Master's Thesis

A goodness–of–fit test for the functional linear model with scalar response

Eduardo García Portugués

Supervised by Wenceslao González Manteiga and Manuel Febrero Bande

Master in Statistical Techniques

University of Santiago de Compostela

June 2012

A goodness–of–fit test for the functional linear model with scalar response

Eduardo García Portugués

Supervised by Wenceslao González Manteiga and Manuel Febrero Bande

Prefacio

En este proyecto de fin de máster se presenta una aportación nueva al campo de los datos funcionales. Esta idea surge a partir de los conocimientos adquiridos al cursar las asignaturas "Contrastes de especificación" y "Datos funcionales" del Máster en Técnicas Estadísticas. En particular, el trabajo realizado para esta última materia en el curso 2011/2012 representa el punto de partida de este proyecto. El presente documento recoge la maduración de esas ideas y su implementación en el paquete `fda.usc`.

Cabe destacar que el documento ha sido redactado en inglés para su uso en una futura publicación y que su versión en formato de artículo está públicamente disponible en el repositorio arXiv: `http://arxiv.org/abs/1205.6167`.

Abstract

In this work, a goodness–of–fit test for the null hypothesis of a functional linear model with scalar response is proposed. The test is based on a generalization to the functional framework of a previous one, designed for the goodness–of–fit of regression models with multivariate covariates using random projections. A simulation study illustrates the finite sample properties of the test for several types of basis and under different alternatives. Finally, the test is applied to two datasets for checking the assumption of the functional linear model.

Keywords: Functional data; Goodness–of–fit; Functional linear model; Bootstrap calibration.

Acknowledgements

I would like to thank the supervisors of my Master's thesis, Prof. Wenceslao González Manteiga and Prof. Manuel Febrero Bande, for their support, advice and valuable contributions. I would also like to especially thank Prof. Rosa M. Crujeiras for her guidance in my research. I also acknowledge the interesting contributions that Prof. César Sánchez Sellero has made in several aspects of this work.

I acknowledge the support of the FPU grant AP2010–0957 from the Spanish Ministry of Education, Project MTM2008–03010 from the Spanish Ministry of Science and Innovation and Project 10MDS207015PR from Dirección Xeral de I+D, Xunta de Galicia.

Contents

Chapter 1

Introduction

Functional data analysis has grown in popularity for the last years due to the increasingly data availability for continuous time processes. Typical examples of functional data include the temperature evolution, stock prices and path trajectories for objects in movement. New statistical methods have been developed to deal with the richer nature of functional data, being Ramsay and Silverman (2005), Ferraty and Vieu (2006) and Ferraty and Romain (2011) some of the main reference books in this area.

In many situations, the functional data is related to a scalar variable. For this cases, it is interesting to assess the relation of the variables via a regression model, which can be used to predict the scalar response from the functional input. Analogue to the multivariate situation, the simplest functional regression model corresponds to the functional linear model with scalar response (see Ramsay and Silverman (2005) for a review).

An interesting methodology approach to deal with functional data is the use of random projections. The objective is to characterize the behaviour of a functional process, which has infinite dimension, via the behaviour of the one dimensional inner products of the functional process with suitable random functions. This method has interesting applications for the goodness–of–fit of the distribution of the process, as it can be seen in Cuesta-Albertos et al. (2007). More recently, Patilea et al. (2012) provide a projection–based test for functional covariate effect in a functional regression model with scalar response. In their paper, the authors adapt the tests of Zheng (1996) and Lavergne and Patilea (2008), based on smoothing techniques, to the context of functional covariates.

In this work, a first goodness–of–fit test for the null hypothesis of the functional linear model, $H_0 : m \in \{\langle\cdot, \beta\rangle : \beta \in \mathbb{H}\}$, being $\mathbb{H}$ the Hilbert space of square integrable functions, is proposed. The statistic test is of a Cramér–von Mises type and is based on a generalization of a previous test of Escanciano (2006) designed for the case of a regression model with multivariate covariates. The test statistic is easy to compute using geometrical arguments and simple to calibrate in its distribution by a wild bootstrap on the residuals. Further, although the test is given for the functional linear model, it can be extended to other functional models with scalar response, as it is based on the residuals of the model.

This work is organized as follows. Some background on functional data, the functional linear model and the random projections paradigm are introduced on Chapter 2. The main part of the work is Chapter 3, where the theoretical arguments of the test, jointly with the bootstrap calibration procedure, are presented. The finite sample properties of the test are illustrated by a simulation study in Chapter 4. Chapter 5 illustrates the application of the test to two datasets and introduces a graphical tool to evaluate the goodness–of–fit of the functional linear model with scalar response. Possible extensions of the test are discussed in Chapter 6 and conclusions are given in Chapter 7. Finally, Appendix A presents the reference manual of the contributed code to the **R** package `fda.usc` (see Febrero-Bande and Oviedo de la Fuente (2012)).

Chapter 2

Background

The main goal of this work is to propose a goodness–of–fit test for the null hypothesis of the functional linear model with scalar response. Bearing in mind the different nature of the functional variables, some background on functional data, the functional linear model and the use of random projections is introduced.

2.1 Functional data

One of the first and most important problems when we deal with functional data is to choose a suitable functional space to work. The most used functional spaces are the metric, the Banach and the Hilbert spaces. This is a sequence of functional spaces with increasing richer structure, where the tools available for the former space are included in the latter. Specifically, in a metric space we can measure distances between functions; in addition, in a Banach space we can also measure the functions and Cauchy sequences are convergent; and finally, in a Hilbert space we have inner product, which allows to consider functional basis.

While there are a lot of types of metrics and norm spaces, the L^p spaces are one of the most used. The $L^p[0,1]$ space, $1 \leq p < \infty$, is defined as the set of all functions $f : [0,1] \to \mathbb{R}$ such that their norm $||f||_p$ is finite, where

$$||f||_p = \left(\int_0^1 |f(t)|^p \, dt \right)^{\frac{1}{p}}.$$

The choice of the interval $[0,1]$ is done only to fix the integration limits and other intervals can be considered without major changes. The most important L^p space corresponds to $p = 2$, because is the only which has an associated inner product $\langle \cdot, \cdot \rangle$ such that $||f||_p = \langle f, f \rangle^{\frac{1}{2}}$. For two functions $f, g \in L^2[0,1]$, their inner product is defined as

$$\langle f, g \rangle = \int_0^1 f(t)g(t) \, dt.$$

In what follows we will consider as our working space the Hilbert space $\mathbb{H} = L^2[0,1]$, bearing in mind that $[0,1]$ can be trivially replaced by another interval. The inner product allows for a basis representation of the elements of $\mathbb{H}$ and, given a functional basis $\{\Psi_j\}_{j=1}^{\infty}$ of $\mathbb{H}$, then any function $\mathcal{X}$ in $\mathbb{H}$ can be expressed by the linear combination:

$$\mathcal{X} = \sum_{j=1}^{\infty} x_j \Psi_j,$$

where $x_j = \langle \mathcal{X}, \Psi_j \rangle$, $j \geq 1$. A basis is said to be orthogonal if $\langle \Psi_i, \Psi_j \rangle = 0$, $i \neq j$ and orthonormal if, in addition, $\langle \Psi_j, \Psi_j \rangle = 1$, $j \geq 1$. Typical examples of basis of $\mathbb{H}$ are the Fourier basis, $\{1, \sin(2\pi jx), \cos(2\pi jx)\}_{j=1}^{\infty}$ and the B–splines basis (see de Boor (2001)).

For the development of the test statistic, we will also need to introduce a p–truncate basis $\{\Psi_j\}_{j=1}^{p}$, which corresponds to the first p elements of the infinite basis $\{\Psi_j\}_{j=1}^{\infty}$. The representation of $\mathcal{X}$ in this truncated basis is denoted by

$$\mathcal{X}^{(p)} = \sum_{j=1}^{p} x_j \Psi_j.$$

We will denote by $\mathbf{x}$ and by $\mathbf{x}_p$ the vector of coefficients of $\mathcal{X}$ in the original and in the p–truncated basis, respectively.

The choice of the number of basis elements p is crucial to have a reliable representation of the function $\mathcal{X}$ by $\mathcal{X}^{(p)}$. Although there exists several methods to select an appropriate p, we will refer to the GCV criteria (see Ramsay and Silverman (2005), page 97) to select p and represent adequately the function $\mathcal{X}$ in $\{\Psi_i\}_{i=1}^{p}$. This criteria will be used in Section 4.1 to select a suitable p for the case of the simple hypothesis.

To deal with functional random projections we will need to define the functional analogue of the euclidean p–sphere $\mathbb{S}^p = \{\mathbf{x} \in \mathbb{R}^p : ||\mathbf{x}||_{\mathbb{R}^p} = 1\}$. In the functional case we have the *functional sphere* of $\mathbb{H}$, defined as $\mathbb{S}_{\mathbb{H}} = \{f \in \mathbb{H} : ||f||_{\mathbb{H}} = 1\}$, and the *functional sphere of dimension p*, which is the set of functions of $\mathbb{H}$ that, expressed in the p–truncated basis, have unit norm: $\mathbb{S}_{\mathbb{H}}^p = \left\{ f = \sum_{j=1}^{p} x_j \Psi_j \in \mathbb{H} : ||f||_{\mathbb{H}} = 1 \right\}$.

The relationship between $\mathbb{S}^p$ and $\mathbb{S}_{\mathbb{H}}^p$ is particularly interesting to develop the test. Let be $\boldsymbol{\Psi} = \left(\langle \Psi_i, \Psi_j \rangle\right)_{ij}$ the matrix of inner products of the p–truncated basis, $\mathbb{S}_{\boldsymbol{\Psi}}^p = \left\{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^T \boldsymbol{\Psi} \mathbf{x} = 1\right\}$ the p–ellipsoid generated by this matrix and $\mathbf{R}^T\mathbf{R}$ the Cholesky decomposition of $\boldsymbol{\Psi}$ (a semi–positive matrix). First of all, we have the trivial isomorphism that maps elements of $\mathbb{S}_{\mathbb{H}}^p$ to elements of $\mathbb{S}_{\boldsymbol{\Psi}}^p$ by means of the functional coefficients: $\phi : f = \sum_{j=1}^{p} x_j \Psi_j \in \mathbb{S}_{\mathbb{H}}^p \mapsto \phi(f) = \mathbf{x} \in \mathbb{S}_{\boldsymbol{\Psi}}^p$. Recall that functions ϕ and ϕ^{-1} are well defined because

$$||f||_{\mathbb{H}}^2 = \left\langle \sum_{j=1}^{p} x_j \Psi_j, \sum_{j=1}^{p} x_j \Psi_j \right\rangle = \mathbf{x}^T \boldsymbol{\Psi} \mathbf{x}.$$

We must consider also a linear transformation from $\mathbb{S}^p$ to $\mathbb{S}_{\boldsymbol{\Psi}}^p$, which is given by $\rho : \mathbf{x} \in \mathbb{S}^p \mapsto \rho(\mathbf{x}) = \mathbf{R}^{-1}\mathbf{x} \in \mathbb{S}_{\boldsymbol{\Psi}}^p$ and whose Jacobian is $|\mathbf{R}|^{-1}$, the determinant of the matrix $\mathbf{R}^{-1}$.

Using these two transformations, the integration of a functional operator T with respect to a functional covariate $\gamma^{(p)}$ in $\mathbb{S}_{\mathbb{H}}^p$ can be reduced to a real integration on the p–sphere:

$$\int_{\mathbb{S}_{\mathbb{H}}^p} T\left(\gamma^{(p)}\right) d\gamma^{(p)} = \int_{\mathbb{S}_{\boldsymbol{\Psi}}^p} T\left(\sum_{j=1}^{p} g_j \Psi_j \right) d\mathbf{g}_p = \int_{\mathbb{S}^p} |\mathbf{R}|^{-1} T\left(\sum_{j=1}^{p} \left(\mathbf{R}^{-1}\mathbf{g}\right)_j \Psi_j \right) d\mathbf{g}_p. \tag{2.1}$$

In the case where the basis is orthonormal, $\boldsymbol{\Psi}$ and $\mathbf{R}$ are the identity matrix of order p. Then the coefficients of $\gamma^{(p)} \in \mathbb{S}_{\mathbb{H}}^p$ in the basis $\{\Psi_j\}_{j=1}^{p}$ belong to $\mathbb{S}^p$ without any transformation.

2.2 Functional linear model

Suppose that $\mathcal{X}$ is a functional random variable in $\mathbb{H}$ and Y is a real random variable. If both variables are centred, i.e., $\mathbb{E}\left[\mathcal{X}(t)\right] = 0$ for a.e. $t \in [0,1]$ and $\mathbb{E}\left[Y\right] = 0$, the Functional Linear Model (FLM) with scalar response claims for the following relation:

$$Y = \langle \mathcal{X}, \beta \rangle + \varepsilon = \int \mathcal{X}(t)\beta(t)\,dt + \varepsilon,$$

where the functional parameter β belongs to $\mathbb{H}$ and ε is a random variable with zero mean, variance σ^2 and such that $\mathbb{E}\left[\mathcal{X}(t)\varepsilon\right] = 0$, $\forall t$. The prediction of Y is done with the conditional expectation of Y given $\mathcal{X}$:

$$m(\mathcal{X}) = \mathbb{E}\left[Y|\mathcal{X}\right] = \langle \mathcal{X}, \beta \rangle .$$

Saying that $(\mathcal{X}, Y)$ share the functional linear model is equivalent to saying that the regression function of Y on $\mathcal{X}$, m, belongs to the family $\mathcal{M} = \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$.

Given a sample $(\mathcal{X}_1, Y_1), \ldots, (\mathcal{X}_n, Y_n)$, the estimation of the functional parameter can be done by minimising the Residual Sum of Squares (RSS):

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{H}} \sum_{i=1}^{n} \left(Y_i - \langle \mathcal{X}_i, \beta \rangle\right)^2 .$$

A possible method to search for the parameter β that minimises the RSS is representing the functional data and the functional parameter in the truncated functional basis $\{\Psi_j\}_{j=1}^{p_{\mathcal{X}}}$ and $\{\theta_j\}_{j=1}^{p_\beta}$, respectively:

$$\mathcal{X}_i = \sum_{j=1}^{p_{\mathcal{X}}} c_{ij}\Psi_j,\ \beta = \sum_{j=1}^{p_\beta} b_j\theta_j,\ i = 1, \ldots, n.$$

Using the vector notation $\mathbf{x} = (\mathcal{X}_i)_i$, $\mathbf{C} = (c_{ij})_{ij}$, $\boldsymbol{\psi} = (\Psi_j)_j$, $\mathbf{b} = (b_j)_j$ and $\boldsymbol{\theta} = (\theta_j)_j$, the previous representation can be expressed as $\mathbf{x} = \mathbf{C}\boldsymbol{\psi}$ and $\beta = \boldsymbol{\theta}^T\mathbf{b}$. The functional linear model results in

$$Y = \langle \mathcal{X}, \beta \rangle + \epsilon \approx \mathbf{CJb} + \epsilon = \mathbf{Zb} + \epsilon, \tag{2.2}$$

where $\mathbf{J} = \left(\langle \Psi_i, \theta_j \rangle\right)_{ij}$. Then, basis representation allows to express the FLM as a standard linear regression, where the estimated coefficients of β in the basis $\{\theta_j\}_{j=1}^{p_\beta}$ are given by $\hat{\mathbf{b}} = (\mathbf{Z}^T\mathbf{Z})^{-1}$ $\mathbf{Z}^T\mathbf{Y}$. Although different combinations of $\{\Psi_j\}_{j=1}^{p_{\mathcal{X}}}$ and $\{\theta_j\}_{j=1}^{p_\beta}$ are possible, the usual choice is $\{\Psi_j\}_{j=1}^{p} = \{\theta_j\}_{j=1}^{p}$, being $\{\Psi_j\}_{j=1}^{p}$ an orthogonal basis because in that case the matrix $\mathbf{J}$ is diagonal.

There are several alternatives to represent the functional process and estimate the parameter β in a truncated basis. For instance, a general review of the estimation based on the use of basis expansions such as Fourier series or B–splines can be found in the book by Ramsay and Silverman (2005) and also has been analysed by Cardot et al. (2003), Li and Hsing (2007) and Crambes et al. (2009), among others. The so called Functional Principal Component regression estimation (FPC) was proposed by Cardot et al. (1999) and also studied by Cardot et al. (2003), Hall and Hosseini-Nasab (2006) and Cai and Hall (2006), among others. The FPC provide an orthogonal data–driven basis that gives the most rapidly convergent representation of the functional dataset predictor when speed of convergence is defined in a L^2 sense (see Hall and Hosseini-Nasab (2006) and Hall and Horowitz (2007)). Preda and Saporta (2002) have proposed the Functional Partial Least Squares regression method (FPLS) that produces iteratively a sequence of orthogonal functions, as the FPC are, but with maximum predictive performance. In order to implement any of the methods shown before, it is required to fix the number of basis elements (or functional principal components, functional partial least squares components) that are used in the estimation.

The optimal number of components, p, has to be fixed based on the information provided by the data. To do this, Hall and Hosseini-Nasab (2006) and Preda and Saporta (2002) use the predictive cross–validation criterion (PCV), Cardot et al. (2003) and Ferraty and Romain (2011) consider the generalized cross–validation criterion (GCV) and Chiou and Müller (2007) and Febrero-Bande et al. (2010) consider those methods based on information approaches: the Akaike Information Criterion (AIC), the Corrected Akaike Information Criterion (AICc) and the Bayesian Information Criterion (BIC).

Let denote by $\hat{Y}_i^{(p)} = \left\langle \mathcal{X}_i^{(p)}, \hat{\beta}^{(p)} \right\rangle$ and $\hat{Y}_{i,(-i)}^{(p)} = \left\langle \mathcal{X}_i^{(p)}, \hat{\beta}_{(-i)}^{(p)} \right\rangle$ the prediction of Y_i using p components with the whole sample and with the whole sample excluding the i–th element, respectively. The PCV is defined as:

$$\mathrm{PCV}(p) = \arg\min_p \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i,(-i)}^{(p)} \right)^2 ,$$

which is computationally expensive because it involves the estimation of the $\hat{\beta}_{(-i)}^{(p)}$ n times. This is especially expensive in the case of data–driven basis (FPC, FPLS) because the basis has to be calculated for every datum. As an alternative, GCV avoids recalculating the $\hat{\beta}^{(p)}$ for every datum introducing a penalty term. The GCV is defined as

$$\mathrm{GCV}(p) = \arg\min_p \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^{(p)} \right)^2}{n \left(1 - \frac{df}{n} \right)}, \tag{2.3}$$

where df is the number of degrees of freedom consumed by the model. GCV is closely related with AIC, AICc and BIC although they come from different perspectives. For example, doing some simple calculations, it is easy to show that

$$n \log \mathrm{GCV}(p) = \mathrm{AIC}(p) + O(n^{-1}).$$

2.3 Random projections

Random projections are becoming quite popular when dealing with high dimensional data, as a way to overcome the well known *curse of the dimensionality*. The main idea behind is to reduce the dimension, and characterize the distribution of the multidimensional data by the distribution of the randomly projected data.

In the goodness–of–fit field, this is specially interesting, as the test procedures tend to become less efficient, less powerful, when the dimension of the model increases. Escanciano (2006) used this technique to develop a goodness–of–fit test for multivariate regression models based on random projections. According to his simulation study, their test has an excellent power performance and has the best empirical power for most situations when comparing to their competitors in the finite dimensional context.

In the functional framework, it is also possible to consider random projections. Usually, this is achieved by considering the inner product of the functional variable $\mathcal{X}$ of $\mathbb{H}$ and a suitable family of projectors, i.e. random functions γ in $\mathbb{H}$. For example, using with this approach Cuesta-Albertos et al. (2007) developed some goodness–of–fit tests for parametric families of functional distributions, which includes goodness–of–fit tests for Gaussianity and for the Black–Scholes model.

A very interesting result on projections can be found in Patilea et al. (2012). In their paper, the authors provide a characterization of the conditional expectation of a scalar variable Y with respect to a functional variable $\mathcal{X}$ given in terms of the conditional expectation of Y with respect to the projected $\mathcal{X}$. The result is stated here in the following lemma.

Lemma 1 (Patilea et al. (2012))**.** *Let Y be a random variable and $\mathcal{X}$ a functional random variable in the functional space $\mathbb{H}$. The following statements are equivalent:*

I. $\mathbb{E}\left[Y|\mathcal{X}=x\right]=0$*, for almost every (a.e.)* $x\in\mathbb{H}$*.*

II. $\mathbb{E}\left[Y|\left\langle\mathcal{X},\gamma\right\rangle=u\right]=0$*, for a.e.* $u\in\mathbb{R}$ *and* $\forall\gamma\in\mathbb{S}_{\mathbb{H}}$*.*

III. $\mathbb{E}\left[Y|\left\langle\mathcal{X},\gamma\right\rangle=u\right]=0$*, for a.e.* $u\in\mathbb{R}$ *and* $\forall\gamma\in\mathbb{S}_{\mathbb{H}}^{p}$*,* $\forall p\geq 1$*.*

Chapter 3

The test

The presentation of the goodness–of–fit test that we propose in this work is divided into three sections. The first and most important presents the theoretical fundamentals, with starting point in Lemma 2. The second shows the effective implementation of the test statistic in practise. Finally, the bootstrap resampling for the calibration of the test distribution is presented in the last section.

3.1 Theoretical arguments

Let Y be a real random variable and $\mathcal{X}$ a functional random variable in the space $\mathbb{H}$. Given a random sample $\{(\mathcal{X}_i, Y_i)\}_{i=1}^n$, we are interested in checking if a functional linear model is suitable to explain the relation between the functional covariate and the scalar response, i.e., test for the composite hypothesis:

$$H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\},$$

versus a general alternative of the form:

$$H_1 : \mathbb{P}\{m \notin \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}\} > 0.$$

Further, the simple hypothesis, i.e. checking for a specific functional linear model:

$$H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle, \text{ for a fixed } \beta_0 \in \mathbb{H},$$

is also of interest as it includes the important case of no interaction between the functional covariate and the scalar response (considering $\beta_0(t) = 0$, $\forall t$). In what follows we will focus on the procedure for the composite hypothesis, given that the simple is obtained in an easier way, just considering that the functional parameter is known and substituting $\hat{\beta}$ and $\hat{\beta}^{(p)}$ by β_0 and $\beta_0^{(p)}$, respectively.

The key point to test the null hypothesis H_0 is the following lemma, an adaptation of the Lemma 1 to our setting, which gives the characterization of H_0 in terms of the random projections of $\mathcal{X}$.

Lemma 2. *Let β be an element of $\mathbb{H}$. The following statements are equivalent:*

I. $m(\mathcal{X}) = \langle \mathcal{X}, \beta \rangle$, $\forall \mathcal{X} \in \mathbb{H}$.

II. $\mathbb{E}[Y - \langle \mathcal{X}, \beta \rangle \,|\mathcal{X} = x] = 0$, *for a.e.* $x \in \mathbb{H}$.

III. $\mathbb{E}[Y - \langle \mathcal{X}, \beta \rangle \,|\, \langle \mathcal{X}, \gamma \rangle = u] = 0$, *for a.e.* $u \in \mathsf{R}$ *and* $\forall \gamma \in \mathbb{S}_{\mathbb{H}}$.

IV. $\mathbb{E}\left[Y - \langle \mathcal{X}, \beta\rangle \mid \langle \mathcal{X}, \gamma\rangle = u\right] = 0$*, for a.e.* $u \in \mathsf{R}$ *and* $\forall \gamma \in \mathbb{S}^p_{\mathbb{H}}$, $\forall p \geq 1$.

V. $\mathbb{E}\left[(Y - \langle \mathcal{X}, \beta\rangle)\mathbb{1}_{\{\langle \mathcal{X}_i, \gamma\rangle \leq u\}}\right] = 0$*, for a.e.* $u \in \mathsf{R}$ *and* $\forall \gamma \in \mathbb{S}_{\mathbb{H}}$.

VI. $\mathbb{E}\left[(Y - \langle \mathcal{X}, \beta\rangle)\mathbb{1}_{\{\langle \mathcal{X}_i, \gamma\rangle \leq u\}}\right] = 0$*, for a.e.* $u \in \mathsf{R}$ *and* $\forall \gamma \in \mathbb{S}^p_{\mathbb{H}}$, $\forall p \geq 1$.

Proof of Lemma 2. Let β be an arbitrary element of $\mathbb{H}$. We will proceed by proving equivalences by pairs.

First of all, equivalence of I and II is immediately by the definition of $m(x) = \mathbb{E}\left[Y|\mathcal{X} = x\right]$. Equivalences of II, III and IV follow by Lemma 1.

The equivalence of III and V is based on the definition of the integrated regression function and is given by a chain of equivalences. Let denote $U_\gamma = \langle X, \gamma\rangle$, for any $\gamma \in \mathbb{S}_{\mathbb{H}}$, $m_\gamma(u) = \mathbb{E}\left[Y|U_\gamma = u\right]$ and $m_{0,\gamma}(u) = \mathbb{E}\left[\langle \mathcal{X}, \beta\rangle |U_\gamma = u\right]$. The integrated regression functions for m_γ and $m_{0,\gamma}$ are given by:

$$\begin{aligned}
I_\gamma(u) &= \mathbb{E}\left[Y\mathbb{1}_{\{U_\gamma \leq u\}}\right] = \mathbb{E}\left[\mathbb{E}\left[Y\mathbb{1}_{\{U_\gamma \leq u\}}|U_\gamma\right]\right] = \mathbb{E}\left[\mathbb{E}\left[Y|U_\gamma\right]\mathbb{1}_{\{U_\gamma \leq u\}}\right] \\
&= \mathbb{E}\left[m_\gamma(U_\gamma)\mathbb{1}_{\{U_\gamma \leq u\}}\right] = \int_{-\infty}^{\infty} m_\gamma(u)\mathbb{1}_{\{u \leq x\}}\,dF_\gamma(u) = \int_{-\infty}^{x} m_\gamma(u)\,dF_\gamma(u), \qquad (3.1) \\
I_{0,\gamma}(u) &= \mathbb{E}\left[\langle \mathcal{X}, \beta\rangle\mathbb{1}_{\{U_\gamma \leq u\}}\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \mathcal{X}, \beta\rangle\mathbb{1}_{\{U_\gamma \leq u\}}|U_\gamma\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \mathcal{X}, \beta\rangle|U_\gamma\right]\mathbb{1}_{\{U_\gamma \leq u\}}\right] \\
&= \mathbb{E}\left[m_{0,\gamma}(U_\gamma)\mathbb{1}_{\{U_\gamma \leq u\}}\right] = \int_{-\infty}^{\infty} m_{0,\gamma}(u)\mathbb{1}_{\{u \leq x\}}\,dF_\gamma(u) = \int_{-\infty}^{x} m_{0,\gamma}(u)\,dF_\gamma(u), \qquad (3.2)
\end{aligned}$$

where F_γ represents the distribution function of U_γ. Statement III can be expressed as

$$m_\gamma(u) = m_{0,\gamma}(u), \text{ for a.e. } u \in \mathbb{R},$$

which by (3.1) and (3.2) is equivalent to

$$I_\gamma(u) = I_{0,\gamma}(u), \text{ for a.e. } u \in \mathbb{R}. \qquad (3.3)$$

As V is equivalent to (3.3), this proofs the equivalence of III and V. The same argument can be applied to prove the equivalence between IV and VI, which ends the proof.

□

Then H_0 is characterized by the null value of the moment $\mathbb{E}\left[(Y - \langle \mathcal{X}, \beta\rangle)\mathbb{1}_{\{\langle \mathcal{X}_i, \gamma\rangle \leq u\}}\right]$, for a.e. $u \in \mathbb{R}$ and $\forall \gamma \in \mathbb{S}_{\mathbb{H}}$ (or $\forall \gamma \in \mathbb{S}^p_{\mathbb{H}}$, $\forall p \geq 1$) and a possible way to measure the deviation of the data from H_0 is by the empirical process arising from the estimation of this moment:

$$R_n(u, \gamma) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left(Y_i - \left\langle \mathcal{X}_i, \hat{\beta}\right\rangle\right) \mathbb{1}_{\{\langle \mathcal{X}_i, \gamma\rangle \leq u\}}, \qquad (3.4)$$

that will be denoted as the *Residual Marked empirical Process based on Projections* (RMPP). The marks of (3.4) are given by the residuals $\left\{Y_i - \langle \mathcal{X}_i, \hat{\beta}\rangle\right\}_{i=1}^n$ and the jumps by the projected functional regressor in the direction γ, $\{\langle \mathcal{X}_i, \gamma\rangle\}_{i=1}^n$. The estimation of β can be done by different methods as described in Chapter 2. Note that the RMPP only depends on the residuals of the model considered (in this case the residuals of the FLM) and therefore it can be easily extended to other regression models (see Chapter 6 for discussion).

To measure the distance of the empirical process (3.4) from zero, two possibilities are the classical Cramér–von Mises and Kolmogorov–Smirnov norms, adapted to the *projected* space $\Pi = \mathbb{R} \times \mathbb{S}_{\mathbb{H}}$:

$$\mathrm{PCvM}_n = \int_{\Pi} R_n(u,\gamma)^2 \, F_{n,\gamma}(du) \, \omega(d\gamma), \tag{3.5}$$

$$\mathrm{PKS}_n = \sup_{(u,\gamma)\in\Pi} |R_n(u,\gamma)|, \tag{3.6}$$

where $F_{n,\gamma}$ is the empirical cumulative distribution function (ecdf) of the projected functional data in the direction γ (i.e. the ecdf of the data $\{\langle \mathcal{X}_i, \gamma\rangle\}_{i=1}^n$) and ω represents a functional measure on $\mathbb{S}_{\mathbb{H}}$.

Unfortunately, the infinite dimension of the space $\mathbb{S}_{\mathbb{H}}$ makes infeasible to compute the functionals (3.5) and (3.6) and some kind of discretization is needed. A solution to this problem is to consider the properties of the Hilbert space $\mathbb{H}$ and use a basis representation.

Up to this end, let us introduce some notation. Let $\{\Psi_j\}_{j=1}^\infty$ be a basis of $\mathbb{H}$ and consider the basis representation of the functions $\mathcal{X}_i$ and γ as $\mathcal{X}_i = \sum_{j=1}^\infty x_{ij}\Psi_j$ and $\gamma = \sum_{j=1}^\infty g_j \Psi_j$, for $i = 1, \ldots, n$. For any integer $p \geq 1$, denote by $\mathcal{X}_i^{(p)} = \sum_{j=1}^p x_{ij}\Psi_j$ and $\gamma^{(p)} = \sum_{j=1}^p g_j\Psi_j$ the representation of the functions $\mathcal{X}_i$ and γ in the p–truncated basis $\{\Psi_j\}_{j=1}^p$, for $i = 1, \ldots, n$. Also, denote by $\boldsymbol{\Psi}$ the matrix of inner products of the p–truncated basis. The vectors of coefficients of $\mathcal{X}_i^{(p)}$ and $\gamma^{(p)}$ are denoted by $\mathbf{x}_{i,p} = (x_{i1}, \ldots, x_{ip})$ and $\mathbf{g}_p = (g_1, \ldots, g_p)$, respectively. Using this, and bearing in mind that $\{\Psi_j\}_{j=1}^\infty$ is any basis, we have that

$$\left\langle \mathcal{X}_i^{(p)}, \gamma^{(p)} \right\rangle = \mathbf{x}_{i,p}^T \, \boldsymbol{\Psi} \, \mathbf{g}_p.$$

By analogy with the previously defined $F_{n,\gamma}$, we will denote $F_{n,\gamma^{(p)}}$ to the ecdf of the projected functional data expressed in the p–truncated basis, both for the projector γ and for the functional data. Then, the RMPP can be expressed in terms of a p–truncated basis, yielding

$$\begin{aligned} R_{n,p}\left(u, \gamma^{(p)}\right) &= n^{-\frac{1}{2}} \sum_{i=1}^n \left(Y_i - \left\langle \mathcal{X}_i^{(p)}, \hat{\beta}^{(p)} \right\rangle\right) \mathbb{1}_{\left\{\left\langle \mathcal{X}_i^{(p)}, \gamma^{(p)}\right\rangle \leq u\right\}} \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n \left(Y_i - \mathbf{x}_{i,p}^T \, \boldsymbol{\Psi} \, \mathbf{b}_p\right) \mathbb{1}_{\left\{\mathbf{x}_{i,p}^T \, \boldsymbol{\Psi} \, \mathbf{g}_p \leq u\right\}} \\ &= R_{n,p}\left(u, \mathbf{g}_p\right), \end{aligned}$$

where $\mathbf{b}_p$ represents the coefficients of $\hat{\beta}$ in the p–truncated basis $\{\Psi_j\}_{j=1}^p$.

Bearing in mind this, our test statistic propose is a modified version of PCvM_n that results from expressing all the functions in a p–truncated basis of $\mathbb{H}$:

$$\mathrm{PCvM}_{n,p} = \int_{\mathbb{S}_{\mathbb{H}}^p \times \mathbb{R}} R_{n,p}\left(u, \gamma^{(p)}\right)^2 F_{n,\gamma^{(p)}}(du) \, \omega(d\gamma^{(p)}). \tag{3.7}$$

We have decided to choose the Cramér–von Mises statistic because, as we will see, presents important computational advantages and can be adapted to the given framework of Escanciano (2006) for the finite dimensional case. The most important advantage is that we can derive an explicit expression where there is no need to compute the RMPP for different projections, property that does not hold for the Kolmogorov–Smirnov statistic.

Using that the integration in the p–sphere of $\mathbb{H}$ can be expressed as the integration in the p–sphere of $\mathbb{R}^p$ via the transformations defined in Section 2.1 and the relation (2.1), we have:

$$\begin{aligned}
\mathrm{PCvM}_{n,p} &= \int_{\mathbb{S}^p_{\boldsymbol{\Psi}}\times\mathbb{R}} R_{n,p}(u,\mathbf{g}_p)^2\, F_{n,\mathbf{g}_p}(du)\,\omega(d\mathbf{g}_p)\\
&= \int_{\mathbb{S}^p\times\mathbb{R}} |\mathbf{R}|^{-1}\, R_{n,p}(u,\mathbf{R}^{-1}\mathbf{g}_p)^2\, F_{n,\mathbf{R}^{-1}\mathbf{g}_p}(du)\,\omega(d\mathbf{g}_p)\\
&= \int_{\mathbb{S}^p\times\mathbb{R}} |\mathbf{R}|^{-1}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\left(Y_i-\mathbf{x}_{i,p}^T\,\boldsymbol{\Psi}\,\mathbf{b}_p\right)\mathbb{1}_{\left\{\mathbf{x}_{i,p}^T\,\mathbf{R}^T\,\mathbf{g}_p\le u\right\}}\right)^2 F_{n,\mathbf{R}^{-1}\mathbf{g}_p}(du)\,\omega(d\mathbf{g}_p),
\end{aligned}$$

where ω now represents a measure in the p–sphere $\mathbb{S}^p$ that, for simplicity purposes, will be considered as the uniform distribution on $\mathbb{S}^p$. Thus, our simplified version of the statistic (3.7) is:

$$\mathrm{PCvM}_{n,p} = \int_{\mathbb{S}^p\times\mathbb{R}} |\mathbf{R}|^{-1}\, R_{n,p}(u,\mathbf{R}^{-1}\mathbf{g}_p)^2\, F_{n,\mathbf{R}^{-1}\mathbf{g}_p}(du)\,d\mathbf{g}_p. \tag{3.8}$$

Essentially, what we have done is to treat the functional process as a p–multivariate process, expressing the functions in a basis of p elements. The methods to choose the number of elements p and to estimate the parameter β both for the simple and for the composite hypothesis are the ones introduced in Chapter 2. These methods will be illustrated in the simulation study of Chapter 4.

3.2 Implementation

Following the steps of Escanciano (2006) it is possible to derive a simpler expression for (3.8). Using the definition of the RMPP in a p–truncate basis and the fact that $F_{n,\mathbf{R}^{-1}\mathbf{g}_p}$ is the ecdf of $\left\{\mathbf{x}_{i,p}^T\boldsymbol{\Psi}\mathbf{R}^{-1}\mathbf{g}_p\right\}_{i=1}^n = \left\{\mathbf{x}_{i,p}^T\mathbf{R}^T\mathbf{g}_p\right\}_{i=1}^n$, by simple algebra:

$$\begin{aligned}
\mathrm{PCvM}_{n,p} &= \int_{\mathbb{S}^p\times\mathbb{R}} |\mathbf{R}|^{-1}\, R_{n,p}(u,\mathbf{R}^{-1}\mathbf{g}_p)^2\, F_{n,\mathbf{R}^{-1}\mathbf{g}_p}(du)\,d\mathbf{g}_p\\
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\varepsilon}_i\hat{\varepsilon}_j\int_{\mathbb{S}^p\times\mathbb{R}} |\mathbf{R}|^{-1}\,\mathbb{1}_{\left\{\mathbf{x}_{i,p}^T\mathbf{R}^T\mathbf{g}_p\le u\right\}}\mathbb{1}_{\left\{\mathbf{x}_{j,p}^T\mathbf{R}^T\mathbf{g}_p\le u\right\}}\, F_{n,\mathbf{R}^{-1}\mathbf{g}_p}(du)\,d\mathbf{g}_p\\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\hat{\varepsilon}_i\hat{\varepsilon}_j\int_{\mathbb{S}^p} |\mathbf{R}|^{-1}\,\mathbb{1}_{\left\{\mathbf{x}_{i,p}^T\mathbf{R}^T\mathbf{g}_p\le\mathbf{x}_{r,p}^T\mathbf{R}^T\mathbf{g}_p\right\}}\mathbb{1}_{\left\{\mathbf{x}_{j,p}^T\mathbf{R}^T\mathbf{g}_p\le\mathbf{x}_{r,p}^T\mathbf{R}^T\mathbf{g}_p\right\}}\,d\mathbf{g}_p\\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\hat{\varepsilon}_i\hat{\varepsilon}_j A_{ijr},
\end{aligned}$$

with $\hat{\varepsilon}_i = Y_i - \left\langle\mathcal{X}_i^{(p)},\hat{\beta}^{(p)}\right\rangle$. The terms A_{ijr} represent the integrals

$$\begin{aligned}
A_{ijr} &= \int_{\mathbb{S}^p} |\mathbf{R}|^{-1}\,\mathbb{1}_{\left\{\mathbf{x}_{i,p}^T\mathbf{R}^T\mathbf{g}_p\le\mathbf{x}_{r,p}^T\mathbf{R}^T\mathbf{g}_p\right\}}\mathbb{1}_{\left\{\mathbf{x}_{j,p}^T\mathbf{R}^T\mathbf{g}_p\le\mathbf{x}_{r,p}^T\mathbf{R}^T\mathbf{g}_p\right\}}\,d\mathbf{g}_p\\
&= \int_{\mathbb{S}^p} |\mathbf{R}|^{-1}\,\mathbb{1}_{\{(\mathbf{R}\mathbf{x}_{i,p}-\mathbf{R}\mathbf{x}_{r,p})^T\mathbf{g}_p\le 0,\,(\mathbf{R}\mathbf{x}_{j,p}-\mathbf{R}\mathbf{x}_{r,p})^T\mathbf{g}_p\le 0\}}\,d\mathbf{g}_p\\
&= |\mathbf{R}|^{-1}\int_{S_{ijr}} d\mathbf{g}_p\\
&= |\mathbf{R}|^{-1}\,S\left(S_{ijr}\right),
\end{aligned}$$

where $S_{ijr} = \left\{ \boldsymbol{\xi} \in \mathbb{S}^p : \frac{\pi}{2} \leq \measuredangle \left(\mathbf{x}'_{i,p} - \mathbf{x}'_{r,p}, \boldsymbol{\xi} \right) \leq \frac{3\pi}{2}, \frac{\pi}{2} \leq \measuredangle \left(\mathbf{x}'_{j,p} - \mathbf{x}'_{r,p}, \boldsymbol{\xi} \right) \leq \frac{3\pi}{2} \right\}$, $S(S_{ijr})$ represents the surface area of S_{ijr} and $\measuredangle(\mathbf{a}, \mathbf{b})$ represents the angle between vectors $\mathbf{a}$ and $\mathbf{b}$. To simplify notation, we denote $\mathbf{x}'_{k,p} = \mathbf{R}\mathbf{x}_{k,p}$ ($\mathbf{x}'_{k,p} = \mathbf{x}_{k,p}$ if the basis is orthonormal) for $k = 1, \ldots, n$. Depending on $\mathbf{x}'_{i,p}$, $\mathbf{x}'_{j,p}$, $\mathbf{x}'_{r,p}$, the region S_{ijr} can be the whole sphere $\mathbb{S}^p$ ($\mathbf{x}'_{i,p} = \mathbf{x}'_{j,p} = \mathbf{x}'_{r,p}$), a hemisphere of $\mathbb{S}^p$ ($\mathbf{x}'_{i,p} = \mathbf{x}'_{j,p}$, $\mathbf{x}'_{i,p} = \mathbf{x}'_{r,p}$ or $\mathbf{x}'_{j,p} = \mathbf{x}'_{r,p}$) or a spherical wedge (see Figure 3.1 for a graphical interpretation) of width angle given by

$$\left| \pi - \arccos \left(\frac{(\mathbf{x}'_{i,p} - \mathbf{x}'_{r,p})^T (\mathbf{x}'_{j,p} - \mathbf{x}'_{r,p})}{||\mathbf{x}'_{i,p} - \mathbf{x}'_{r,p}|| \cdot ||\mathbf{x}'_{j,p} - \mathbf{x}'_{r,p}||} \right) \right|. \tag{3.9}$$

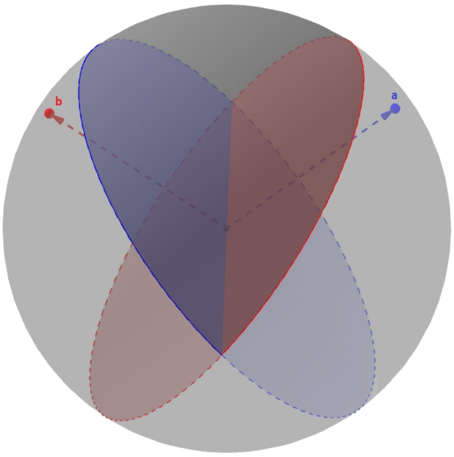

Figure 3.1: Spherical wedge $S_{\mathbf{a},\mathbf{b}} = \left\{ \boldsymbol{\xi} \in \mathbb{S}^p : \frac{\pi}{2} \leq \measuredangle(\boldsymbol{\xi}, \mathbf{a}) \leq \frac{3\pi}{2}, \frac{\pi}{2} \leq \measuredangle(\boldsymbol{\xi}, \mathbf{b}) \leq \frac{3\pi}{2} \right\}$ defined by points $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{S}^2$.

Thus A_{ijr} is the product of the surface area of a spherical wedge of angle $A_{ijr}^{(0)}$ times $|\mathbf{R}|^{-1}$, and is given by

$$A_{ijr} = A_{ijr}^{(0)} \frac{\pi^{p/2-1}}{\Gamma\left(\frac{p}{2}+1\right)} |\mathbf{R}|^{-1},$$

where $A_{ijr}^{(0)}$ is given by

$$A_{ijr}^{(0)} = \begin{cases} 2\pi, & \mathbf{x}'_{i,p} = \mathbf{x}'_{j,p} = \mathbf{x}'_{r,p}, \\ \pi, & \mathbf{x}'_{i,p} = \mathbf{x}'_{j,p}, \mathbf{x}'_{i,p} = \mathbf{x}'_{r,p} \text{ or } \mathbf{x}'_{j,p} = \mathbf{x}'_{r,p}, \\ (3.9), & \text{else.} \end{cases}$$

We also have a symmetric property, $A_{ijr} = A_{jir}$, which simplifies the evaluation of the test statistic from $O(n^3)$ to $O\left(\frac{n^3+n^2}{2}\right)$ computations. The memory requirement is expensive, because we need store the $\frac{n^3+n^2}{2}$ elements of the three dimensional array $\mathbf{A}$, which is symmetric in its two first indexes. However, this requirement can be stretched if we consider the following expression for the statistic:

$$\mathrm{PCvM}_{n,p} = n^{-2} \hat{\boldsymbol{\varepsilon}}^T \mathbf{A}_{\bullet} \hat{\boldsymbol{\varepsilon}}, \tag{3.10}$$

where $\mathbf{A}_{\bullet} = \left(\sum_{r=1}^n A_{ijr}\right)_{ij}$ is a $n \times n$ matrix and $\hat{\boldsymbol{\varepsilon}}$ is the vector of the residuals. By the definition of $A_{ijr}^{(0)}$ and its symmetry in the first two entries, the matrix $\mathbf{A}_{\bullet}$ is symmetric and its diagonal terms are given by $(n+1)\pi$. Although the order of computations remains similar, $O\left(\frac{n^3-n^2}{2}\right)$, the memory required for storing the matrix $\mathbf{A}_\bullet$ is substantially lower and drops to $\frac{n^2-n+2}{2}$ elements. This fact improves drastically the time of computation of the statistic and allows to apply the test to larger datasets.

Again, let us remark that the expression derived for the $\mathrm{PCvM}_{n,p}$ statistic remains valid for any functional regression model with scalar response and not just for the FLM, as the expression is based on the residuals of the model.

3.3 Bootstrap resampling

To calibrate the distribution of (3.8), a wild bootstrap on the residuals is applied. This bootstrap procedure is consistent in the finite dimensional case, as it was shown in Stute et al. (1998), and is adequate to situations with potential heterocedasticity, quite common in functional data. The resampling process for the case of the composite hypothesis, given an initial estimation $\hat{\beta}$ of the functional parameter, is the following:

I. Construct the estimated residuals: $\hat{\varepsilon}_i = Y_i - \left\langle \mathcal{X}_i^{(p)}, \hat{\beta}^{(p)} \right\rangle$, $i = 1, \ldots, n$.

II. Draw independent random variables $V_1^*, \ldots, V_n^*$ satisfying

$$\mathbb{E}^*\left[V_i^*\right] = 0 \text{ and } \mathbb{E}^*\left[V_i^{*2}\right] = 1.$$

For example, if V^* is a discrete random variable with distribution weights

$$\mathbb{P}\left\{V^* = \frac{1-\sqrt{5}}{2}\right\} = \frac{5+\sqrt{5}}{10} \text{ and } \mathbb{P}\left\{V^* = \frac{1+\sqrt{5}}{2}\right\} = \frac{5-\sqrt{5}}{10},$$

we have the *golden section bootstrap.*

III. Construct the bootstrap residuals $\varepsilon_i^* = V_i^* \hat{\varepsilon}_i$, $i = 1, \ldots, n$.

IV. Set $Y_i^* = \left\langle \mathcal{X}_i^{(p)}, \hat{\beta}^{(p)} \right\rangle + \varepsilon_i^*$, $i = 1, \ldots, n$ and estimate $\beta^{*,(p)}$ for the sample $\{(\mathcal{X}_i, Y_i^*)\}_{i=1}^n$.

V. Obtain the estimated bootstrap residues $\hat{\varepsilon}_i^* = Y_i^* - \left\langle \mathcal{X}_i^{(p)}, \hat{\beta}^{*,(p)} \right\rangle$, $i = 1, \ldots, n$.

Then, the procedure to calibrate the test is the following. In step I we compute the test statistic with the residuals under H_0 using the implementation (3.10) of the previous section:

$$\mathrm{PCvM}_{n,p} = n^{-2} \hat{\boldsymbol{\varepsilon}}^T \mathbf{A}_\bullet \hat{\boldsymbol{\varepsilon}}.$$

Then repeat steps II–V for $b = 1, \ldots, B$, computing each time the bootstrap statistic

$$\mathrm{PCvM}_{n,p}^{*,b} = n^{-2} \hat{\boldsymbol{\varepsilon}}^{*,b,T} \mathbf{A}_\bullet \hat{\boldsymbol{\varepsilon}}^{*,b}$$

and estimate the p–value of the test by Monte Carlo: p–value $\approx \#\left\{\mathrm{PCvM}_{n,p} \leq \mathrm{PCvM}_{n,p}^{*,b}\right\}/B$. For computational efficiency, it is important to note that we do not have to compute again the matrix $\mathbf{A}_\bullet$ in the bootstrap replicates.

A very interesting fact of the FLM is that step V can be easily performed using the properties of the estimation of $\hat{\beta}^{(p)}$. From (2.2) it is clear that the vector of coefficients of $\hat{\beta}^{(p)}$ is estimated throughout $\hat{\mathbf{b}} = \left(\mathbf{Z}^T\mathbf{Z}\right)\mathbf{Z}^T\mathbf{Y}$. Then, the estimated bootstrap residuals, represented by the vector $\hat{\boldsymbol{\varepsilon}}^*$, can be obtained as

$$\hat{\boldsymbol{\varepsilon}}^* = \left(\mathbf{I}_p - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)\mathbf{Z}^T\right)\mathbf{Y}^*,$$

where $\mathbf{Y}^*$ is the vector of bootstrap responses given by step IV and $\mathbf{I}_p$ is the identity matrix of order p. The projection matrix $\left(\mathbf{I}_p - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)\mathbf{Z}^T\right)$ remains the same for all the bootstrap replicates, so it can be stored without the need of computing it again. Obtaining the residuals in this way implies a significative computational saving.

The bootstrap resampling in the case of the simple hypothesis is easier: just replace $\hat{\beta}^{(p)}$ by $\beta_0^{(p)}$ and omit steps IV and V, considering $\hat{\varepsilon}_i^* = \varepsilon_i^*$, $i = 1, \ldots, n$.

Chapter 4

Simulation study

To illustrate the finite sample properties of the proposed test, a simulation study was carried out for the simple and the composite hypotheses. Before starting with these two cases, we describe briefly the simulation setting.

The functional process considered for the functional covariate $\mathcal{X}$ is an Ornstein–Uhlenbeck process in $[0,1]$, which is the solution to the stochastic differential equation

$$d\mathcal{X}(t) = \theta(\mu(t) - \mathcal{X}(t))dt + \sigma dB(t), \tag{4.1}$$

where B is a Brownian motion, μ is the functional mean and θ and σ are positive parameters. This process corresponds to a Brownian motion with functional mean μ and covariance function given by

$$\mathrm{Cov}(\mathcal{X}(s), \mathcal{X}(t)) = \frac{\sigma^2}{2\theta} e^{-\theta(s+t)} \left(e^{2\theta \min(s,t)} - 1\right).$$

We have considered $\theta = \frac{1}{3}$, $\sigma = 1$ and the functional mean $\mu(t) = 0$, $\forall t \in [0,1]$. Figure 4.1 shows a set of 100 simulated observations of the functional process (4.1) and their representation in three different functional basis: B–splines, FPC and FPLS. As said before, the choice of the right basis type and number of elements is crucial to capture correctly the structure of the functional process.

Let us remark that all the functional data in this simulation study is represented in 201 equidistant points in the interval $[0,1]$ and that, in the following, the number of bootstrap replicates considered will be $B = 1000$ and the number of Monte Carlo replicates for determining the empirical sizes and powers will be $M = 1000$. The sample size, except otherwise stated, is $n = 100$. Lastly, in order to properly compare the effect of the kind of basis, the number of elements and the sample sizes, the initial seed for the random generation of the functional underlying process is the same for each model.

4.1 Testing for simple hypothesis

The simulation study for the simple hypothesis is centred on the case $H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle$, where $\beta_0(t) = 0$, $t \in [0,1]$. This is equivalent to test that the functional covariate $\mathcal{X}$ has no effect on the scalar response, i.e., test the null hypothesis $H_0 : m(\mathcal{X}) = 0$. There is an extensive collection of goodness–of–fit tests for finite dimensional covariates (see González-Manteiga and Crujeiras (2011)), although for the case of functional covariates the literature is more limited. Therefore, we will focus on the competing procedures of Delsol et al. (2011) and González-Manteiga et al.

(2012) to compare the different tests in terms of level and power. Let us describe briefly these two test statistics.

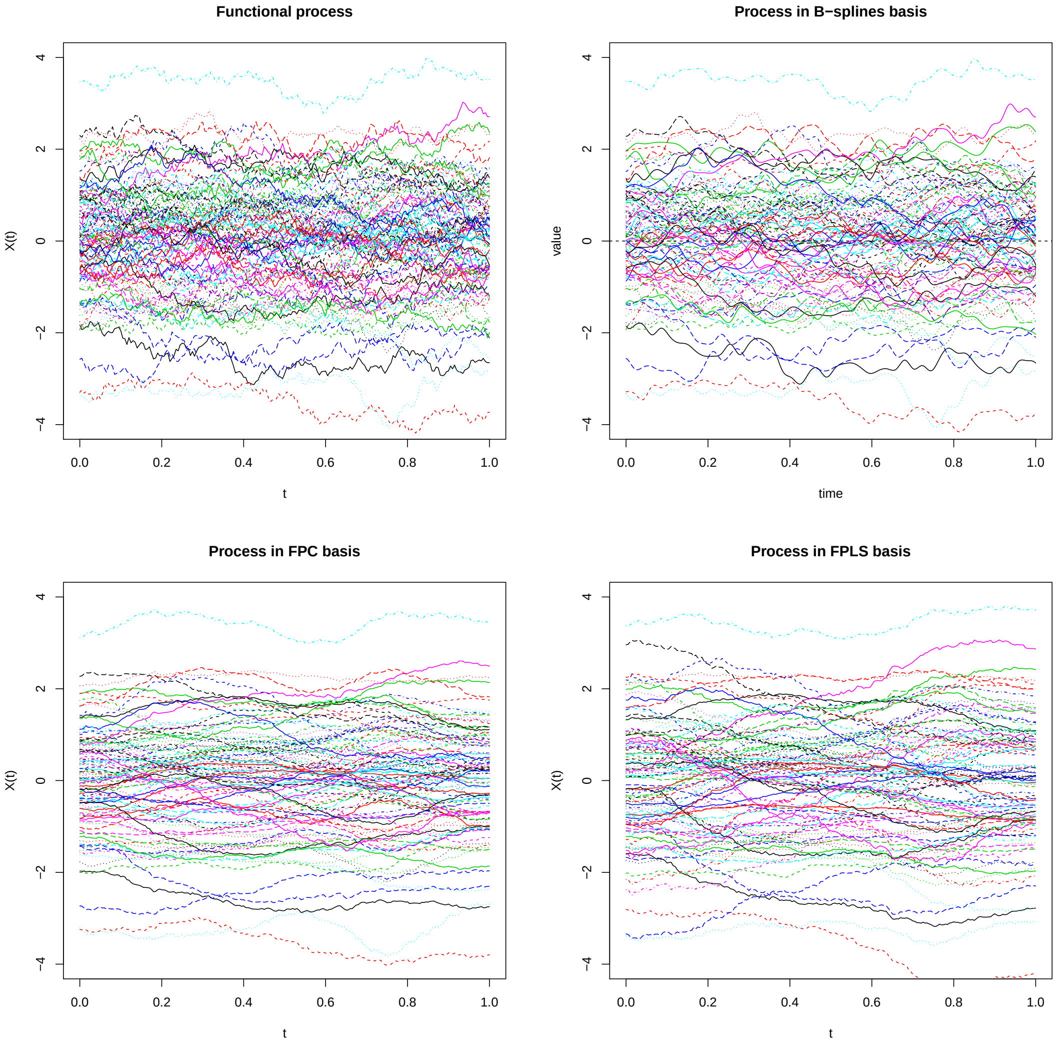

Figure 4.1: From up to down and left to right: simulated data process from the Ornstein–Uhlenbeck process (4.1); representation in a B–splines basis of 50 elements; representation in a FPC basis of 5 elements; representation in a FPLS basis of 5 elements, using an independent scalar response distributed as a $\mathcal{N}(0, 1)$.

Delsol et al. (2011) propose a test statistic for $H_0 : m(\mathcal{X}) = m_0(\mathcal{X})$, deriving its asymptotic law and giving a bootstrap procedure based on the residuals. The statistic, inspired in the propose of Härdle and Mammen (1993), is

$$T_n = \int \left(\sum_{i=1}^{n} (Y_i - m_0(\mathcal{X}_i)) K \left(\frac{d(\mathcal{X}, \mathcal{X}_i)}{h} \right) \right)^2 \omega(\mathcal{X}) dP_{\mathcal{X}}(\mathcal{X}),$$

where K is a kernel function, d is a semimetric and h is the bandwidth parameter. $P_{\mathcal{X}}$ represents the probability distribution of the functional process and ω is a suitable weight function. The test used in our implementation results from considering no functional effect, i.e. $H_0 : m_0(\mathcal{X}) = 0$, and from approximating the integral with respect to $dP_{\mathcal{X}}$ by the empirical mean of the sample:

$$T_n = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} Y_i K\left(\frac{d(\mathcal{X}_j,\mathcal{X}_i)}{h}\right)\right)^2 \omega(\mathcal{X}_j).$$

We have also considered the kernel $K(t) = 2\phi(|t|)$, $t \in \mathbb{R}$, being ϕ the density of a $\mathcal{N}(0,1)$, the L^2 distance in $\mathbb{H}$ for d and the uniform weight function. For the crucial choice of the bandwidth parameter we have considered the grid of bandwidths 0.25, 0.50, 0.75 and 1.00. Implementation of bootstrap resampling was done using golden wild bootstrap.

The other competing test is the one proposed by González-Manteiga et al. (2012) and is based on the idea of extending the covariance to functional–scalar data:

$$D_n = \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{X}_i - \bar{\mathcal{X}}\right)\left(Y_i - \bar{Y}\right)\right\|_{\mathbb{H}},$$

where $\bar{\mathcal{X}}$ is the functional mean of $\{\mathcal{X}_i\}_{i=1}^n$ and is $\bar{Y}$ the usual scalar mean of $\{Y_i\}_{i=1}^n$. The authors extend the ideas of the classical F–test to the functional framework, resulting a statistic to test the null hypothesis of no interaction *inside* the functional linear model. The test is consistent and the authors derived the asymptotic distribution of the process $\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{X}_i - \bar{\mathcal{X}}\right)\left(Y_i - \bar{Y}\right)$, resulting in a Brownian motion with mean $\mathbb{E}\left[(\mathcal{X}-\mu_{\mathcal{X}})(Y-\mu_Y)\right]$ and a particular covariance structure. This test can be viewed as a possible benchmark in our simulation study and, recalling its similarity with the classical F–test, will be denoted as the *functional F–test*. The bootstrap resampling was also performed using golden wild bootstrap.

As said before, the null hypothesis will be $H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta_0\rangle$, where $\beta_0(t) = 0$, $t \in [0,1]$. Two different kinds of deviations from the null are considered. The first one represents a deviation inside the linear model, i.e., considering different functions $\beta_{j,k}$, $j = 1,2$, $k = 1,2,3$, instead of β_0. The linear functions $\beta_{1,k}(t) = \gamma_k \cdot (t - 0.5)$, $k = 1,2,3$ with coefficients $\gamma_1 = 0.25$, $\gamma_2 = 0.65$ and $\gamma_3 = 1.00$ represent the first block of alternatives, $H_{1,k}$. The other block, $H_{2,k}$, is formed by the sinusoidal functions $\beta_{2,k}(t) = \eta_k \cdot \sin(2\pi t^3)^3$, $k = 1,2,3$, with $\eta_1 = 0.10$, $\eta_2 = 0.20$ and $\eta_3 = 0.50$. The upper row of Figure 4.2 shows the deviations of $\beta_{1,k}$ and $\beta_{2,k}$ from β_0.

The second kind of deviation from the null hypothesis consists on adding a *second order* term $\langle \mathcal{X}, \mathcal{X}\rangle$ to the regression function, thus the model is no longer linear. Different weights for the second term are represented in the alternatives

$$H_{3,k}: \quad Y = \langle \mathcal{X}, \beta_0\rangle + \delta_k \langle \mathcal{X}, \mathcal{X}\rangle + \varepsilon,$$

where $k = 1,2,3$ is the index for the deviation from the null and the deviation coefficients are $\delta_1 = 0.005$, $\delta_2 = 0.010$ and $\delta_3 = 0.015$. The difficulty to distinguish between the null hypothesis and the alternatives is reflected on the difference between the densities of the response (see Figure 4.2). As before, the larger the index of the deviation, the easier to distinguish from the null hypothesis and the more different the densities under the null and under the deviation are. The estimation of the densities of the response has been done with kernel smoothing from a sample of 1000 observations. The bandwidth is the same in the four densities of each model, and is computed by the method of Sheather and Jones (1991), for the case of the null hypothesis.

Further, we can measure the relation between the variance of the response with respect to the variance of the error using the following *signal–to–noise ratio*: $\mathrm{snr} = \sigma^2/\left(\sigma^2 + \mathbb{E}\left[m(\mathcal{X})^2\right]\right)$. For Model 1 the signal–to–noise ratios of the alternatives are 0.956, 0.765 and 0.579, respectively for $H_{1,k}$, $k = 1, 2, 3$. For Model 2, the snr's of the alternatives are 0.981, 0.850 and 0.671. For Model 3, we have 0.985, 0.914 and 0.728.

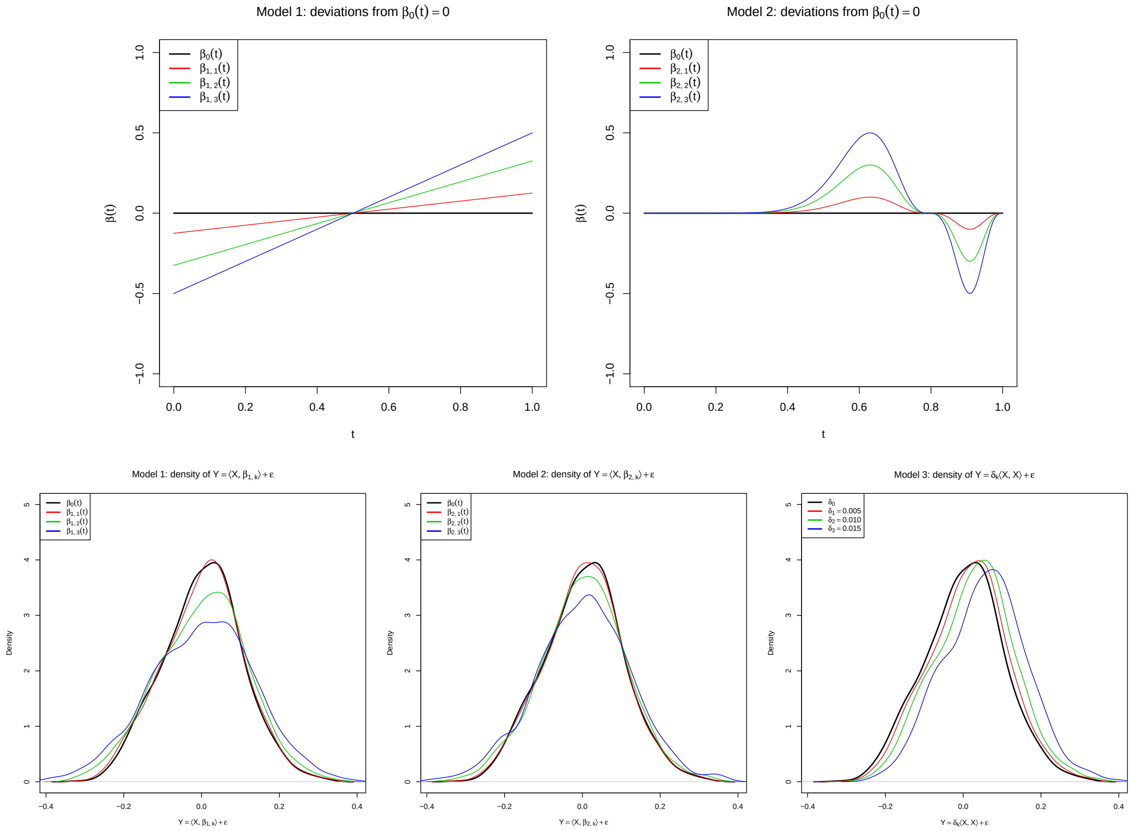

Figure 4.2: Upper row: functional coefficient deviations of the simple null hypothesis for $H_{1,k}$ (left) and $H_{2,k}$ (right), $k = 1, 2, 3$. Lower row: densities of the scalar response under the null hypothesis (H_0) and for the three deviations ($H_{j,k}$, $k = 1, 2, 3$, for each model $j = 1, 2, 3$).

In the case of the simple hypothesis there is no estimation of the parameter β_0, as it is known. However, it is necessary to express the functional process p and the function β_0 in a suitable basis in order to compute the test statistic. Up to this end, we consider a B–splines basis and we choose its number of elements by the GCV criteria commented in Section 2.1.

The results of the study for the simple hypothesis are collected in Tables 4.1, 4.2 and 4.3. Firstly, Table 4.1 shows the empirical sizes and powers of the functional F–test, the Delsol's test and the PCvM test for simple hypothesis, under the null hypothesis and for the three blocks of deviations from the null. The noise considered has a normal distribution with zero mean and standard deviation 0.10. All of the tests seem to calibrate well the significance level, $\alpha = 0.05$. With respect to the power, the functional F–test has in average a superior behaviour in the alternatives $H_{1,k}$ and $H_{2,k}$, $k = 1, 2, 3$, which represents deviations from the null *inside* the linear model. The test of Delsol performs also well, but the choice of the bandwidth has an important impact on the power performance (for example, in $H_{1,k}$ the bandwidth with more power is $h = 0.25$ but in $H_{2,k}$ is $h = 0.50$). As expected, the PCvM test performs worse than the functional F–test for alternatives $H_{1,k}$ and $H_{2,k}$ and similarly to the Delsol's test. Nevertheless, for alternatives that are not in the linear model, the functional F–test is not a benchmark any more, resulting the PCvM test the most powerful. Delsol's test also performs well, but seems to have less power.

Table 4.2 shows the same comparison of Table 4.1 but with a noise with a re–centred exponential distribution (i.e. the random variable $X - \lambda^{-1}$, where the density function of X is $\lambda e^{-\lambda x}$, $x > 0$) with parameter $\lambda^{-1} = 0.10$. The results are quite similar to Table 4.1, and show that the test for the simple hypothesis is robust with respect to a non symmetric random error.

Models	F–test	PCvM test	Delsol's test			
			$h = 0.25$	$h = 0.50$	$h = 0.75$	$h = 1.00$
H_0	0.058	0.043	0.055	0.050	0.047	0.041
$H_{1,1}$	0.063	0.069	0.070	0.069	0.069	0.066
$H_{1,2}$	0.166	0.079	0.358	0.115	0.063	0.055
$H_{1,3}$	0.422	0.137	0.819	0.245	0.087	0.066
$H_{2,1}$	0.253	0.053	0.068	0.078	0.068	0.055
$H_{2,2}$	0.952	0.336	0.314	0.447	0.392	0.273
$H_{2,3}$	1.000	0.904	0.795	0.887	0.870	0.775
$H_{3,1}$	0.036	0.173	0.051	0.096	0.116	0.126
$H_{3,2}$	0.057	0.691	0.207	0.361	0.457	0.523
$H_{3,3}$	0.056	0.998	0.802	0.928	0.956	0.977

Table 4.1: Empirical power of the competing tests for the simple hypothesis $H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle$, $\beta_0(t) = 0$, $\forall t$ and significance level $\alpha = 0.05$. Noise has a normal distribution with zero mean and standard deviation 0.10.

Models	F–test	PCvM test	Delsol's test			
			$h = 0.25$	$h = 0.50$	$h = 0.75$	$h = 1.00$
H_0	0.042	0.051	0.034	0.053	0.057	0.057
$H_{1,1}$	0.054	0.052	0.052	0.056	0.050	0.052
$H_{1,2}$	0.196	0.087	0.337	0.134	0.063	0.059
$H_{1,3}$	0.461	0.166	0.779	0.265	0.099	0.073
$H_{2,1}$	0.269	0.071	0.051	0.093	0.074	0.071
$H_{2,2}$	0.933	0.343	0.312	0.459	0.408	0.306
$H_{2,3}$	0.999	0.900	0.746	0.876	0.857	0.775
$H_{3,1}$	0.057	0.125	0.035	0.066	0.077	0.094
$H_{3,2}$	0.052	0.725	0.135	0.351	0.445	0.527
$H_{3,3}$	0.061	1.000	0.806	0.985	0.993	0.994

Table 4.2: Empirical power of the competing tests for the simple hypothesis $H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle$, $\beta_0(t) = 0$, $\forall t$ and significance level $\alpha = 0.05$. Noise has a centred exponential distribution with $\lambda^{-1} = 0.10$.

Finally, Table 4.3 gives the trace of the PCvM test for the simple hypothesis as a function of the number of FPC considered in the representation of the functional process. The trace is computed from one to six FPC, for the null hypothesis and for the *intermediate* deviations of the three models, i.e, $H_{j,2}$, $j = 1, 2, 3$. Unlike smoothing tests, the dependence on the equivalent of the smoothing parameter, the number of FPC components p, is very low. It turns out that, for the considered scenarios, there is no remarkable difference in terms of power and calibration for $p \geq 3$.

Models	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
H_0	0.035	0.037	0.037	0.036	0.036	0.037
$H_{1,2}$	0.043	0.098	0.085	0.080	0.074	0.074
$H_{2,2}$	0.529	0.407	0.375	0.359	0.350	0.347
$H_{3,2}$	0.657	0.703	0.707	0.708	0.709	0.709

Table 4.3: Empirical power of the PCvM test for the simple hypothesis $H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle$, $\beta_0(t) = 0$, $\forall t$, for different numbers p of FPC. The significance level is $\alpha = 0.05$ and noise has a normal distribution with zero mean and standard deviation 0.10.

4.2 Testing for composite hypothesis

To see the performance of the test under the composite hypothesis $H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$ we have considered three different null models of the form

$$H_{j,0}: \quad Y = \langle \mathcal{X}, \beta_j \rangle + \varepsilon, \tag{4.2}$$

with $j = 1, 2, 3$ being the index of the three different models. The functional coefficients of the three FLM are $\beta_1(t) = \sin(2\pi t) - \cos(2\pi t)$, $\beta_2(t) = t - (t - 0.75)^2$ and $\beta_3(t) = t + \cos(2\pi t)$, $t \in [0, 1]$. The first one corresponds to a difference between trigonometric functions that can not be perfectly represented in a B–splines basis. On the other hand, the second function is a polynomial of order two that can be exactly described by B–splines. The third one is the sum of a linear and a trigonometric function and is also not perfectly described in the B–splines basis. The upper row of Figure 4.3 shows these three functions.

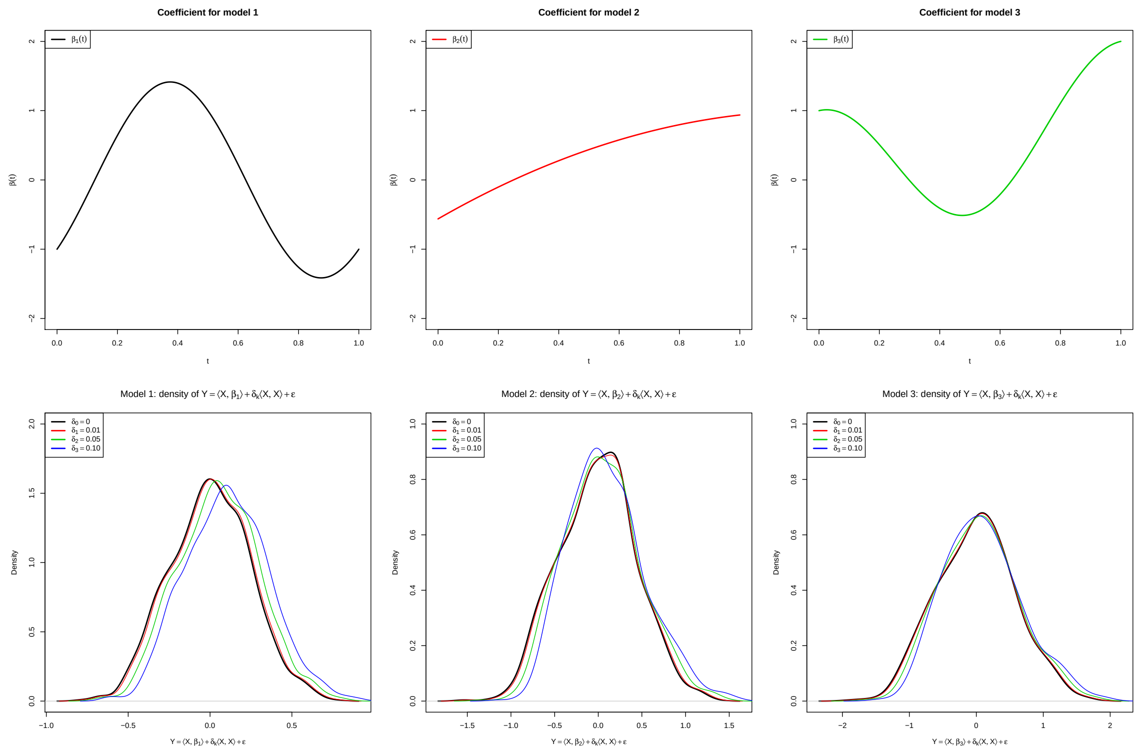

Figure 4.3: Upper row: functional coefficients of the linear models for the composite hypothesis. Lower row: densities of the scalar response under the null hypothesis ($H_{j,0}$, for each model $j = 1, 2, 3$) and for the three quadratic deviations ($H_{j,k}$, $k = 1, 2, 3$, for each model $j = 1, 2, 3$).

In order to check the power performance of the test, a set of possible deviations from the linear regression model is considered. Again, a second order term $\langle \mathcal{X}, \mathcal{X} \rangle$ is introduced to transform the model into a non–linear one. Three different weights for this term are considered, representing the alternatives $H_{j,k}$:

$$H_{j,k}: \quad Y = \langle \mathcal{X}, \beta_j \rangle + \delta_k \langle \mathcal{X}, \mathcal{X} \rangle + \varepsilon. \tag{4.3}$$

The index for the model is denoted by $j = 1, 2, 3$ and $k = 1, 2, 3$ is the index that measures the degree of the deviation from the null hypothesis. The weights of the quadratic term are $\delta_1 = 0.01$, $\delta_2 = 0.05$ and $\delta_3 = 0.10$. The lower row of Figure 4.2 shows how these deviations affect the densities of the scalar response, giving an idea of how difficult are to distinguish from the null hypothesis. The densities are computed in the way described for the simple hypothesis. The snr's for Model 1 are 0.177, 0.176, 0.166 and 0.140, respectively for $H_{1,k}$, $k = 0, 1, 2, 3$. For Model 2, the snr's are 0.050, 0.050, 0.050 and 0.047. For Model 3, we have 0.029, 0.029, 0.029 and 0.028.

Four estimation methods for the functional parameter β will be considered. The first three ones are designed in order to provide automatic selectors of the number of elements considered in the basis estimation of β. The fourth method is the FPC estimation described in Section 2.2, for a fixed number p of FPC. So, the first automatic method considered is the *optimal* representation in a B–splines basis of the functional process $\{\mathcal{X}_i\}_{i=1}^n$. The number of elements p in the basis is chosen by the GCV criteria (2.3) and then the β is estimated as a linear combination of p B–splines. Secondly, FPC estimation relies on the BIC criteria to choose the optimal number of elements in the FPC basis derived from the process $\{\mathcal{X}_i\}_{i=1}^n$ to estimate β. Finally, the FPLS method also uses PCV to select the adequate number of elements in the FPLS basis derived from the joint sample $\{(\mathcal{X}_i, Y_i)\}_{i=1}^n$.

	Coefficient estimation								
Models	B–splines estimation			FPC estimation			FPLS estimation		
	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
$H_{1,0}$	0.119	0.059	0.012	0.102	0.051	0.008	0.104	0.061	0.017
$H_{1,1}$	0.160	0.095	0.024	0.118	0.056	0.014	0.151	0.081	0.023
$H_{1,2}$	0.845	0.750	0.512	0.418	0.352	0.211	0.809	0.716	0.467
$H_{1,3}$	1.000	0.997	0.986	0.474	0.435	0.396	1.000	0.997	0.972
$H_{2,0}$	0.111	0.055	0.015	0.090	0.046	0.010	0.092	0.049	0.014
$H_{2,1}$	0.161	0.082	0.023	0.148	0.072	0.020	0.155	0.074	0.019
$H_{2,2}$	0.847	0.748	0.512	0.814	0.717	0.489	0.812	0.724	0.494
$H_{2,3}$	0.997	0.997	0.986	0.999	0.997	0.984	0.999	0.997	0.984
$H_{3,0}$	0.109	0.053	0.007	0.093	0.049	0.006	0.100	0.044	0.008
$H_{3,1}$	0.157	0.081	0.018	0.155	0.076	0.014	0.147	0.074	0.014
$H_{3,2}$	0.856	0.765	0.516	0.839	0.750	0.498	0.834	0.752	0.485
$H_{3,3}$	0.999	0.999	0.988	0.999	0.997	0.988	0.999	0.998	0.985

Table 4.4: Empirical power of the PCvM test for the composite hypothesis $H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$ and for three estimating methods of β. Noise has a normal distribution with zero mean and standard deviation 0.10.

Table 4.4 shows the rejection frequencies of the null hypothesis for the test computed from observations of the null models (4.2) and from models (4.3) (H_0 false), for the significance levels $\alpha = 0.10,\ 0.05,\ 0.01$. The rejection rates were computed for the three types of estimation of the functional coefficient and basis representation, in order to see the possible effects of estimation method in the power performance. At sight of the rejection frequencies for the three models, several comments must be done. Firstly, the test respects the significance levels for the null hypothesis in the three models considered. Secondly, as it is expected, the power increases when the alternatives are spreading apart from the null. Finally, in general, there seems to be no big differences in the rejection frequencies but for two exceptions: B–splines estimation tends to have slightly larger powers and, more notoriously, FPC estimation gives considerably lower rates in alternatives $H_{1,k}$.

As far as we know, this problem is caused by the way that optimal FPC are chosen. The optimal FPC obtained by the BIC criteria does not have to be ordered (in the sense of percentage of variance explained) and, for example, the components PC3, PC2 and PC4 could be the optimal to estimate β for predicting Y. However, these components could lead to a conservative test if they are not able to capture properly the deviations from the null, as they are less informative about the functional process. If instead of these FPC we consider PC1, PC2 and PC3, although the estimation of β is not optimal, the test will detect better the deviations from the null. This is clearly seen in the comparison of Tables 4.4 and 4.7. While for the former the empirical power for FPC is almost the half, for the latter the powers are quite similar to the other estimation methods. It is also important to note that for the FPLS estimation method this problem does not appear. The components of the FPLS are optimal to estimate the β but take into account the response Y, and therefore detect deviations from the model correctly. Therefore, for practical implementation of the test procedure is important to keep in mind this disadvantage of the FPC method, which can be solved by considering a fixed number of components p.

Analogously to the simple hypothesis, in Table 4.5 we show the rejection frequencies of the three estimation methods but with a non symmetric random noise. It turns out that the test for the composite hypothesis is robust with respect to a non symmetric random error.

	Coefficient estimation								
Models	B–splines estimation			FPC estimation			FPLS estimation		
	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
$H_{1,0}$	0.103	0.040	0.005	0.083	0.033	0.003	0.105	0.043	0.006
$H_{1,1}$	0.145	0.072	0.020	0.098	0.040	0.004	0.144	0.076	0.020
$H_{1,2}$	0.825	0.736	0.498	0.431	0.366	0.233	0.804	0.720	0.481
$H_{1,3}$	0.998	0.996	0.987	0.479	0.453	0.427	0.996	0.996	0.983
$H_{2,0}$	0.088	0.039	0.009	0.089	0.038	0.009	0.089	0.035	0.010
$H_{2,1}$	0.155	0.076	0.016	0.147	0.079	0.018	0.133	0.078	0.016
$H_{2,2}$	0.831	0.743	0.493	0.811	0.716	0.481	0.813	0.719	0.493
$H_{2,3}$	0.995	0.994	0.978	0.996	0.995	0.979	0.995	0.994	0.978
$H_{3,0}$	0.096	0.048	0.007	0.092	0.042	0.006	0.087	0.041	0.004
$H_{3,1}$	0.129	0.072	0.016	0.119	0.061	0.017	0.110	0.062	0.014
$H_{3,2}$	0.831	0.735	0.498	0.830	0.733	0.486	0.825	0.724	0.484
$H_{3,3}$	0.999	0.998	0.988	0.999	0.998	0.988	0.999	0.998	0.984

Table 4.5: Empirical power of the PCvM test for the composite hypothesis $H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$ and for three estimating methods of β. Noise has a centred exponential distribution with $\lambda^{-1} = 0.10$.

The behaviour of the test for different sample sizes is shown in Table 4.6. B–splines and FPLS estimation methods have very similar rejection ratios, although for the B–splines are slightly better. Further, when the sample sizes increases, the rejection rates also. Nevertheless, FPC estimation continues showing a bad behaviour in alternatives $H_{1,k}$ for the different sample sizes.

		Coefficient estimation								
Models	n	B–splines estimation			FPC estimation			FPLS estimation		
		α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
$H_{1,0}$	50	0.140	0.059	0.008	0.116	0.051	0.008	0.123	0.064	0.011
	100	0.119	0.059	0.012	0.102	0.051	0.008	0.104	0.061	0.017
	200	0.111	0.061	0.009	0.092	0.051	0.006	0.107	0.056	0.010
$H_{1,1}$	50	0.157	0.074	0.016	0.098	0.040	0.005	0.139	0.069	0.010
	100	0.160	0.095	0.024	0.118	0.056	0.014	0.151	0.081	0.023
	200	0.212	0.121	0.041	0.149	0.083	0.022	0.209	0.118	0.034
$H_{1,2}$	50	0.607	0.477	0.177	0.351	0.255	0.081	0.551	0.418	0.161
	100	0.845	0.750	0.512	0.418	0.352	0.211	0.809	0.716	0.467
	200	0.982	0.969	0.890	0.537	0.500	0.436	0.978	0.960	0.871
$H_{1,3}$	50	0.957	0.904	0.664	0.505	0.433	0.294	0.933	0.875	0.647
	100	1.000	0.997	0.986	0.474	0.435	0.396	1.000	0.997	0.972
	200	1.000	1.000	1.000	0.549	0.503	0.459	1.000	1.000	0.999

Table 4.6: Empirical power of the PCvM test for the composite hypothesis $H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$ and for different sample sizes n. Noise has a normal distribution with zero mean and standard deviation 0.10.

Finally, the determination of the effect of the number of basis elements in the power performance, more important in the case of the composite hypothesis than in the simple, is studied throughout the number of FPC considered in the estimation of β. Then, the following table shows the rejection ratios for different numbers p of FPC's in the first block of alternatives. Recall that when the number of FPC components is fixed at p, the components considered will be always the p first FPC: PC1, …, PCp. We can conclude that there is a moderate dependence of the power on the number of basis elements, with increasing power for larger p's.

Models	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$
$H_{1,0}$	0.044	0.053	0.049	0.056	0.060	0.062
$H_{1,1}$	0.054	0.062	0.079	0.079	0.085	0.089
$H_{1,2}$	0.192	0.410	0.685	0.743	0.755	0.757
$H_{1,3}$	0.577	0.911	0.996	0.997	0.997	0.997

Table 4.7: Empirical power of the PCvM test for the composite hypothesis $H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$, for different numbers p of FPC. The significance level is $\alpha = 0.05$ and noise has a normal distribution with zero mean and standard deviation 0.10.

Chapter 5

Data application and graphical tool

In this chapter we apply our proposed testing procedure to two datasets, in order to check if the FLM is enough well supported by the data. Further, a graphical tool to check the FLM model by means of the empirical process is provided.

The Tecator dataset is a well known dataset in the literature of functional data analysis (see, for example, Ferraty and Vieu (2006)). It contains data from 215 meat samples, consisting of a 100 channel spectrum of absorbances measured by a spectrometer and the contents of water, fat and protein. When trying to explain the content of fat in the meat samples throughout the spectrometric curves, it is common to transform the original curves into the first derivatives or the second derivatives, in order to properly capture the wavy effects in the absorbances of the meat samples with high percentage of fat (see Figure 5.1).

We have applied our goodness–of–fit test with $B = 5000$ bootstrap replicates for the original dataset and for the dataset of the first and second derivatives. The p–values obtained are 0.004, 0.000 and 0.000, respectively. Thus we have significative evidences against the null hypothesis of FLM. The test was applied with the FPLS estimation method and with automatic selection of the number of FPLS by PCV. As the case of no interaction is a particular case of a FLM, we can conclude that in the Tecator dataset there exists a significative dependence between the functional covariate and the scalar response, although this dependence is not a linear one.

The other dataset considered is the AEMET dataset, which is available in the **R** package `fda.usc` (see Febrero-Bande and Oviedo de la Fuente (2012)). It is formed by the daily summaries of 73 Spanish weather stations during the period 1980–2009. Among others, the functional covariate is the daily temperature in each weather station, and the scalar response is the daily wind speed (both variables are averaged over 1980–2009). Left plot of Figure 5.1 represents the functional observations of the daily temperature, where the lonely upper curves are the weather stations from the Canary Islands, a Spanish region with a warmer weather. Before applying the tests, four functional outliers corresponding to the 5% less depth curves according to the Fraiman and Muniz (2001) depth were removed.

The resulting p–value from the goodness–of–fit test is 0.121, thus there is no significative evidences to reject the null hypothesis of the FLM for the AEMET dataset. The test is applied with the FPLS estimation method and with $B = 5000$ bootstrap replicates. The right plot of Figure 5.1 shows the estimation of the functional parameter β, resulting from a basis of 2 FPLS. Once we have determined that the FLM is a suitable model, we can check if the estimated coefficient β is significantly different from zero with the available tests for the simple hypothesis: the functional F–test, the Delsol's test (with the grid of bandwidths corresponding to the quantiles

0.10, 0.15, 0.25, 0.50 and 0.75 of the L^2 distances of the functional data) and our test for the simple null hypothesis of no interaction. The p–values obtained are: 0.002, 0.000 (for all the bandwidths) and 0.062, respectively. The first two tests reject the null, whereas for the PCvM, it is accepted with a limiting p–value. At sight of this and the R^2 of the FLM, 0.42, we can conclude that the curves of the temperature and the average wind speed show a mild relation.

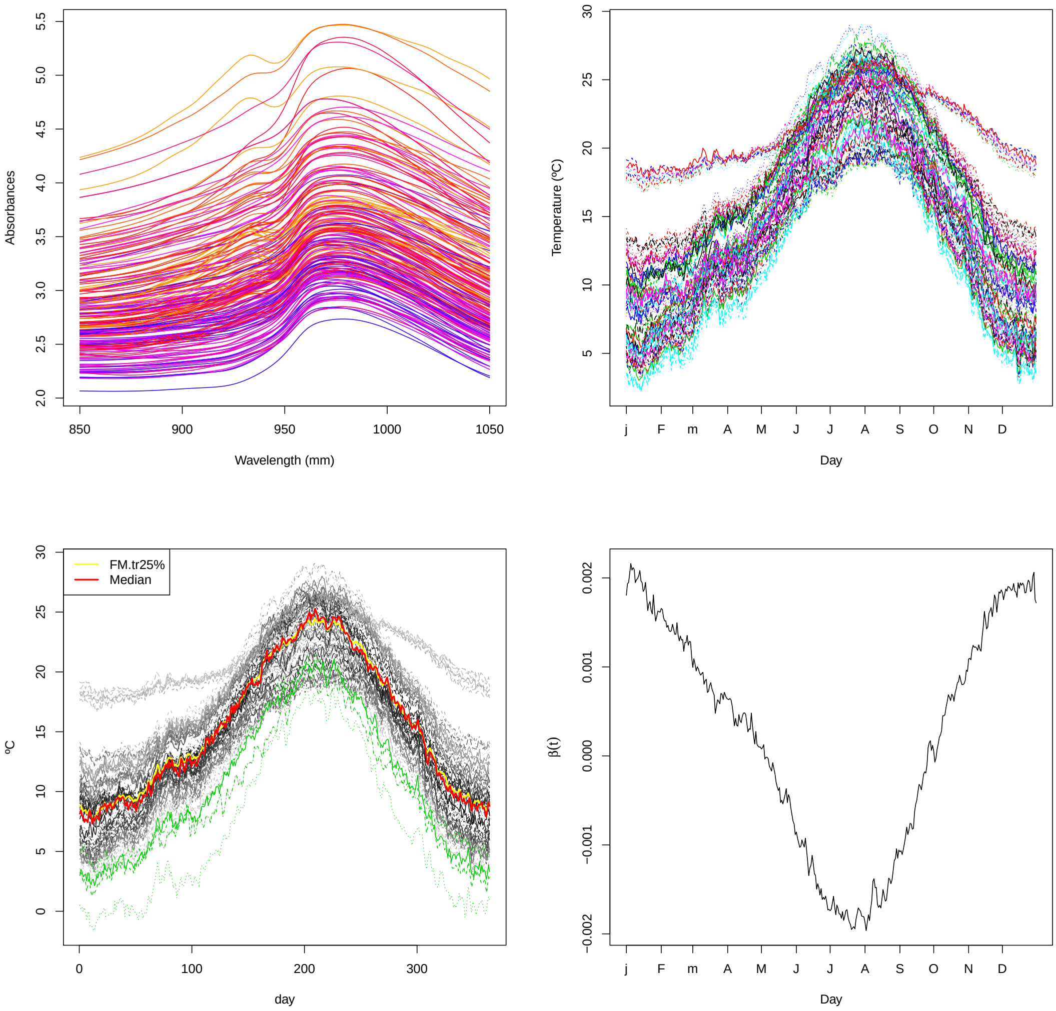

Figure 5.1: From up to down and left to right: Tecator dataset with spectrometric curves coloured according to their percentage of fat (red for larger content of fat and blue for lower); AEMET temperatures for the 73 Spanish weather stations; outliers of the temperature curves with respect to the Fraiman and Muniz (2001) depth; estimated functional coefficient by the FPLS method for the AEMET dataset (functional covariate is the average daily temperature and the scalar response is the average wind speed).

We conclude this chapter showing a graphical tool to visualize the goodness–of–fit of the FLM to a dataset that can be useful to practitioners. The key idea is to compare graphically the process (3.4) obtained with the residuals of the fitted model with the processes obtained with the bootstrapped residuals under the null hypothesis. The path of the RMPP depends on the random projections γ and therefore it is difficult to compare two trajectories of the process. However, integrating with respect to γ results a process that does not depend on the projections.

Further, this integration is easily approximated by Monte Carlo:

$$R_n(u) = \int_{\mathbb{S}_{\mathbb{H}}} R_n(u,\gamma)\,\omega(d\gamma) \approx \frac{1}{G}\sum_{g=1}^{G} R_n(u,\gamma_g),$$

being γ_g functions in $\mathbb{S}_{\mathbb{H}}$ and G the number of Monte Carlo replicates. For γ_g, a possibility is to consider stationary Gaussian processes with unit norm. Figure 5.2 shows the comparison of the observed process R_n and $B = 100$ bootstrapped processes under the null, for the two studied datasets. Consistently with the obtained p–values, the observed processes for the Tecator dataset seem to be significantly different, whereas for the AEMET dataset the observed process is just an *ordinary* trajectory of the bootstrapped ones.

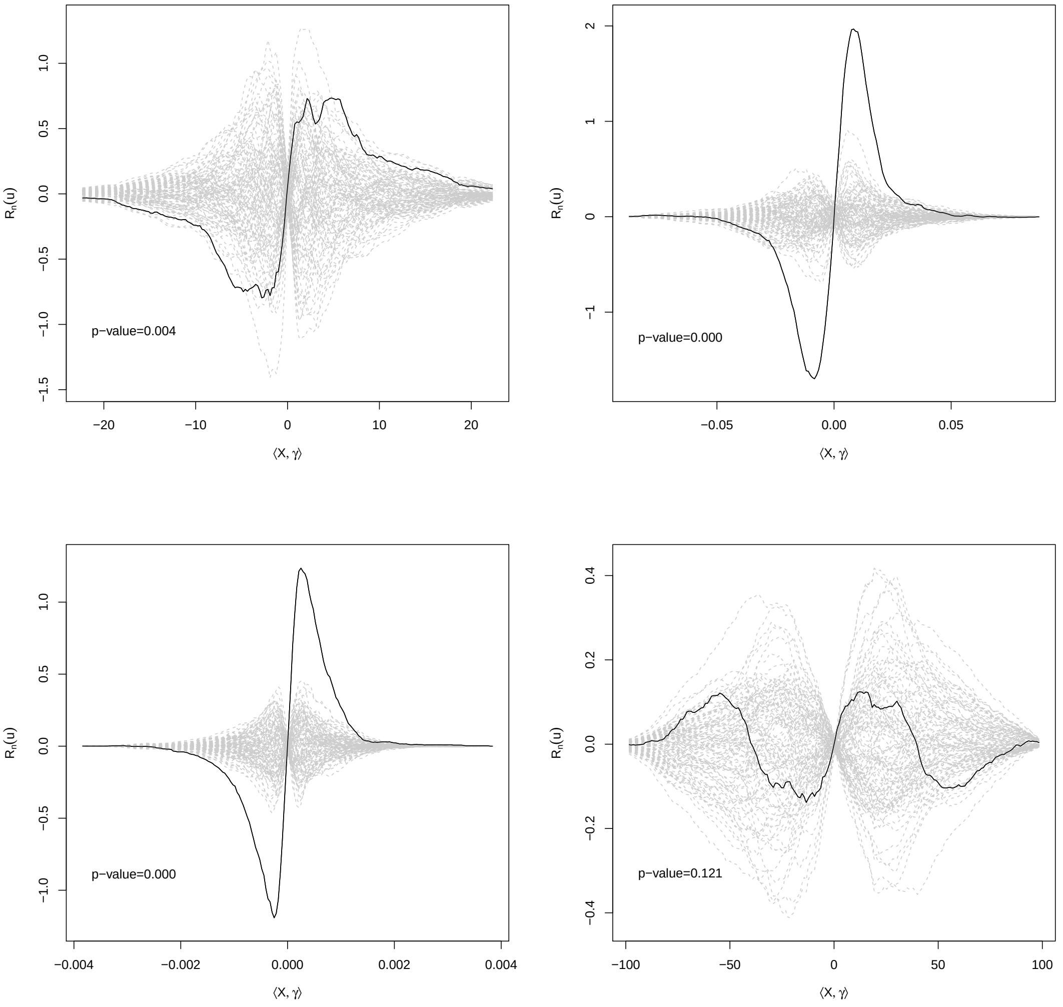

Figure 5.2: From up to down and left to right: R_n process observed (solid line) and $B = 100$ generated process under the null hypothesis $H_0 : m \in \{\langle\cdot,\beta\rangle : \beta \in \mathbb{H}\}$ (dashed lines), for the Tecator dataset, the Tecator dataset of considering the first and second derivatives of the curves and the AEMET dataset. The number of Monte Carlo replicates for the projections is $G = 200$.

Chapter 6

Extensions

So far we have focused on the testing the goodness–of–fit of the FLM, but, as we have pointed out in several times in this work, the procedure we present here can be extended to other regression models. This chapter is devoted to present and comment ideas for some possible extensions of the Projected Cramér–von Mises test statistic to different settings.

The first and more obvious expansion of our testing propose is the goodness–of–fit of any parametric functional regression model with scalar response:

$$Y = m_\theta(\mathcal{X}) + \varepsilon, \tag{6.1}$$

where m_θ is the parametric regression function of the random functional variable $\mathcal{X}$ (in $\mathbb{H}$) over the scalar random variable Y. m_θ depends on the parameter $\theta \in \Theta$ (that can be scalar, functional, ...) and ε is the random error that satisfies $\mathbb{E}\left[\varepsilon|\mathcal{X}\right] = 0$.

A general way to test if the model (6.1) holds is to check if $\mathbb{E}\left[\varepsilon|\mathcal{X}\right] = 0$ by fitting the parametric regression $m_{\hat{\theta}}$, computing the fitted residuals $\hat{\varepsilon}_i$, $i = 1, \ldots, n$ and examining $\mathbb{E}\left[\hat{\varepsilon}|\mathcal{X}\right]$. For example, this idea is used in Patilea et al. (2012) considering an adaptation of the test of Zheng (1996). The following result, based on Lemma 1 and with analogous proof to the one of Lemma 2, gives the characterization of the statement $\mathbb{E}\left[\varepsilon|\mathcal{X}\right] = 0$ by means of the *projected* integrated regression.

Lemma 3. *Let ε be a random variable and $\mathcal{X}$ a functional random variable in the functional space $\mathbb{H}$. The following statements are equivalent:*

I. $\mathbb{E}\left[\varepsilon|\mathcal{X} = x\right] = 0$*, for almost every (a.e.)* $x \in \mathbb{H}$.

II. $\mathbb{E}\left[\varepsilon \mathbb{1}_{\{\langle \mathcal{X}_i, \gamma\rangle \leq u\}}\right] = 0$*, for a.e.* $u \in \mathbb{R}$ *and* $\forall \gamma \in \mathbb{S}_{\mathbb{H}}$.

III. $\mathbb{E}\left[\varepsilon \mathbb{1}_{\{\langle \mathcal{X}_i, \gamma\rangle \leq u\}}\right] = 0$*, for a.e.* $u \in \mathbb{R}$ *and* $\forall \gamma \in \mathbb{S}_{\mathbb{H}}^p$, $\forall p \geq 1$.

Then, the RMPP in this general framework is

$$R_n(u, \gamma) = n^{-\frac{1}{2}} \sum_{i=1}^n \hat{\varepsilon}_i \mathbb{1}_{\{\langle \mathcal{X}_i, \gamma\rangle \leq u\}},$$

and we will work with the RMPP that arises from considering the functions $\{\mathcal{X}_i\}_{i=1}^n$ and the projections γ expressed in a p–truncate basis of $\mathbb{H}$ (for example FPC, FPLS or a B–splines basis):

$$R_{n,p}\left(u, \gamma^{(p)}\right) = n^{-\frac{1}{2}} \sum_{i=1}^n \hat{\varepsilon}_i \mathbb{1}_{\left\{\left\langle \mathcal{X}_i^{(p)}, \gamma^{(p)}\right\rangle \leq u\right\}}.$$

The closed expression (3.10) holds for the PCvM statistic of the marked empirical process and the wild bootstrap can be applied to approximate the distribution of the statistic under the null hypothesis. The differences with respect to the particular case of the FLM appear in the estimation of the parameter θ. Obviously, nor the expression $\hat{\varepsilon}_i = Y_i - \mathbf{x}_{i,p}\boldsymbol{\Psi}\mathbf{b}_p$ nor the specific bootstrap resampling of the FLM can be used with a generic parametric model.

Some interesting models that could be checked with this approach are the functional linear model with several functional covariates,

$$Y = \langle \mathcal{X}_1, \beta_1 \rangle + \cdots + \langle \mathcal{X}_r, \beta_r \rangle + \varepsilon$$

and the functional quadratic model:

$$\begin{aligned} Y &= \langle \mathcal{X}(t), \beta(t) \rangle + \langle \mathcal{X}(t), \langle \mathcal{X}(s), h(s,t) \rangle \rangle + \varepsilon \\ &= \int_0^1 \mathcal{X}(t)\beta(t)\, dt + \int_0^1 \int_0^1 \mathcal{X}(t)\mathcal{X}(s)h(s,t)\, ds\, dt + \varepsilon. \end{aligned}$$

In both of them, Y and $\mathcal{X}$ are scalar and functional centred variables, respectively.

The second extension is related with the testing of regression models with functional covariate and functional response:

$$\mathcal{Y}(t) = m_\theta(\mathcal{X}(s))(t) + \varepsilon(t),\ s,t \in [0,1], \tag{6.2}$$

where $\mathcal{Y}$ and $\mathcal{X}$ are random functional variables in $\mathbb{H}$ and ε now plays the role of a functional error, also in $\mathbb{H}$. Examples of this kind of models include the prediction of the daily temperature curve from the one of the previous day or the prediction of the price evolution of a financial asset from the evolution of other asset.

Our idea to check the model (6.2) is to consider two kinds of projections: for the regressor and for the response functions, denoted by $\gamma_\mathcal{X}$ and $\gamma_\mathcal{Y}$, respectively. Then the following marked empirical process can be considered:

$$R_n(u, \gamma_\mathcal{X}, \gamma_\mathcal{Y}) = n^{-\frac{1}{2}} \sum_{i=1}^n \langle \hat{\varepsilon}_i, \gamma_\mathcal{Y} \rangle \mathbb{1}_{\{\langle \mathcal{X}_i, \gamma_\mathcal{X} \rangle \leq u\}}.$$

The marks of the process are given by the *projected residuals* $\{\langle \hat{\varepsilon}_i, \gamma_\mathcal{Y} \rangle\}_{i=1}^n$ and the jumps by the projected functional regressor in the direction $\gamma_\mathcal{X}$. The PCvM statistic of this process is:

$$\mathrm{PCvM}_n = \int_{\mathbb{S}_\mathbb{H} \times \mathbb{S}_\mathbb{H} \times \mathbb{R}} R_n(u, \gamma_\mathcal{X}, \gamma_\mathcal{Y})^2\, F_{n,\gamma_\mathcal{X}}(du)\, \omega_\mathcal{X}(d\gamma_\mathcal{X})\, \omega_\mathcal{Y}(d\gamma_\mathcal{Y}),$$

with $\omega_\mathcal{X}$ and $\omega_\mathcal{Y}$ being suitable measures in the functional sphere $\mathbb{S}_\mathbb{H}$.

Following analogous ideas to those of the Chapter 3, we can consider the $p_\mathcal{X}$–basis $\left\{\Psi_k^\mathcal{X}\right\}_{k=1}^{p_\mathcal{X}}$ and the $p_\mathcal{Y}$–basis $\left\{\Psi_l^\mathcal{Y}\right\}_{l=1}^{p_\mathcal{Y}}$ for representing $\{\mathcal{X}_i\}_{i=1}^n$ and $\{\mathcal{Y}_i\}_{i=1}^n$, respectively. Then we have the following expression for the RMPP:

$$R_{n,p_\mathcal{X},p_\mathcal{Y}}\left(u, \gamma_\mathcal{X}^{(p_\mathcal{X})}, \gamma_\mathcal{Y}^{(p_\mathcal{Y})}\right) = n^{-\frac{1}{2}} \sum_{i=1}^n \left\langle \hat{\varepsilon}_i^{(p_\mathcal{Y})}, \gamma_\mathcal{Y}^{(p_\mathcal{Y})} \right\rangle \mathbb{1}_{\left\{\left\langle \mathcal{X}_i^{(p_\mathcal{X})}, \gamma_\mathcal{X}^{(p_\mathcal{X})} \right\rangle \leq u\right\}},$$

where the estimated functional residuals $\hat{\varepsilon}_i$ are expressed in the basis of the functional response. The PCvM statistic for this empirical process has again the advantage of having a closed expression if we consider the uniform measures on the hyperspheres $\mathbb{S}_\mathbb{H}^{p_\mathcal{Y}}$ and $\mathbb{S}_\mathbb{H}^{p_\mathcal{X}}$. Applying the relation (2.1) for the two functional integrals and with calculus analogous to the Section 3.2, we have:

$$\begin{aligned}
\mathrm{PCvM}_{n,p_{\mathcal{X}},p_{\mathcal{Y}}} &= \int_{\mathbb{S}_{\mathbb{H}}^{p_{\mathcal{Y}}}\times\mathbb{S}_{\mathbb{H}}^{p_{\mathcal{X}}}\times\mathbb{R}} R_{n,p_{\mathcal{X}},p_{\mathcal{Y}}}\left(u,\gamma_{\mathcal{X}}^{(p_{\mathcal{X}})},\gamma_{\mathcal{Y}}^{(p_{\mathcal{Y}})}\right)^2 F_{n,\gamma_{\mathcal{X}}^{(p_{\mathcal{X}})}}(du)\,\omega_{\mathcal{X}}\big(d\gamma_{\mathcal{X}}^{(p_{\mathcal{X}})}\big)\,\omega_{\mathcal{Y}}\big(d\gamma_{\mathcal{Y}}^{(p_{\mathcal{Y}})}\big)\\
&= \int_{\mathbb{S}^{p_{\mathcal{Y}}}\times\mathbb{S}^{p_{\mathcal{X}}}\times\mathbb{R}} |\mathbf{R}_{\mathcal{X}}|^{-1}\,|\mathbf{R}_{\mathcal{Y}}|^{-1}\,R_{n,p_{\mathcal{X}},p_{\mathcal{Y}}}\left(u,\mathbf{R}_{\mathcal{X}}^{-1}\mathbf{g}_{\mathcal{X},p_{\mathcal{X}}},\mathbf{R}_{\mathcal{Y}}^{-1}\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\right)^2\\
&\cdot F_{n,\mathbf{R}_{\mathcal{X}}^{-1}\mathbf{g}_{\mathcal{X},p_{\mathcal{X}}}}(du)\,d\mathbf{g}_{\mathcal{X},p_{\mathcal{X}}}\,d\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\\
&= \int_{\mathbb{S}^{p_{\mathcal{Y}}}\times\mathbb{S}^{p_{\mathcal{X}}}\times\mathbb{R}} |\mathbf{R}_{\mathcal{X}}|^{-1}\,|\mathbf{R}_{\mathcal{Y}}|^{-1}\left(n^{-\frac{1}{2}}\sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}^T\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\mathbb{1}_{\left\{\mathbf{x}_{i,p_{\mathcal{X}}}^T\mathbf{g}_{\mathcal{X},p_{\mathcal{X}}}\leq u\right\}}\right)^2\\
&\cdot F_{n,\mathbf{R}_{\mathcal{X}}^{-1}\mathbf{g}_{\mathcal{X},p_{\mathcal{X}}}}(du)\,d\mathbf{g}_{\mathcal{X},p_{\mathcal{X}}}\,d\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\\
&= n^{-2}\sum_{i=1}^n\sum_{j=1}^n\sum_{r=1}^n A_{ijr}\,|\mathbf{R}_{\mathcal{Y}}|^{-1}\int_{\mathbb{S}^{p_{\mathcal{Y}}}} \hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}^T\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\hat{\boldsymbol{\varepsilon}}_{j,p_{\mathcal{Y}}}^T\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\,d\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}},
\end{aligned}$$

where the subscripts $\mathcal{X}$ and $\mathcal{Y}$ stands for the terms related to that covariates and $\hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}$ is the vector of coefficients of $\hat{\varepsilon}_i^{(p_{\mathcal{Y}})}$ in the basis $\left\{\Psi_l^{\mathcal{Y}}\right\}_{l=1}^{p_{\mathcal{Y}}}$. The integral of the last term can be computed using some integration techniques on the $p_{\mathcal{Y}}$–sphere, yielding

$$\int_{\mathbb{S}^{p_{\mathcal{Y}}}} \hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}^T\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\hat{\boldsymbol{\varepsilon}}_{j,p_{\mathcal{Y}}}^T\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}}\,d\mathbf{g}_{\mathcal{Y},p_{\mathcal{Y}}} = \frac{\pi^{p_{\mathcal{Y}}/2-1}}{\Gamma\left(\frac{p_{\mathcal{Y}}}{2}+1\right)p_{\mathcal{Y}}}\hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}^T\hat{\boldsymbol{\varepsilon}}_{j,p_{\mathcal{Y}}}.$$

Joining these terms results the closed and easily computable expression of the statistic:

$$\begin{aligned}
\mathrm{PCvM}_{n,p_{\mathcal{X}},p_{\mathcal{Y}}} &= n^{-2}\sum_{i=1}^n\sum_{j=1}^n\sum_{r=1}^n A_{ijr}\,|\mathbf{R}_{\mathcal{Y}}|^{-1}\frac{\pi^{p_{\mathcal{Y}}/2-1}}{\Gamma\left(\frac{p_{\mathcal{Y}}}{2}+1\right)p_{\mathcal{Y}}}\hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}^T\hat{\boldsymbol{\varepsilon}}_{j,p_{\mathcal{Y}}}\\
&= |\mathbf{R}_{\mathcal{Y}}|^{-1}\frac{\pi^{p_{\mathcal{Y}}/2-1}}{\Gamma\left(\frac{p_{\mathcal{Y}}}{2}+1\right)p_{\mathcal{Y}}}n^{-2}\sum_{i=1}^n\sum_{j=1}^n(\mathbf{A}_\bullet)_{ij}\hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}^T\hat{\boldsymbol{\varepsilon}}_{j,p_{\mathcal{Y}}}\\
&= |\mathbf{R}_{\mathcal{Y}}|^{-1}\frac{\pi^{p_{\mathcal{Y}}/2-1}}{\Gamma\left(\frac{p_{\mathcal{Y}}}{2}+1\right)p_{\mathcal{Y}}}n^{-2}\sum_{l=1}^{p_{\mathcal{Y}}}\hat{\mathbf{E}}_{p_{\mathcal{Y}}}^{(l),T}\mathbf{A}_\bullet\hat{\mathbf{E}}_{p_{\mathcal{Y}}}^{(l)},
\end{aligned}$$

where $\hat{\mathbf{E}}_{p_{\mathcal{Y}}}$ is the $n\times p_{\mathcal{Y}}$ matrix whose i–th row is the vector of coefficients of $\hat{\varepsilon}_i^{(p_{\mathcal{Y}})}$ in the basis $\left\{\Psi_l^{\mathcal{Y}}\right\}_{l=1}^{p_{\mathcal{Y}}}$, $\hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}}$, and $\mathbf{B}^{(l)}$ stands for the l–th column of the matrix $\mathbf{B}$.

The simplest and most known parametric model of the form (6.2) is the functional linear model with functional response,

$$\mathcal{Y}(t) = \langle\mathcal{X}(s),\beta(s,t)\rangle+\varepsilon(t) = \int_0^1\mathcal{X}(s)\beta(s,t)\,ds+\varepsilon(t),\ s,t\in[0,1], \tag{6.3}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are centred functional variables in $\mathbb{H}$. The usual estimator of the bivariate function β is given by the following bilinear (see Ramsay and Silverman (2005)) combination

$$\hat{\beta}^{(p_{\mathcal{X}},p_{\mathcal{Y}})}(s,t) = \sum_{k=1}^{p_{\mathcal{X}}}\sum_{l=1}^{p_{\mathcal{Y}}} b_{kl}\Psi_k^{\mathcal{X}}(s)\Psi_l^{\mathcal{Y}}(t). \tag{6.4}$$

For this model, an explicit expression for $\hat{\mathbf{E}}_{p_{\mathcal{Y}}}$ can be derived. By simple algebra, $\hat{\boldsymbol{\varepsilon}}_{i,p_{\mathcal{Y}}} = \mathbf{y}_{i,p_{\mathcal{Y}}} - \mathbf{B}^T\boldsymbol{\Psi}^{\mathcal{X}}\mathbf{x}_{i,p_{\mathcal{X}}}$, where $\mathbf{B}$ is the $p_{\mathcal{X}}\times p_{\mathcal{Y}}$ matrix with the estimated coefficients of β that arises from expressing (6.3) as a linear model with multivariate response. The bootstrap resampling for the case of functional response is not so clear and will need to be investigated carefully. Two naive approaches could be to apply a wild bootstrap in the projected residuals or directly in the functional residuals:

$$\varepsilon_i^*(t) = \hat{\varepsilon}_i(t)V_i^*,\ t \in [0,1],\ i = 1, \ldots, n.$$

Once it is possible to test the FLM with functional response, a possible step could be the checking of an autoregressive Hilbertian process $\mathrm{AR}(p)_{\mathbb{H}}$:

$$\mathcal{X}_r(t) = \langle \mathcal{X}_{r-1}(s), \beta_1(s,t)\rangle + \cdots + \langle \mathcal{X}_{r-p}(s), \beta_p(s,t)\rangle + \varepsilon(t),\ s,t \in [0,1].$$

This model is potentially interesting for the prediction of the daily curves of temperature or the daily prices of financial assets. However, the bootstrap resampling for this kind of test also requires further investigation in order to reproduce the dependence structure properly.

Chapter 7

Conclusions

We have presented a goodness–of–fit test for the null hypothesis of the functional linear model. The test is constructed adapting the propose of Escanciano (2006) to the functional scheme with a basis representation. Different estimation methods for the functional parameter were considered, showing in general a similar behaviour in the performance of the test. The simulation study shows that the test behaves well in practise: respects the significance level and has good power. The test was applied to two real datasets to determine if the FLM was plausible, rejecting the null hypothesis for the first and finding no evidences for rejecting in the second.

Although in this work we have focused on the functional linear model, the proposed test can be extended to checking for any other regression model with functional covariate and scalar response. As shown in Chapter 6, the practical implementation and the wild bootstrap calibration remain the same. Therefore, obvious extensions could be the testing of FLM with several functional covariates or the testing of the quadratic functional model. Further, if a consistent bootstrap resampling is considered, the test could be extended to the case of functional response.

Finally, let us remark that the code for the implementation of the goodness–of–fit test in the simple and composite cases is available throughout the function `flm.test` of the **R** library `fda.usc`. This function also shows the graphical tool introduced in Chapter 5. To speed up the computation of the test statistic, the critical parts of the test implementation have been programmed in FORTRAN.

Appendix A

Contributed code to `fda.usc`

This appendix contains the reference manual of the contributed functions to the 0.9.8.1 version of the **R** package `fda.usc` (see Febrero-Bande and Oviedo de la Fuente (2012)). The most important function is `flm.test`, which implements the proposed testing procedure for the FLM for different estimating procedures and options. The function is based on the functions `PCvM.statistic` and `Adot`, which compute the test statistic and represent the computationally hard part of the procedure. In order to speed up computations, several optimization techniques have to be applied, being the most obvious the use of the compiled programming language FORTRAN. Therefore the functions `PCvM.statistic` and `Adot` are nothing but **R** wrappers of the FORTRAN functions `pcvm_statistic` and `adot`.

Further, as the competitive procedures described in Chapter 4 had not been implemented in any **R** package, they were added to `fda.usc` throughout the functions `flm.Ftest` and `dfv.test`. These functions, as well as `flm.test`, make use of `rber.gold` to simulate the Bernoulli random variable of the golden section bootstrap.

The **R** and FORTRAN code for this functions is available at CRAN: `http://cran.r-project.org/web/packages/fda.usc/index.html` `dfv.test` *Delsol, Ferraty and Vieu test for no functional-scalar interaction*

Description

The function `dfv.test` tests the null hypothesis of no interaction between a functional covariate and a scalar response in a general framework. The null hypothesis is

$$H_0 : m(X) = 0,$$

where $m(\cdot)$ denotes the regression function of the functional variate X over the centred scalar response Y ($E[Y] = 0$). The way of testing the null hypothesis is via the smoothed integrated square error of the response (see Details).

Usage

```
dfv.statistic (X.fdata, Y, h=quantile(x=metric.lp(X.fdata),
               probs=c(0.05,0.10,0.15,0.25,0.50)),
               K=function(x)2*dnorm(abs(x)),
               weights=rep(1,dim(X.fdata$data)[1]),d=metric.lp,
               dist=NULL)

dfv.test (X.fdata, Y, B=5000, h=quantile(x=metric.lp(X.fdata),
          probs=c(0.05,0.10,0.15,0.25,0.50)),
          K=function(x)2*dnorm(abs(x)),
          weights=rep(1,dim(X.fdata$data)[1]),d=metric.lp,
          show.prog=TRUE)
```

Arguments

`X.fdata` Functional covariate. The object must be in the class `fdata`.

`Y` Scalar response. Must be a vector with the same number of elements as functions are in `X.fdata`.

`h` Bandwidth parameter for the kernel smoothing. This is a crucial parameter that affects the power performance of the test. One possibility to choose it is considering the Cross-validatory bandwidth of the nonparametric functional regression, given by the function `fregre.np` (see Examples). Other possibility is to consider a grid of bandwidths. This is the default option, considering the grid given by the quantiles 0.05, 0.10, 0.15, 0.25 and 0.50 of the functional L^2 distances of the data.

`B` Number of bootstrap replicates to calibrate the distribution of the test statistic. `B=5000` replicates are the recommended for carry out the test, although for exploratory analysis (**not inferential**), an acceptable less time-consuming option is `B=500`.

`K` Kernel function. If no specified it is taken to be the rescaled right part of the normal density.

`weights` A vector of weights for the sample data. The default is the uniform weights `rep(1,dim(X.fdata$data)[1])`.

`d` Semimetric to use in the kernel smoothers. By default is the L^2 distance given by `metric.lp`.

`dist` Matrix of distances of the functional data, used to save time in the bootstrap calibration. If not given, the matrix is automatically computed using the semimetric `d`.

`show.prog` Either to show or not information about computing progress.

Details

The Delsol, Ferraty and Vieu statistic is defined as

$$T_n = \int \left(\sum_{i=1}^{n} (Y_i - m(X_i)) K\left(\frac{d(X, X_i)}{h}\right) \right)^2 \omega(X) dP_X(X)$$

and in the case of no interaction with **centred** scalar response (when $H_0 : m(X) = 0$ holds), its sample version is computed from

$$T_n = \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{i=1}^{n} Y_i K\left(\frac{d(X_j, X_i)}{h}\right) \right)^2 \omega(X_j).$$

The sample version implemented here does not consider a splitting of the sample, as the authors comment in their paper. The statistic is computed by the function `dfv.statistic` and, before applying the test, the response Y is centred. The distribution of the test statistic is approximated by a wild bootstrap on the residuals, using the *golden section bootstrap.*

Please note that if a grid of bandwidths is passed, a harmless warning message will prompt at the end of the test (it comes from returning several p-values in the `htest` class).

For further details of the test and its implementation see Delsol *et al.* (2011) and Garcia-Portugues *et al.* (2012), respectively.

Value

The value of `dfv.statistic` is a vector of length `length(h)` with the values of the statistic for each bandwidth. The value of `dfv.test` is an object with class `"htest"` whose underlying structure is a list containing the following components:

`statistic` The value of the Delsol, Ferraty and Vieu test statistic.

`boot.statistics`
A vector of length `B` with the values of the bootstrap test statistics.

`p.value` The p-value of the test.

`method` The character string “Delsol, Ferraty and Vieu test for no functional-scalar interaction”.

`B` The number of bootstrap replicates used.

`h` Bandwidth parameters for the test.

`K` Kernel function used.

`weights` The weights considered.

`d` Matrix of distances of the functional data.

`data.name` The character string “Y=0+e”

Note

No NA's are allowed neither in the functional covariate nor in the scalar response.

Author(s)

Eduardo Garcia-Portugues. Please, report bugs and suggestions to
<eduardo.garcia@usc.es>

References

Delsol, L., Ferraty, F. and Vieu, P. (2011). Structural test in regression on functional variables. Journal of Multivariate Analysis, 102, 422-447. http://dx.doi.org/10.1016/j.jmva.2010.10.003

Delsol, L., Ferraty, F. and Vieu, P. No effect tests in regression on functional variable and some applications to spectrometric studies.

Garcia-Portugues, E., Gonzalez-Manteiga, W. and Febrero-Bande, M. (2012). A goodness–of–fit test for the functional linear model with scalar response. http://arxiv.org/abs/1205.6167

See Also

rber.gold, flm.test, flm.Ftest, fregre.np

Examples

```
## Simulated example ##

X=rproc2fdata(n=50,t=seq(0,1,l=101),sigma="OU")

beta0=fdata(mdata=rep(0,length=101)+rnorm(101,sd=0.05),
argvals=seq(0,1,l=101),rangeval=c(0,1))
beta1=fdata(mdata=cos(2*pi*seq(0,1,l=101))-(seq(0,1,l=101)-0.5)^2+
rnorm(101,sd=0.05),argvals=seq(0,1,l=101),rangeval=c(0,1))

# Null hypothesis holds
Y0=drop(inprod.fdata(X,beta0)+rnorm(50,sd=0.1))

# Null hypothesis does not hold
Y1=drop(inprod.fdata(X,beta1)+rnorm(50,sd=0.1))

# We use the CV bandwidth given by fregre.np
# Do not reject H0
dfv.test(X,Y0,h=fregre.np(X,Y0)$h.opt,B=100)
# dfv.test(X,Y0,B=5000)

# Reject H0
dfv.test(X,Y1,B=100)
# dfv.test(X,Y1,B=5000)
```

`flm.Ftest` *F-test for the Functional Linear Model with scalar response*

Description

The function `flm.Ftest` tests the null hypothesis of no interaction between a functional covariate and a scalar response inside the Functional Linear Model (FLM): $Y = \big\langle X, \beta \big\rangle + \epsilon$. The null hypothesis is $H_0 : \beta = 0$ and the alternative is $H_1 : \beta \neq 0$. The way of testing the null hypothesis is via a functional extension of the classical F-test (see Details).

Usage

```
Ftest.statistic (X.fdata, Y)
flm.Ftest (X.fdata, Y, B=5000, show.prog=TRUE)
```

Arguments

`X.fdata`	Functional covariate for the FLM. The object must be in the class `fdata`.
`Y`	Scalar response for the FLM. Must be a vector with the same number of elements as functions are in `X.fdata`.
`B`	Number of bootstrap replicates to calibrate the distribution of the test statistic. `B=5000` replicates are the recommended for carry out the test, although for exploratory analysis (**not inferential**), an acceptable less time-consuming option is `B=500`.
`show.prog`	Either to show or not information about computing progress.

Details

The Functional Linear Model with scalar response (FLM), is defined as $Y = \big\langle X, \beta \big\rangle + \epsilon$, for a functional process X such that $E[X(t)] = 0$, $E[X(t)\epsilon] = 0$ for all t and for a scalar variable Y such that $E[Y] = 0$. The *functional F-test* is defined as

$$T_n = \left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \right\|,$$

where $\bar{X}$ is the functional mean of X, $\bar{Y}$ is the ordinary mean of Y and $\|\cdot\|$ is the L^2 functional norm. The statistic is computed with the function `Ftest.statistic`. The distribution of the test statistic is approximated by a wild bootstrap on the residuals, using the *golden section bootstrap*.

For further details of the test and its implementation see Gonzalez-Manteiga *et al.* (2012) and Garcia-Portugues *et al.* (2012), respectively.

Value

The value for `Ftest.statistic` is simply the F-test statistic. The value for `flm.Ftest` is an object with class `"htest"` whose underlying structure is a list containing the following components:

`statistic`	The value of the F-test statistic.

`boot.statistics`
A vector of length `B` with the values of the bootstrap F-test statistics.

`p.value` The p-value of the test.

`method` The character string “Functional Linear Model F-test”.

`B` The number of bootstrap replicates used.

`data.name` The character string “Y=<X,0>+e”

Note

No NA’s are allowed neither in the functional covariate nor in the scalar response.

Author(s)

Eduardo Garcia-Portugues. Please, report bugs and suggestions to
`<eduardo.garcia@usc.es>`

References

Garcia-Portugues, E., Gonzalez-Manteiga, W. and Febrero-Bande, M. (2012). A goodness–of–fit test for the functional linear model with scalar response. `http://arxiv.org/abs/1205.6167`

Gonzalez-Manteiga, W., Gonzalez-Rodriguez, G., Martinez-Calvo, A. and Garcia-Portugues, E. Bootstrap independence test for functional linear models.

See Also

`rber.gold`, `flm.test`, `dfv.test`

Examples

```
## Simulated example ##

X=rproc2fdata(n=50,t=seq(0,1,l=101),sigma="OU")

beta0=fdata(mdata=rep(0,length=101)+rnorm(101,sd=0.05),
argvals=seq(0,1,l=101),rangeval=c(0,1))
beta1=fdata(mdata=cos(2*pi*seq(0,1,l=101))-(seq(0,1,l=101)-0.5)^2+
rnorm(101,sd=0.05),argvals=seq(0,1,l=101),rangeval=c(0,1))

# Null hypothesis holds
Y0=drop(inprod.fdata(X,beta0)+rnorm(50,sd=0.1))

# Null hypothesis does not hold
Y1=drop(inprod.fdata(X,beta1)+rnorm(50,sd=0.1))

# Do not reject H0
flm.Ftest(X,Y0,B=100)
# flm.Ftest(X,Y0,B=5000)

# Reject H0
flm.Ftest(X,Y1,B=100)
# flm.Ftest(X,Y1,B=5000)
```

flm.test *Goodness-of-fit test for the Functional Linear Model with scalar response*

Description

The function `flm.test` tests the composite null hypothesis of a Functional Linear Model with scalar response (FLM),

$$H_0 : Y = \big\langle X, \beta \big\rangle + \epsilon,$$

versus a general alternative. If $\beta = \beta_0$ is provided, then the simple hypothesis $H_0 : Y = \big\langle X, \beta_0 \big\rangle + \epsilon$ is tested. The way of testing the null hypothesis is via a Projected Cramer-von Mises test (see Details).

Usage

```
flm.test (X.fdata, Y, beta0.fdata = NULL, B = 5000, est.method = "pls",
          p = NULL, type.basis = "bspline", show.prog = TRUE,
          plot.it = TRUE, B.plot = 100, G = 200, ...)
```

Arguments

`X.fdata` Functional covariate for the FLM. The object must be in the class `fdata`.

`Y` Scalar response for the FLM. Must be a vector with the same number of elements as functions are in `X.fdata`.

`beta0.fdata` Functional parameter for the simple null hypothesis, in the `fdata` class. Recall that the `argvals` and `rangeval` arguments of `beta0.fdata` must be the same of `X.fdata`. A possibility to do this is to consider, for example for $\beta_0 = 0$ (the simple null hypothesis of no interaction),
`beta0.fdata=fdata(mdata=rep(0,length(X.fdata$argvals)),`
`argvals=X.fdata$argvals,rangeval=X.fdata$rangeval)`.
If `beta0.fdata=NULL` (default), the function will test for the composite null hypothesis.

`B` Number of bootstrap replicates to calibrate the distribution of the test statistic. `B=5000` replicates are the recommended for carry out the test, although for exploratory analysis (**not inferential**), an acceptable less time-consuming option is `B=500`.

`est.method` Estimation method for the unknown parameter β, only used in the composite case. Mainly, there are two options: specify the number of basis elements for the estimated β by `p` or optimally select `p` by a data-driven criteria (see Details section for discussion). Then, it must be one of the following methods:

- `"pc"` If `p`, the number of basis elements, is given, then β is estimated by `fregre.pc`. Otherwise, an optimum `p` is chosen using `fregre.pc.cv` and the `"SIC"` (BIC) criteria.
- `"pls"` If `p` is given, β is estimated by `fregre.pls`. Otherwise, an optimum `p` is chosen using `fregre.pls.cv` and the `"CV"` criteria. This is the default argument as it has been checked empirically that provides

a good balance between the performance of the test and the estimation of β.

- `"basis"` If `p` is given, β is estimated by `fregre.basis`. Otherwise, an optimum `p` is chosen using `fregre.basis.cv` and the `"GCV.S"` criteria. In these functions, the same basis for the arguments `basis.x` and `basis.b` is considered. The type of basis used will be the given by the argument `type.basis` and must be one of the class of `create.basis`. Further arguments passed to `create.basis` (not `rangeval` that is taken as the `rangeval` of `X.fdata`), can be passed throughout `...` .

`p` Number of elements of the basis considered. If it is not given, an optimal `p` will be chosen using a specific criteria (see `est.method` and `type.basis` arguments).

`type.basis` Type of basis used to represent the functional process. Depending on the hypothesis it will have a different interpretation:

- Simple hypothesis. One of these options:
  - `"bspline"` If `p` is given, the functional process is expressed in a basis of `p` B-splines. If not, an optimal `p` will be chosen by `min.basis`, using the `"GCV.S"` criteria.
  - `"fourier"` If `p` is given, the functional process is expressed in a basis of `p` fourier functions. If not, an optimal `p` will be chosen by `min.basis`, using the `"GCV.S"` criteria.
  - `"pc"` `p` must be given. Expresses the functional process in a basis of `p` PC.
  - `"pls"` `p` must be given. Expresses the functional process in a basis of `p` PLS.

  Although other of the basis supported by `create.basis` are possible too, `"bspline"` and `"fourier"` are recommended. Other basis may cause incompatibilities.
- Composite hypothesis. This argument is only used when `est.method="basis"` and, in this case, claims for the type of basis used in the basis estimation method of the functional parameter. Again, basis `"bspline"` and `"fourier"` are recommended, as other basis may cause incompatibilities.

`show.prog` Either to show or not information about computing progress.

`plot.it` Either to show or not a graph of the observed trajectory, and the bootstrap trajectories under the null composite hypothesis, of the process $R_n(\cdot)$ (see Details). Note that if `plot.it=TRUE`, the function takes more time to run.

`B.plot` Number of bootstrap trajectories to show in the resulting plot of the test. As the trajectories shown are the first `B.plot` of `B`, `B.plot` must be lower or equal to `B`.

`G` Number of projections used to compute the trajectories of the process $R_n(\cdot)$ by Monte Carlo.

`...` Further arguments passed to `create.basis`.

Details

The Functional Linear Model with scalar response (FLM), is defined as $Y = \big\langle X, \beta \big\rangle + \epsilon$, for a functional process X such that $E[X(t)] = 0$, $E[X(t)\epsilon] = 0$ for all t and for a scalar variable Y such that $E[Y] = 0$. Then, the test assumes that `Y` and `X.fdata` are **centred** and will automatically center them. So, bear in mind that when you apply the test for `Y` and `X.fdata`, actually, you are applying it to `Y-mean(Y)` and `fdata.cen(X.fdata)$Xcen`.

The test statistic corresponds to the Cramer-von Mises norm of the *Residual Marked empirical Process based on Projections* $R_n(u, \gamma)$ defined in Garcia-Portugues *et al.* (2012). The expression of this process in a p-truncated basis of the space $L^2[0, T]$ leads to the p-multivariate process $R_{n,p}\big(u, \gamma^{(p)}\big)$, whose Cramer-von Mises norm is easily computed.

The choice of an appropriate p to represent the functional process X, in case that is not provided, is done via the estimation of β for the composite hypothesis. For the simple hypothesis, as no estimation of β is done, the choice of p depends only on the functional process X. As the result of the test may change for different p's, we recommend to use an automatic criterion to select p instead of provide a fixed one. The distribution of the test statistic is approximated by a wild bootstrap on the residuals, using the *golden section bootstrap*.

Finally, the graph shown if `plot.it=TRUE` represents the observed trajectory, and the bootstrap trajectories under the null, of the process RMPP *integrated on the projections*:

$$R_n(u) \approx \frac{1}{G} \sum_{g=1}^{G} R_n(u, \gamma_g),$$

where γ_g are simulated as Gaussians processes. This gives a graphical idea of how *distant* is the observed trajectory from the null hypothesis.

For further details see Garcia-Portugues *et al.* (2012).

Value

An object with class `"htest"` whose underlying structure is a list containing the following components:

`statistic` The value of the test statistic.

`boot.statistics` A vector of length `B` with the values of the bootstrap test statistics.

`p.value` The p-value of the test.

`method` The method used.

`B` The number of bootstrap replicates used.

`type.basis` The type of basis used.

`beta.est` The estimated functional parameter β in the composite hypothesis. For the simple hypothesis, the given `beta0.fdata`.

`p` The number of basis elements passed or automatically chosen.

`ord` The optimal order for PC and PLS given by `fregre.pc.cv` and `fregre.pls.cv`. For other methods is setted to `1:p`.

`data.name` The character string “Y=<X,b>+e”

Note

No NA’s are allowed neither in the functional covariate nor in the scalar response.

Author(s)

Eduardo Garcia-Portugues. Please, report bugs and suggestions to
<eduardo.garcia@usc.es>

References

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. Econometric Theory, 22, 1030-1051. http://dx.doi.org/10.1017/S0266466606060506

Garcia-Portugues, E., Gonzalez-Manteiga, W. and Febrero-Bande, M. (2012). A goodness–of–fit test for the functional linear model with scalar response. http://arxiv.org/abs/1205.6167

See Also

Adot, PCvM.statistic, rber.gold, flm.Ftest, dfv.test, fregre.pc, fregre.pls, fregre.basis, fregre.pc.cv, fregre.pls.cv, fregre.basis.cv, min.basis, create.basis

Examples

```
## Simulated example ##

X=rproc2fdata(n=100,t=seq(0,1,l=101),sigma="OU")
beta0=fdata(mdata=cos(2*pi*seq(0,1,l=101))-(seq(0,1,l=101)-0.5)^2+
rnorm(101,sd=0.05),argvals=seq(0,1,l=101),rangeval=c(0,1))
Y=inprod.fdata(X,beta0)+rnorm(100,sd=0.1)

dev.new(width=21,height=7)
par(mfrow=c(1,3))
plot(X,main="X")
plot(beta0,main="beta0")
plot(density(Y),main=``Density of Y",xlab="Y",ylab="Density")
rug(Y)

# Composite hypothesis: do not reject FLM
pcvm.sim=flm.test(X,Y,B=50,B.plot=50,G=100,plot.it=TRUE)
pcvm.sim
# flm.test(X,Y,B=5000)

# Estimated beta
dev.new()
plot(pcvm.sim$beta.est)

# Simple hypothesis: do not reject beta=beta0
flm.test(X,Y,beta0.fdata=beta0,B=50,B.plot=50,G=100)
# flm.test(X,Y,beta0.fdata=beta0,B=5000)


## AEMET dataset ##

# data(aemet)

## Remove the 5% of the curves with less depth (i.e. 4 curves)
```

```
# dev.new()
# res.FM=depth.FM(aemet$temp,draw=TRUE)
# qu=quantile(res.FM$dep,prob=0.05)
# l=which(res.FM$dep<=qu)
# lines(aemet$temp[l],col=3)
# aemet$df$name[l]

## Data without outliers
# wind.speed=apply(aemet$wind.speed$data,1,mean)[-l]
# temp=aemet$temp[-l]

## Exploratory analysis: accept the FLM
# pcvm.aemet=flm.test(temp,wind.speed,est.method="pls",B=100,B.plot=50,G=100)
# pcvm.aemet

## Estimated beta
# dev.new()
# plot(pcvm.aemet$beta.est,lwd=2,col=2)

## B=5000 for more precision on calibration of the test: also accept the FLM
# flm.test(temp,wind.speed,est.method="pls",B=5000)

## Simple hypothesis: rejection of beta0=0? Limiting p-value...
# dat=rep(0,length(temp$argvals))
# flm.test(temp,wind.speed, beta0.fdata=fdata(mdata=dat,argvals=temp$argvals,
# rangeval=temp$rangeval),B=100)
# flm.test(temp,wind.speed, beta0.fdata=fdata(mdata=dat,argvals=temp$argvals,
# rangeval=temp$rangeval),B=5000)


## Tecator dataset ##

# data(tecator)
# names(tecator)
# absorp=tecator$absorp.fdata
# ind=1:129 # or ind=1:215
# x=absorp[ind,]
# y=tecator$y$Fat[ind]
# tt=absorp[["argvals"]]

## Exploratory analysis for composite hypothesis with automatic choose of p
# pcvm.tecat=flm.test(x,y,B=100,B.plot=50,G=100)
# pcvm.tecat

## B=5000 for more precision on calibration of the test: also reject the FLM
# flm.test(x,y,B=5000)

## Plot of the estimated functional parameters
# plot(pcvm.tecat$beta.est,lwd=2,col=2)
# for(i in 1:100) lines(pcvm.tecat$boot.beta.est[[i]])
# lines(pcvm.tecat$beta.est,lwd=2,col=2)
# legend("topright",legend=c("Estimated","Bootstrap"),col=1:2,lwd=2)

## Distribution of the PCvM statistic
# plot(density(pcvm.tecat$boot.statistics),lwd=2,xlim=c(0,10),
# main="PCvM distribution", xlab="PCvM*",ylab="Density")
# rug(pcvm.tecat$boot.statistics)
# abline(v=pcvm.tecat$statistic,col=2,lwd=2)
# legend("top",legend=c("PCvM observed"),lwd=2,col=2)

## Simple hypothesis: fixed p
# dat=rep(0,length(x$argvals))
# flm.test(x,y,beta0.fdata=fdata(mdata=dat,argvals=x$argvals,
# rangeval=x$rangeval),B=100,p=11)

## Simple hypothesis, automatic choose of p
# flm.test(x,y,beta0.fdata=fdata(mdata=dat,argvals=x$argvals,
# rangeval=x$rangeval),B=100)
# flm.test(x,y,beta0.fdata=fdata(mdata=dat,argvals=x$argvals,
# rangeval=x$rangeval),B=5000)
```

PCvM.statistic *PCvM statistic for the Functional Linear Model with scalar response*

Description

Projected Cramer-von Mises statistic (PCvM) for the Functional Linear Model with scalar response (FLM): $Y = \langle X, \beta \rangle + \varepsilon$.

Usage

```
PCvM.statistic (X, residuals, p, Adot.vec)
Adot (X, inpr)
```

Arguments

`X` Functional covariate for the FLM. The object must be either in the class `fdata` or in the class `fd`. It is used to compute the matrix of inner products.

`residuals` Residuals of the estimated FLM.

`p` Number of elements of the functional basis where the functional covariate is represented.

`Adot.vec` Output from the `Adot` function (see Details). Computed if not given.

`inpr` Matrix of inner products of `X`. Computed if not given.

Details

In order to optimize the computation of the statistic, the critical parts of these two functions are programmed in FORTRAN. The hardest part corresponds to the function `Adot`, which involves the computation of a symmetric matrix of dimension $n \times n$ where each entry is a sum of n elements. As this matrix is symmetric, the order of the method can be reduced from $O(n^3)$ to $O\left(\frac{n^3-n^2}{2}\right)$. The memory requirement can also be reduced to $O\left(\frac{n^2-n+2}{2}\right)$. The value of `Adot` is a vector of length $\frac{n^2-n+2}{2}$ where the first element is the common diagonal element and the rest are the lower triangle entries of the matrix, sorted by rows (see Examples).

Value

For `PCvM.statistic`, the value of the statistic. For `Adot`, a suitable output to be used in the argument `Adot.vec`.

Note

No NA's are allowed in the functional covariate.

Author(s)

Eduardo Garcia-Portugues. Please, report bugs and suggestions to
`<eduardo.garcia@usc.es>`

References

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. Econometric Theory, 22, 1030-1051. `http://dx.doi.org/10.1017/S0266466606060506`

Garcia-Portugues, E., Gonzalez-Manteiga, W. and Febrero-Bande, M. (2012). A goodness–of–fit test for the functional linear model with scalar response. `http://arxiv.org/abs/1205.6167`

See Also

`flm.test`

Examples

```
# Functional process
X=rproc2fdata(n=10,t=seq(0,1,l=101))

# Adot
Adot.vec=Adot(X)

# Obtain the entire matrix Adot
Ad=diag(rep(Adot.vec[1],dim(X$data)[1]))
Ad[upper.tri(Ad,diag=FALSE)]=Adot.vec[-1]
Ad=t(Ad)
Ad=Ad+t(Ad)-diag(diag(Ad))
Ad

# Statistic
PCvM.statistic(X,residuals=rnorm(10),p=5)
```

`rber.gold` *Gold section bootstrap sampling*

Description

Sampling from a binomial variable with values $\left\{\frac{1-\sqrt{5}}{2}, \frac{1+\sqrt{5}}{2}\right\}$ and probabilities $\left\{\frac{5+\sqrt{5}}{10}, \frac{5-\sqrt{5}}{10}\right\}$, respectively.

Usage

```
rber.gold (n)
```

Arguments

`n`	Number of observations.

Details

For the construction of wild bootstrap residuals, sampling from a random variable V such that $E[V^2] = 0$ and $E[V] = 0$ is needed. A simple and suitable V is obtained with a binomial variable of the form:

$$P\left\{V = \frac{1-\sqrt{5}}{2}\right\} = \frac{5+\sqrt{5}}{10} and P\left\{V = \frac{1+\sqrt{5}}{2}\right\} = \frac{5-\sqrt{5}}{10},$$

which leads to the *golden section bootstrap.* If `e` denotes a vector of `n` residuals, the wild bootstrap residuals would be computed as `e*rber.gold(n)`.

Value

A sample of length `n` of the random variable V.

Author(s)

Eduardo Garcia-Portugues. Please, report bugs and suggestions to
`<eduardo.garcia@usc.es>`

See Also

`rbinom`, `flm.test`, `flm.Ftest`, `dfv.test`

Examples

```
# Sampling
samp=rber.gold(100)
mean(samp)
sd(samp)
samp

# Construction of wild bootstrap residuals
e=rnorm(200)
e.boot=e*rber.gold(200)
summary(e)
summary(e.boot)
```

Bibliography

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.*, 34(5):2159–2179.

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003). Testing hypotheses in the functional linear model. *Scand. J. Statist.*, 30(1):241–255.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statist. Probab. Lett.*, 45(1):11–22.

Chiou, J.-M. and Müller, H.-G. (2007). Diagnostics for functional regression via residual processes. *Comput. Statist. Data Anal.*, 51(10):4849–4863.

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.*, 37(1):35–72.

Cuesta-Albertos, J. A., del Barrio, E., Fraiman, R., and Matrán, C. (2007). The random projection method in goodness of fit for functional data. *Comput. Statist. Data Anal.*, 51(10):4814–4831.

de Boor, C. (2001). *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition.

Delsol, L., Ferraty, F., and Vieu, P. (2011). Structural test in regression on functional variables. *J. Multivariate Anal.*, 102(3):422–447.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2010). Measures of influence for the functional linear model with scalar response. *J. Multivariate Anal.*, 101(2):327–339.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). *fda.usc: Functional Data Analysis and Utilities for Statistical Computing (fda.usc)*. URL `http://cran.r-project.org/web/packages/fda.usc/`. R package version 0.9.8.1.

Ferraty, F. and Romain, Y. (2011). *The Oxford Handbook of functional data analysis.* Oxford University Press.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

González-Manteiga, W. and Crujeiras, R. (2011). A general view of the goodness–of–fit tests for statistical models. In Pardo, L., Balakrishnan, N., and Gil, M., editors, *Modern Mathematical Tools and Techniques in Capturing Complexity*, volume 72 of *Understanding Complex Systems*, pages 3–16. Springer Berlin / Heidelberg.

González-Manteiga, W., González-Rodríguez, G., Martínez-Calvo, A., and García-Portugués, E. (2012). Bootstrap independence test for functional linear models. Unpublished paper.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.*, 35(1):70–91.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):109–126.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21(4):1926–1947.

Lavergne, P. and Patilea, V. (2008). Breaking the curse of dimensionality in nonparametric testing. *J. Econometrics*, 143(1):103–122.

Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *J. Multivariate Anal.*, 98(9):1782–1804.

Patilea, V., Sellero, C. S., and Saumard, M. (2012). Projection–based nonparametric testing for functional covariate effect. arXiv:1205.5578.

Preda, C. and Saporta, G. (2002). Régression pls sur un processus stochastique. *Revue de statistique appliquée*, 50(2):27–46.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis.* Springer Series in Statistics. Springer, New York, second edition.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, 53(3):683–690.

Stute, W., González Manteiga, W., and Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.*, 93(441):141–149.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics*, 75(2):263–289.