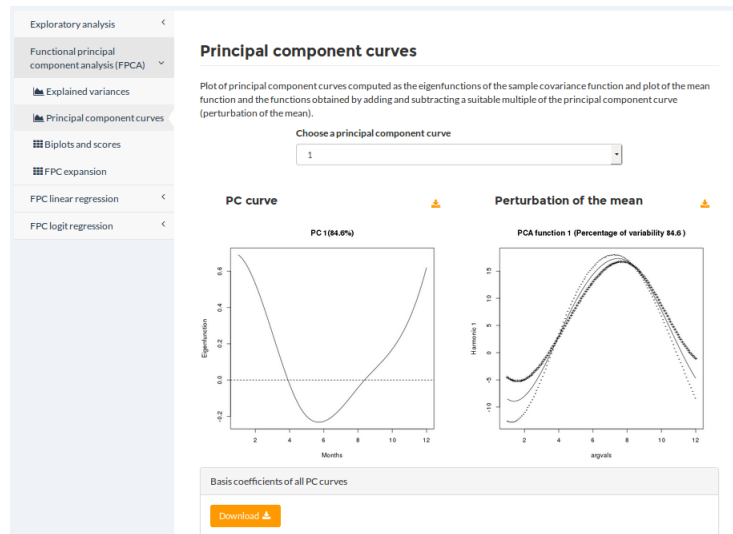
In this section, the application provide a plot and a table of explained variance.



The user can download plot and table.

## Functional principal component analysis: Principal component curves

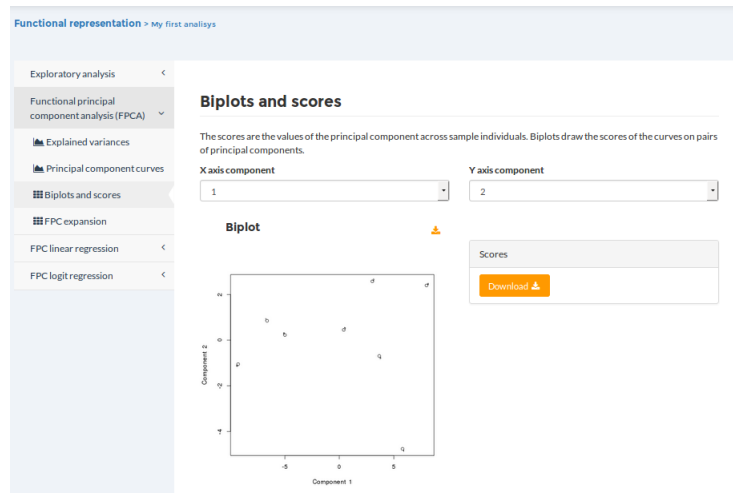
In this section, the application provide a plot of eigenfunction of the functional principal component analysis and the perturbation of mean function by this eigenfunction. The user can select the eigenfunction and its corresponding mean perturbation to show. Every time the user change the eigenfunction to show, the plot changes. Finally the basis coefficients of eigenfunctions are provided.



The user can download all plots and the matrix of basis coefficients in .csv format with columns separated by comma and with period as decimal point.

## Functional principal component analysis: Biplots and scores

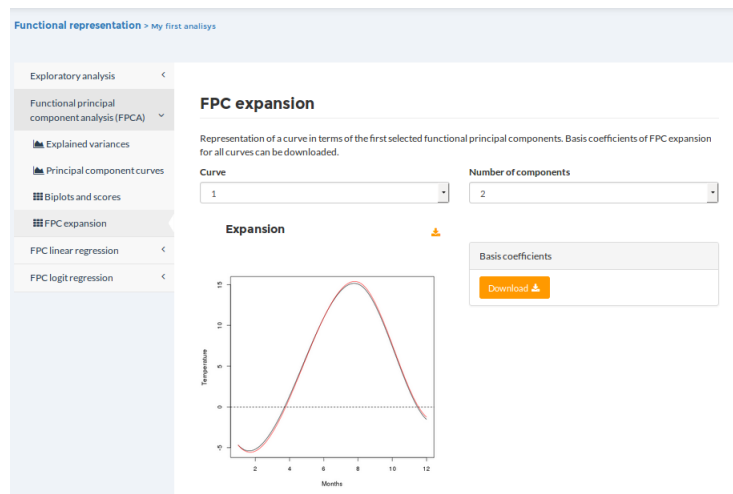
In this section, the application provide the classical biplot of components. The user can select the components to show in each axis, so every time the user choose an option the plot changes. Finally the functional principal components scores are provided.



The user can download all plots and the matrix of scores in .csv format with columns separated by comma and with period as decimal point.

## Functional principal component analysis: Functional representation

In this section, the application provide the representation of curves in terms of a set of functional principal components. The user can select the curve to show and the number of functional principal components to use in the representation. Finally the basis coefficients of the represented curves are provided in .csv format. Every time the user changes the number of components to use, the plot and .csv change. The plot also changes when it is modified the curve-to-show option.



The user can download all plots and the matrix of basis coefficients in .csv format with columns separated by comma and with period as decimal point.

## Functional principal component linear regression: upload scalar response

Functional principal component linear regression needs a scalar response to fit the model (see theory of functional principal component linear regression).

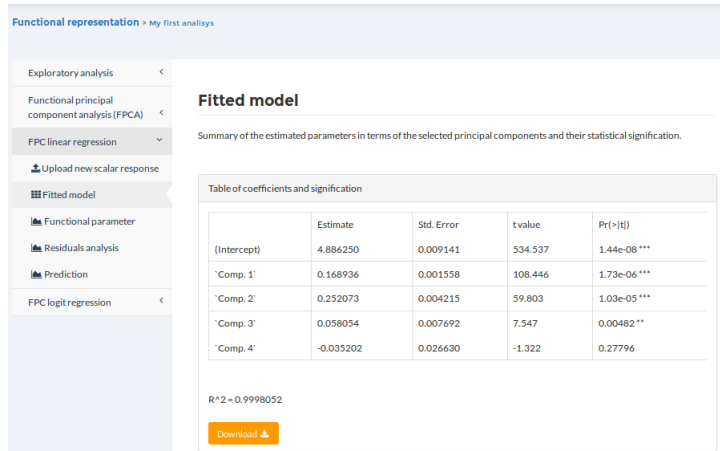
The user must upload the file with the response. **It is important the file format to be correct.** The file must be a .csv with only one column with no names in rows or columns. **It is important the number of rows to be the same as the file of discretized curves has.** The system returns a warning if it detects differences in both files.

4.63  
6.08  
5.44  
6.73  
5.16  
3.01  
4.00  
4.04

The user must also select the number of functional principal components to use as covariates of the model.

## Functional principal component linear regression: Fitted model

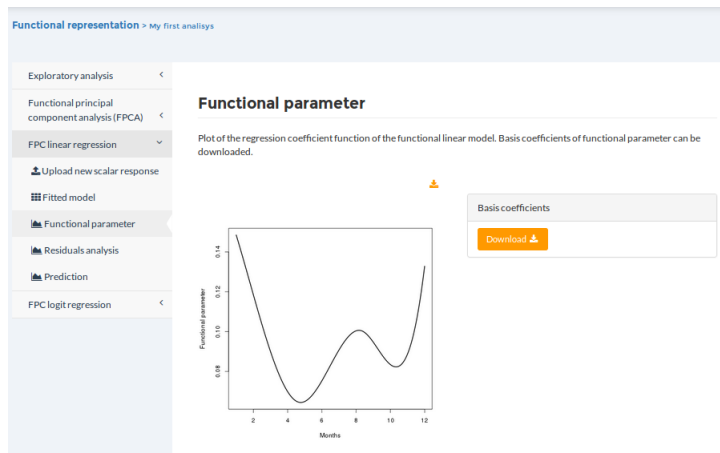
In this section, the application provide the fitted model of the functional principal component linear regression model. More exactly the table of parameters corresponding to each functional principal component.



The user can download this table more completely in .txt format, in the form that usually return R software.

## Functional principal component linear regression: Functional parameter

In this section, the application provide a plot of the functional parameter of the functional principal component linear regression model, and the vector of basis coefficients of this functional parameter.

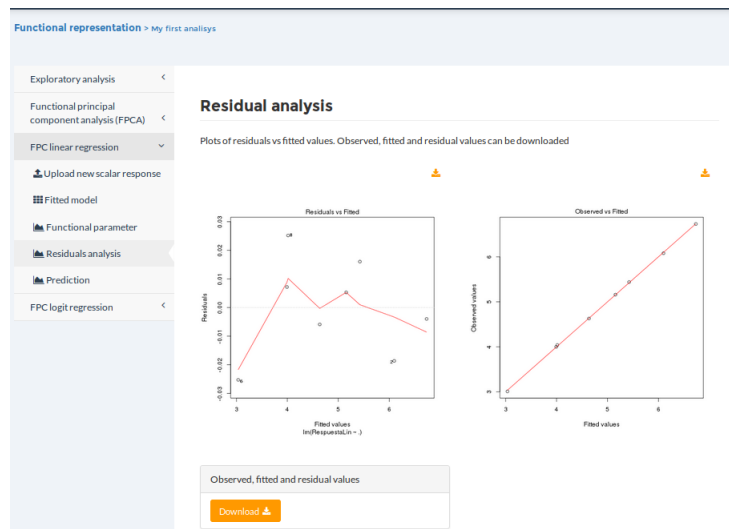


The user can download the plot and the vector of basis coefficients in .csv format with columns separated by comma and with period as decimal point.

## Functional principal component linear regression: Residual analysis

In this section, the application provide two classical plots of a residual analysis in R. A matrix of observed, fitted and residual values is also provided.





The user can download the plots and the matrix in .csv format with columns separated by comma and with period as decimal point.

## Functional principal component linear regression: prediction

Functional principal component linear regression prediction needs a new curve to get the prediction that the model gives.

The user must upload the file with the new curve. **It is important the file format to be correct.** The file must be a .csv with only one row and no names in rows or columns. The system returns a error if the format is not correct. **For prediction, the application is designed differently depending of the format of discretized curves previously uploaded. For cases 1 and 2, one file with one row and as many columns of discretized curves uploaded has, is necessary**

-4.65, -5.33, -2.53, 1.26, 5.79, 10.79, 15.21, 15.28, 11.62, 7.02, 2.95, -1.85

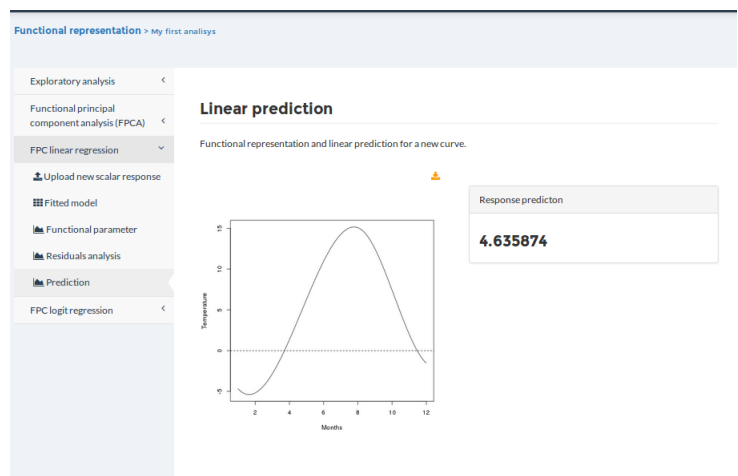
**For case 3 there are necessary two files: one with the row of discretized curves**

-12.78, -11.28, -4.67, 3.29, 10.91, 16.38, 19.05, 17.65

**and the other one with the sampling points. The number of columns of both files must be the same**

2,3,5,6,7,8,10,11

The result of the analysis is the plot of the new curve and the prediction of the response



The user can download the plot.

Functional principal component logit regression: upload scalar response

Functional principal component logit regression needs a scalar response of 0's and 1's to fit the model (see theory of functional principal component logit regression).

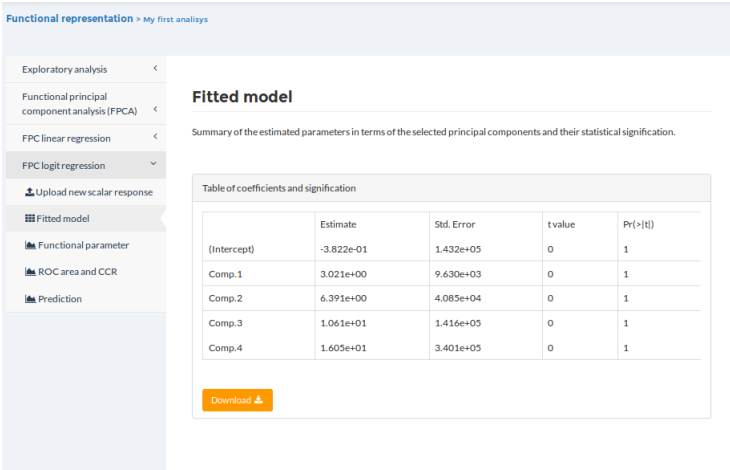
The user must upload the file with the response. **It is important the file format to be correct.** The file must be a .csv with only one column and no names in rows or columns. **It is important the number of rows to be the same as the file of discretized curves has.** The system returns a warning if it detects differences in both files.

0  
1  
1  
1  
1  
1  
0  
0  
0

The user must also select the number of functional principal components to use as covariates of the model.

## Functional principal component logit regression: Fitted model

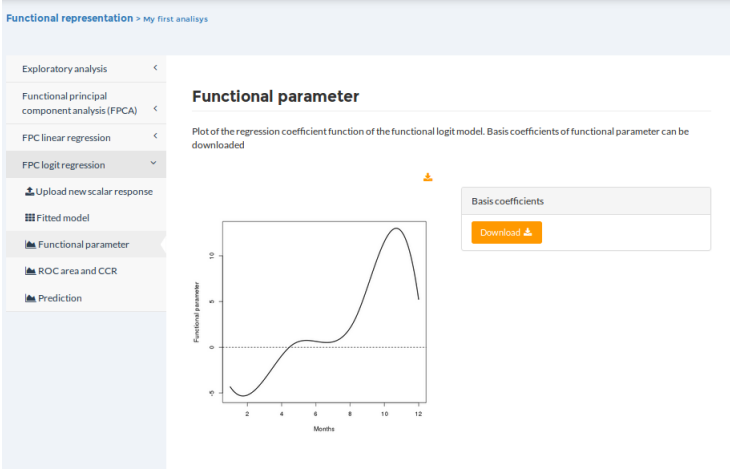
In this section, the application provide the fitted model of the functional principal component logit regression model. More exactly the table of parameters corresponding to each functional principal component.



The user can download this table more completely in .txt format, in the form that usually return R software.

## Functional principal component logit regression: Functional parameter

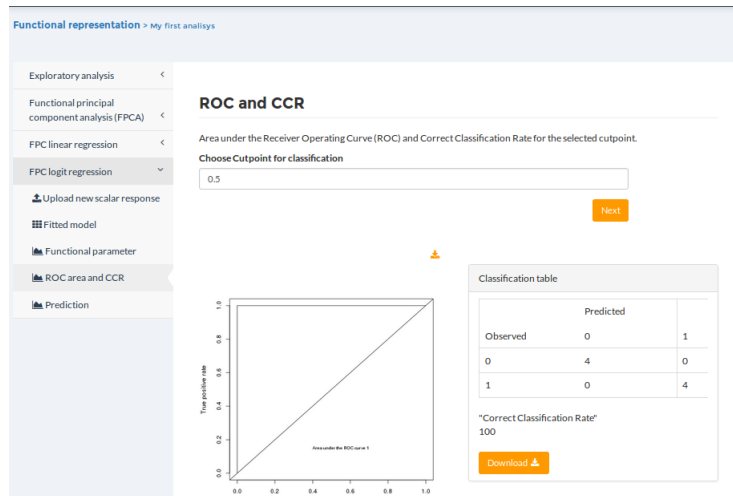
In this section, the application provide a plot of the functional parameter of the functional principal component logit regression model, and the vector of basis coefficients of this functional parameter.



The user can download the plot and the vector of basis coefficients in .csv format with columns separated by comma and with period as decimal point.

## Functional principal component logit regression: ROC area and CCR

In this section, the application provide the plot of the ROC curve, and the classification table and CCR. The cutpoint for the classification table and CCR can be changed by user. Every time the cutpoint is changed, the classification table changes.



The user can download the plot and the table in .txt format.

## Functional principal component logit regression: prediction

Functional principal component logit regression prediction needs a new curve to get the prediction that the model gives.

The user must upload the file with the new curve. **It is important the file format to be correct.** The file must be a .csv with only one row with no names in rows or columns. The system returns a error if the format is not correct. **For prediction the application is designed differently depending of the format of discretized curves previously uploaded. For cases 1 and 2, one file with one row and as many columns of discretized curves uploaded has is necessary**

-4.65, -5.33, -2.53, 1.26, 5.79, 10.79, 15.21, 15.28, 11.62, 7.02, 2.95, -1.85

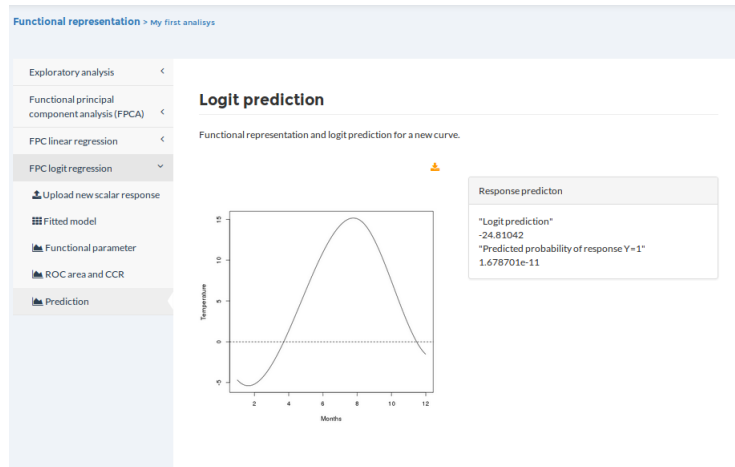
**For case 3 there are necessary two files: one with the row of discretized curve**

-12.78, -11.28, -4.67, 3.29, 10.91, 16.38, 19.05, 17.65

**and the other one with the sampling points. The number of columns of both files must be the same**

2, 3, 5, 6, 7, 8, 10, 11

The result of the analysis is the plot of the new curve and the prediction of the response



The user can download the plot.

## The history

The application provides a history page where the user can use previous analysis made without uploading data again.

Description	Lower bound	Upper bound	Data (csv)	Time points (csv)	Linear regression response (by col)	Logit regression response (by col)	Date
My first analysys	1.0	12.0	<a href="#">Download</a>	-	<a href="#">Download</a>	<a href="#">Download</a>	24-07-2016

It is important to note that the number of functional principal components to use in functional principal component linear and logit regression do not keep saved in the history, so the user must select this number every time he/she wants to get one of these analysis. If user do not actualize this parameter the models will be fitted in terms of the first functional principal component only.

## Important information

Remember that usually .csv files are automatically open with Excell in Windows systems and Open office Calc in Linux Systems. Remember also that Excell and Open office Calc use comma as decimal point, so if user wants to use these data, he/she will have to replace periods by commas. The search/replace ability of these programs can help to do it.