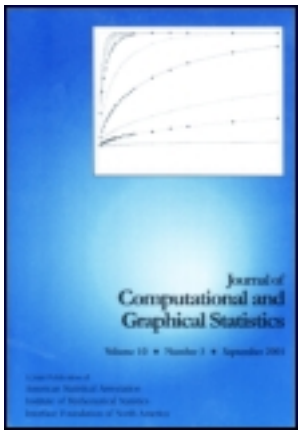

Journal of Computational and Graphical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

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

Eduardo García-Portugués [a], Wenceslao González-Manteiga [a] & Manuel Febrero-Bande [a]
[a] Department of Statistics and Operations Research, University of Santiago de Compostela
Accepted author version posted online: 27 Jun 2013.

To cite this article: Journal of Computational and Graphical Statistics (2013): A goodness-of-fit test for the functional linear model with scalar response, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2013.812519

To link to this article: http://dx.doi.org/10.1080/10618600.2013.812519

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

Eduardo García–Portugués*, Wenceslao González–Manteiga and Manuel Febrero–Bande†

Department of Statistics and Operations Research, University of Santiago de Compostela

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. The test statistic is easy to compute using geometrical and matrix arguments, and simple to calibrate in its distribution by a wild bootstrap on the residuals. The finite sample properties of the test are illustrated by a simulation study 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 and a graphical tool is introduced. Supplementary materials are available online.

Keywords: Functional Data; Goodness-of-Fit Methods; Bootstrap/resampling; Statistical Computing.

*Corresponding author. e–mail: eduardo.garcia@usc.es

†The authors acknowledge the support of Project MTM2008–03010, from the Spanish Ministry of Science and Innovation, Project 10MDS207015PR from Dirección Xeral de I+D, Xunta de Galicia and IAP network StUDyS, from Belgian Science Policy. Work of E. García–Portugués has been supported by FPU grant AP2010–0957 from the Spanish Ministry of Education. The authors also acknowledge the suggestions by two anonymous referees that helped improving this paper.

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 in Section 2. The main part of this work is Section 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 Section 4. Section 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. Final comments and conclusions are given in Section 6. An appendix in supplementary materials, available online, contains omitted proofs, tables and figures.

2 Background

The main goal of this paper 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 $\|f\|_p = \left(\int_0^1 |f(t)|^p \, dt\right)^{\frac{1}{p}}$. is finite. 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 $\left\{\Psi_j\right\}_{j=1}^{\infty}$ 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 $x_j = \left\langle \mathcal{X}, \Psi_j \right\rangle$, $j \geq 1$. A basis is said to be orthogonal if $\left\langle \Psi_i, \Psi_j \right\rangle = 0$, $i \neq j$ and orthonormal if, in addition, $\left\langle \Psi_j, \Psi_j \right\rangle = 1$, $j \geq 1$. 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–truncated basis $\left\{\Psi_j\right\}_{j=1}^{p}$, which corresponds to the first p elements of the infinite basis $\left\{\Psi_j\right\}_{j=1}^{\infty}$. The representation of $\mathcal{X}$ in this truncated basis is denoted by $\mathcal{X}^{(p)} = \sum_{j=1}^{p} x_j \Psi_j$. 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 $\mathbf{\Psi} = \left(\left\langle \Psi_i, \Psi_j \right\rangle\right)_{ij}$ the matrix of inner products of the p–truncated basis, $\mathbb{S}_{\mathbf{\Psi}}^{p} = \left\{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^T \mathbf{\Psi} \mathbf{x} = 1\right\}$ the p–ellipsoid generated by this matrix and $\mathbf{R}^T\mathbf{R}$ the Cholesky decomposition of $\mathbf{\Psi}$ (a semi–positive matrix). First of all, we have the trivial isomorphism that maps elements of $\mathbb{S}_{\mathbb{H}}^{p}$ to elements of $\mathbb{S}_{\mathbf{\Psi}}^{p}$ by means of the functional coefficients: $\phi : f = \sum_{j=1}^{p} x_j \Psi_j \in \mathbb{S}_{\mathbb{H}}^{p} \mapsto \phi(f) = \mathbf{x} \in \mathbb{S}_{\mathbf{\Psi}}^{p}$. Recall that functions ϕ and ϕ^{-1} are well defined because $\|f\|_{\mathbb{H}}^2 = \left\langle \sum_{j=1}^{p} x_j \Psi_j, \sum_{j=1}^{p} x_j \Psi_j \right\rangle = \mathbf{x}^T \mathbf{\Psi} \mathbf{x}$. We must consider also a linear transformation from $\mathbb{S}^p$ to $\mathbb{S}_{\mathbf{\Psi}}^{p}$, which is given by $\rho : \mathbf{x} \in \mathbb{S}^p \mapsto \rho(\mathbf{x}) = \mathbf{R}^{-1}\mathbf{x} \in \mathbb{S}_{\mathbf{\Psi}}^{p}$ 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:

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

In the case where the basis is orthonormal, $\mathbf{\Psi}$ 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 $\left\{\Psi_j\right\}_{j=1}^{p}$ 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., $\mathbb{E}[\mathcal{X}(t)] = 0$ for a.e. $t \in [0,1]$ and $\mathbb{E}[Y] = 0$, 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 $\mathbb{E}[\mathcal{X}(t)\varepsilon] = 0,\ \forall t$. The prediction of Y is done with the conditional expectation of Y given $\mathcal{X}$:

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

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):

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

A possible method to search for the parameter β that minimises the RSS is representing the func-

tional data and the functional parameter in the truncated functional basis $\left\{\Psi_j\right\}_{j=1}^{p_\mathcal{X}}$ and $\left\{\theta_j\right\}_{j=1}^{p_\beta}$, respectively:

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

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

$$Y = \langle\mathcal{X},\beta\rangle + \varepsilon \approx \mathbf{CJb} + \varepsilon = \mathbf{Zb} + \varepsilon, \tag{1}$$

where $\mathbf{J} = \left(\left\langle\Psi_i,\theta_j\right\rangle\right)_{ij}$. Then, basis representation allows to express the FLM as a standard linear regression, where the estimated coefficients of β in the basis $\left\{\theta_j\right\}_{j=1}^{p_\beta}$ are given by $\hat{\mathbf{b}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$. Although different combinations of $\left\{\Psi_j\right\}_{j=1}^{p_\mathcal{X}}$ and $\left\{\theta_j\right\}_{j=1}^{p_\beta}$ are possible, the usual choice is $\left\{\Psi_j\right\}_{j=1}^{p} = \left\{\theta_j\right\}_{j=1}^{p}$, being $\left\{\Psi_j\right\}_{j=1}^{p}$ 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). The so called Functional Principal Component (FPC) regression estimation, proposed by Cardot et al. (1999), provide an orthogonal data–driven basis that gives the most rapidly convergent representation of the functional dataset predictor in a L^2 sense (see Hall and Horowitz (2007)). Preda and Saporta (2002) have proposed the Functional Partial Least Squares (FPLS) regression method that produces iteratively a sequence of orthogonal functions, as the FPC are, but with maximum predictive performance. To implement any of the methods shown before, it is required to fix the number of basis elements (or 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 the AIC, AICc and BIC information approaches.

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:

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

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 recalculate for every datum. As an alternative, GCV avoids recalculating the $\hat{\beta}^{(p)}$ for every datum by introducing a penalty term. The GCV is defined as

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

where df is the number of degrees of freedom consumed by the model, typically given by the trace of the matrix $\mathbf{Z}$. GCV is closely related with AIC, AICc and BIC although they come from different perspectives.

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 original distribution of the multidimensional data by the distribution of the randomly projected univariate 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, the test has an excellent power performance and has the best empirical power for most situations when comparing to their competitors.

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

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

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

3 The test

The presentation of the goodness–of–fit test that we propose in this paper is divided into three sections. The first and most important presents the theoretical fundamentals of the test, with starting point in Lemma 2, which proof is detailed in the appendix. The second derives the effective implementation of the test statistic in practise considering some geometrical and matrix arguments. Finally, the bootstrap resampling for the calibration of the test statistic 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 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. $\mathbb{E}\left[Y - \langle \mathcal{X}, \beta \rangle | \mathcal{X} = x\right] = 0$, *for a.e.* $x \in \mathbb{H}$.

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

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

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

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

Then H_0 is characterized by the null value of the moment $\mathbb{E}\left[(Y - \langle \mathcal{X}, \beta \rangle) \mathbb{1}_{\{\langle \mathcal{X}, \gamma \rangle \leq u\}}\right]$, for a.e. $u \in \mathbb{R}$ and $\forall \gamma \in \mathbb{S}_{\mathbb{H}}$ (or $\forall \gamma \in \mathbb{S}_{\mathbb{H}}^p$, $\forall p \geq 1$) 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:

$$R_n(u, \gamma) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left(Y_i - \left\langle \mathcal{X}_i, \hat{\beta} \right\rangle\right) \mathbb{1}_{\{\langle \mathcal{X}_i, \gamma \rangle \leq u\}}, \tag{3}$$

that will be denoted as the *Residual Marked empirical Process based on Projections* (RMPP). The marks of (3) are given by the residuals $\{Y_i - \langle \mathcal{X}_i, \hat{\beta} \rangle\}_{i=1}^n$ 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 Section 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 Section 6 for discussion).

To measure the distance of the empirical process (3) from zero, two possibilities are the classical Cramér–von Mises and Kolmogorov–Smirnov norms, adapted to the *projected* space

$\Pi = \mathbb{R} \times \mathbb{S}_{\mathbb{H}}$:

$$\text{PCvM}_n = \int_{\Pi} R_n(u,\gamma)^2 \, F_{n,\gamma}(du)\, \omega(d\gamma), \tag{4}$$

$$\text{PKS}_n = \sup_{(u,\gamma)\in\Pi} |R_n(u,\gamma)|\,, \tag{5}$$

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 measure on $\mathbb{S}_{\mathbb{H}}$. Unfortunately, the infinite dimension of the space $\mathbb{S}_{\mathbb{H}}$ makes infeasible to compute the functionals (4) and (5) 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 $\left\{\Psi_j\right\}_{j=1}^{\infty}$ be a basis of $\mathbb{H}$ and consider the p–truncated basis $\left\{\Psi_j\right\}_{j=1}^{p}$, with matrix of inner products $\mathbf{\Psi}$. Denote by $\mathcal{X}_i^{(p)}$ and $\gamma^{(p)}$ the representation of the functions $\mathcal{X}_i$ and γ in the p–truncated basis, with vectors of coefficients $\mathbf{x}_{i,p}$ and $\mathbf{g}_p$, respectively, and for $i = 1, \ldots, n$. Using this, as $\left\{\Psi_j\right\}_{j=1}^{\infty}$ is any basis, we have that

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

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

$$R_{n,p}\left(u, \gamma^{(p)}\right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_{i,p}^T \, \mathbf{\Psi} \, \mathbf{b}_p\right) \mathbb{1}_{\left\{\mathbf{x}_{i,p}^T \, \mathbf{\Psi} \, \mathbf{g}_p \le u\right\}} = R_{n,p}\left(u, \mathbf{g}_p\right),$$

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

Bearing in mind this, our test statistic propose is a modified version of (4) that results from

expressing all the functions in a p–truncated basis of $\mathbb{H}$:

$$\text{PCvM}_{n,p} = \int_{\mathbb{S}^p_{\mathbb{H}}\times\mathbb{R}} R_{n,p}\left(u,\gamma^{(p)}\right)^2 \, F_{n,\gamma^{(p)}}(du)\,\omega(d\gamma^{(p)}). \tag{6}$$

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, we have:

$$\begin{aligned}
\text{PCvM}_{n,p} &= \int_{\mathbb{S}^p_{\mathbf{\Psi}}\times\mathbb{R}} R_{n,p}(u,\mathbf{g}_p)^2 \, F_{n,\mathbf{g}_p}(du)\,\omega(d\mathbf{g}_p) \\
&= \int_{\mathbb{S}^p\times\mathbb{R}} |\mathbf{R}|^{-1} \, R_{n,p}(u,\mathbf{R}^{-1}\mathbf{g}_p)^2 \, F_{n,\mathbf{R}^{-1}\mathbf{g}_p}(du)\,\omega(d\mathbf{g}_p) \\
&= \int_{\mathbb{S}^p\times\mathbb{R}} |\mathbf{R}|^{-1} \left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_{i,p}^T\,\mathbf{\Psi}\,\mathbf{b}_p\right)\mathbb{1}_{\left\{\mathbf{x}_{i,p}^T\,\mathbf{R}^T\,\mathbf{g}_p\le u\right\}}\right)^2 \, F_{n,\mathbf{R}^{-1}\mathbf{g}_p}(du)\,\omega(d\mathbf{g}_p),
\end{aligned} \tag{7}$$

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 $\mathbb{S}^p$.

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 Section 2. These methods will be illustrated in Section 4.

3.2 Implementation

Following the steps of Escanciano (2006) it is possible to derive a simpler expression for (7). Using the definition of the RMPP in a p–truncated basis, the fact that $F_{n,\mathbf{R}^{-1}\mathbf{g}_p}$ is the ecdf of $\{\mathbf{x}_{i,p}^T \mathbf{\Psi}\mathbf{R}^{-1}\mathbf{g}_p\}_{i=1}^n = \{\mathbf{x}_{i,p}^T \mathbf{R}^T\mathbf{g}_p\}_{i=1}^n$ and some simple algebra, we have:

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

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

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

where $S_{ijr} = \left\{\boldsymbol{\xi}\in\mathbb{S}^p : \frac{\pi}{2}\le \measuredangle\left(\mathbf{x}'_{i,p}-\mathbf{x}'_{r,p},\boldsymbol{\xi}\right)\le\frac{3\pi}{2},\ \frac{\pi}{2}\le \measuredangle\left(\mathbf{x}'_{j,p}-\mathbf{x}'_{r,p},\boldsymbol{\xi}\right)\le\frac{3\pi}{2}\right\}$ 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 B1 in appendix) of width angle given by

$$\left|\pi - \arccos\left(\frac{(\mathbf{x}'_{i,p}-\mathbf{x}'_{r,p})^T(\mathbf{x}'_{j,p}-\mathbf{x}'_{r,p})}{\|\mathbf{x}'_{i,p}-\mathbf{x}'_{r,p}\|\cdot\|\mathbf{x}'_{j,p}-\mathbf{x}'_{r,p}\|}\right)\right|. \tag{8}$$

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

$$A_{ijr} = A_{ijr}^{(0)} \frac{\pi^{p/2-1}}{\Gamma\left(\frac{p}{2}\right)} |\mathbf{R}|^{-1}, \quad A_{ijr}^{(0)} = \begin{cases} 2\pi, & \mathbf{x}'_{i,p} = \mathbf{x}'_{j,p} = \mathbf{x}'_{r,p}, \\ \pi, & \mathbf{x}'_{i,p} = \mathbf{x}'_{j,p}, \mathbf{x}'_{i,p} = \mathbf{x}'_{r,p} \text{ or } \mathbf{x}'_{j,p} = \mathbf{x}'_{r,p}, \\ (8), & \text{else.} \end{cases}$$

We also have a symmetric property, $A_{ijr} = A_{jir}$, which simplifies the evaluation of the test statistic from $O(n^3)$ to $O\left((n^3 + n^2)/2\right)$ computations. The memory requirement is expensive, because we need to store the $(n^3 + n^2)/2$ 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:

$$\text{PCvM}_{n,p} = n^{-2} \hat{\boldsymbol{\varepsilon}}^T \mathbf{A}_\bullet \hat{\boldsymbol{\varepsilon}}, \tag{9}$$

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, $O\left((n^3 - n^2)/2\right)$, the memory required for storing the matrix $\mathbf{A}_\bullet$ is substantially lower and drops to $(n^2 - n + 2)/2$ 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 $\text{PCvM}_{n,p}$ 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 the statistic $\text{PCvM}_{n,p}$ under the null hypothesis, 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 $\hat{\beta}^{(p)}$ 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 $\mathbb{E}^*\left[V_i^*\right] = 0$ and $\mathbb{E}^*\left[V_i^{*2}\right] = 1$. For example, if V^* is a discrete random variable with distribution weights $\mathbb{P}\left\{V^* = \frac{1-\sqrt{5}}{2}\right\} = \frac{5+\sqrt{5}}{10}$ $\mathbb{P}\left\{V^* = \frac{1+\sqrt{5}}{2}\right\} = \frac{5-\sqrt{5}}{10}$, 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 $\left\{\left(\mathcal{X}_i, Y_i^*\right)\right\}_{i=1}^n$.

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

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 (9) of the previous section. 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: $\#\left\{\mathrm{PCvM}_{n,p} \leq \mathrm{PCvM}_{n,p}^{*,b}\right\}/B$. 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 (1) it is clear that the vector of coefficients of $\hat{\beta}^{(p)}$ is estimated throughout $\hat{\mathbf{b}} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Y}$. Then, the estimated bootstrap residuals, represented by the vector $\hat{\boldsymbol{\varepsilon}}^*$, can be obtained as $\hat{\boldsymbol{\varepsilon}}^* = \left(\mathbf{I}_p - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\right)\mathbf{Y}^*$, where $\mathbf{Y}^*$ is the vector of bootstrap responses given in step IV and $\mathbf{I}_p$ is the identity matrix of order p. The projection matrix $\left(\mathbf{I}_p - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\right)$ 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 $\hat{\varepsilon}_i^* = \varepsilon_i^*,\ i = 1, \ldots, n$.

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. The functional process considered for the functional covariate $\mathcal{X}$ is an Ornstein–Uhlenbeck process in $[0, 1]$, which 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 $\mu(t) = 0,\ \forall t \in [0, 1]$. See Figure B2 in appendix for further details.

All the functional data in this simulation study is represented in 201 equidistant points in the interval $[0, 1]$. The number of bootstrap replicates considered is $B = 1000$ and the number of Monte Carlo replicates for determining the empirical sizes and powers, $M = 1000$. 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.

Several lengthy tables have been reduced in this section for space saving. The reader is referred to the appendix in supplementary materials to see the whole tables as well as other explanatory figures.

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 $H_0 : m(\mathcal{X}) = 0$. Although there is an extensive collection of goodness–of–fit tests for finite dimensional covariates (see González-Manteiga and Crujeiras (2011)), the literature for the case of functional covariates 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.

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 by the proposal 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. 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. The bandwidth parameter is given by the PCV criterion and 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.

Three different blocks of deviations from the null are considered. The first two blocks represent 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 are $\beta_{1,k}(t) = \gamma_k \cdot (t - 0.5)$, with coefficients $\gamma_1 = 0.25$, $\gamma_2 = 0.65$ and $\gamma_3 = 1.00$ for $H_{1,k}$ and $\beta_{2,k}(t) = \eta_k \cdot \sin(2\pi t^3)^3$, with $\eta_1 = 0.10$, $\eta_2 = 0.20$ and $\eta_3 = 0.50$ for $H_{2,k}$. The second block of deviations 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 $H_{3,k} : Y = \langle \mathcal{X}, \beta_0 \rangle + \delta_k \langle \mathcal{X}, \mathcal{X} \rangle + \varepsilon$, where $\delta_1 = 0.005$, $\delta_2 = 0.010$ and $\delta_3 = 0.015$. The relation between the variance of the response with respect to the variance of the error can be measured by the *signal–to–noise ratio*: $\text{snr} = \mathbb{E}\left[m(\mathcal{X})^2\right] / (\mathbb{E}\left[m(\mathcal{X})^2\right] + \sigma^2)$. For block 1 the snr's of the alternatives are 0.044, 0.235 and 0.421, respectively for $H_{1,k}$, $k = 1, 2, 3$. For block 2, the snr's are 0.019, 0.150 and 0.329. For block 3, we have 0.015, 0.086 and 0.272.

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. To this end, we consider a B–splines basis and we choose automatically 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 Table 1, which shows the empirical sizes and powers of the functional F–test, the test of Delsol et al. and the PCvM test for simple hypothesis, for the models previously commented. All of the tests seem to calibrate the significance level $\alpha = 0.05$. With respect to the power, the functional F–test has in average a superior behaviour in the alternatives $H_{2,k}$, which represents deviations from the null *inside* a linear model. The test of Delsol et al. performs also well with the cross–validatory bandwidth, being the most competitive for the block $H_{1,k}$. The PCvM test has lower power than the functional F–test for alternatives $H_{1,k}$ and $H_{2,k}$ and similar or lower power to the test of Delsol et al., which is in part favoured by an over–rejection of the null hypothesis. Nevertheless, for alternatives that are not inside the linear model, the PCvM test results the most powerful. Metrics L^1, L^∞ and four L^2–based semi–metrics were considered also for the test of Delsol et al., without obtaining better results than with the L^2 metric. Similar results are obtained with a noise given by a recentred exponential distribution with parameter $\lambda = 10$.

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

$$H_{j,0}: \quad Y = \left\langle \mathcal{X}, \beta_j \right\rangle + \varepsilon, \tag{10}$$

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 second functional coefficient is chosen to be perfectly described by B–splines, whereas this is

not the case for β_1 and β_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 $H_{j,k}$:

$$H_{j,k}: \quad Y = \left\langle \mathcal{X}, \beta_j \right\rangle + \delta_k \langle \mathcal{X}, \mathcal{X} \rangle + \varepsilon. \tag{11}$$

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 snr's for model 1 are 0.823, 0.824, 0.834 and 0.861, respectively for $H_{1,k}$, $k = 0, 1, 2, 3$. For model 2, 0.949, 0.949, 0.950 and 0.953. For model 3, we have 0.971, 0.971, 0.971 and 0.972.

Three estimation methods for the functional parameter β will be considered. All of them are designed in order to provide automatic selectors of the number of elements considered in the basis estimation of β. So, the first automatic method considered is the estimation of β as a linear combination of a B–splines basis of p elements, where p is chosen by the GCV criteria (2). Secondly, FPC estimation relies on the BIC criteria to choose the optimal number of elements in the FPC basis derived from the process 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 $\{(\mathcal{X}_i, Y_i)\}_{i=1}^n$.

Table 2 shows the rejection frequencies of the null hypothesis for the test computed from observations of the null hypotheses (10) and deviations (11), for the significance level $\alpha = 0.05$. 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 the 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 for the three estimation methods considered. Secondly, there seems to be no big differences in terms of power for the three methods, although it can be observed that the FPC and FPLS estimation methods are slightly more conservative. Finally, at sight of the similarities between the response under the null and under the alternatives (see Figure B4 in appendix), the results of Table 2 point toward a quite competitive test. Similar results are obtained with a non symmetric random noise.

The behaviour of the test for different sample sizes is shown in Table 3. As in the previous tables, the three estimating methods have very similar rejection ratios and we can see that B–splines estimation has again larger rejection ratios for all the models. As expected, when the sample sizes increases, the rejection rates also do.

5 Data application and graphical tool

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 of the meat samples with high percentage of fat (see the left plot of Figure 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). The center plot of Figure 1 represents the functional observations of the daily temperature. 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 1 shows the estimated 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 test of Delsol et al. (with PCV bandwidth) and our test for the simple null hypothesis of no interaction. The p–values obtained are: 0.002, 0.000 and 0.000, respectively. All the tests reject the null, so we can conclude that the curves of the temperature and the average wind speed show a non–trivial linear relation.

We conclude this section 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) 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 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.

6 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 using 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.

The asymptotic distribution of the statistics PCvM_n and $\text{PCvM}_{n,p}$, quadratic functionals of the processes R_n and $R_{n,p}$, respectively, is an open problem. The convergence of both processes remains as a problem of great relevance to be considered in the future, taking into account that these

processes are indexed in $\mathbb{R} \times \mathbb{H}$ and that it does not exist, up to our knowledge, any results of weak functional convergence of empirical processes indexed in infinite dimensional spaces.

Although in this paper 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 the statistic is based on the residuals, the practical implementation and the wild bootstrap calibration given in Section 3 will remain the same: we just have to consider suitable estimators for the parameters of the regression model to compute the residuals. Therefore, obvious extensions could be the testing of FLM with several covariates or the testing of the quadratic functional model.

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` since version 0.9.8. This function also shows the graphical tool introduced in Section 5. To speed up the computation of the test statistic, the critical parts of the test implementation have been programmed in FORTRAN.

SUPPLEMENTAL MATERIALS

Appendix: Contains the proof of Lemma 2, explaining figures and more detailed tables for the results of the simulation study. (pdf file)

R-package for `flm.test`, `flm.Ftest` and `dfv.test` routines: R-package `fda.usc` containing code to perform the testing methods described in the article. The package also contains the AEMET and Tecator datasets used as examples in the article. (GNU zipped tar file)

References

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.

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 1.0.3.

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. arXiv:1210.1072.

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.

Patilea, V., Sánchez-Sellero, C., 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.

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.

Models	F–test	PCvM	Delsol et al.	F–test	PCvM	Delsol et al.
H_0	0.060	0.041	0.065	0.043	0.051	0.066
$H_{1,1}$	0.060	0.069	0.098	0.056	0.052	0.072
$H_{1,2}$	0.163	0.078	0.309	0.180	0.085	0.285
$H_{1,3}$	0.401	0.138	0.772	0.442	0.166	0.719
$H_{2,1}$	0.248	0.053	0.080	0.265	0.071	0.089
$H_{2,2}$	0.951	0.336	0.403	0.932	0.343	0.420
$H_{2,3}$	1.000	0.904	0.877	0.999	0.901	0.848
$H_{3,1}$	0.034	0.173	0.165	0.052	0.125	0.128
$H_{3,2}$	0.038	0.691	0.554	0.034	0.721	0.558
$H_{3,3}$	0.019	0.998	0.932	0.012	1.000	0.967

Table 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 follows a $\mathcal{N}(0, 0.10^2)$ and a recentred Exp(10).

Models	B–splines	FPC	FPLS	B–splines	FPC	FPLS
$H_{1,0}$	0.061	0.052	0.059	0.039	0.046	0.046
$H_{1,1}$	0.094	0.082	0.078	0.074	0.072	0.077
$H_{1,2}$	0.747	0.732	0.715	0.737	0.721	0.720
$H_{1,3}$	0.997	0.997	0.996	0.996	0.997	0.996
$H_{2,0}$	0.058	0.045	0.050	0.041	0.035	0.033
$H_{2,1}$	0.086	0.071	0.074	0.081	0.080	0.078
$H_{2,2}$	0.745	0.722	0.720	0.743	0.724	0.718
$H_{2,3}$	0.997	0.996	0.997	0.994	0.995	0.994
$H_{3,0}$	0.054	0.046	0.044	0.052	0.040	0.038
$H_{3,1}$	0.082	0.077	0.075	0.072	0.062	0.062
$H_{3,2}$	0.764	0.752	0.750	0.735	0.737	0.721
$H_{3,3}$	0.999	0.998	0.998	0.998	0.998	0.997

Table 2: 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 at significance level $\alpha = 0.05$ with noise $\mathcal{N}(0, 0.10^2)$ (first three columns) and recentred Exp(0.10) (last three).

	$H_{1,0}$			$H_{1,1}$			$H_{1,2}$			$H_{1,3}$		
Method	50	100	200	50	100	200	50	100	200	50	100	200
B–spline	0.076	0.061	0.062	0.093	0.094	0.121	0.484	0.747	0.966	0.900	0.997	1.000
FPC	0.059	0.052	0.059	0.064	0.082	0.123	0.442	0.732	0.963	0.893	0.997	1.000
FPLS	0.062	0.059	0.058	0.069	0.078	0.115	0.414	0.715	0.961	0.873	0.996	1.000

Table 3: 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 is a $\mathcal{N}(0, 0.10^2)$.

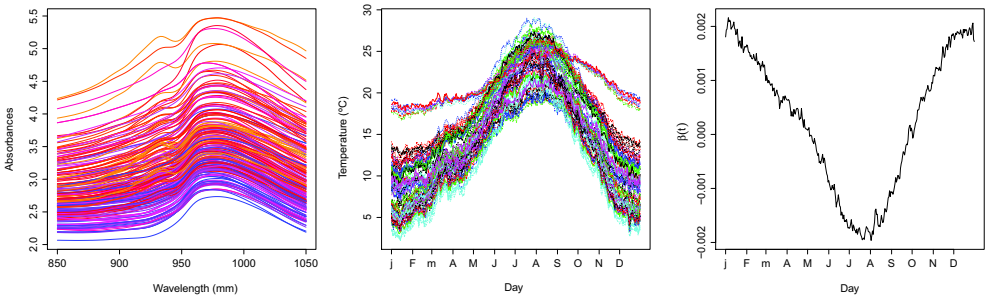

Figure 1: From left to right: Tecator dataset with spectrometric curves coloured according to their content of fat (red for larger and blue for lower); AEMET temperatures for the 73 Spanish weather stations; estimated functional coefficient by the FPLS method for the AEMET dataset.

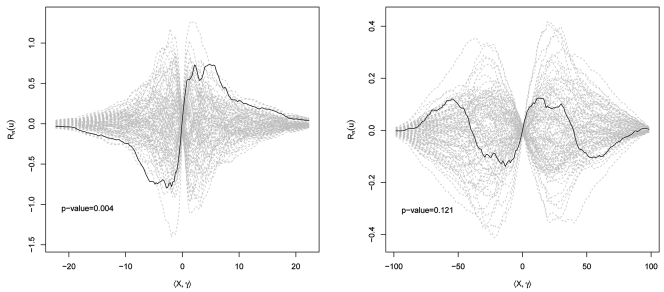

Figure 2: R_n process observed (solid line) and $B = 100$ generated process under the null hypothesis $H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$ (dashed lines), for the Tecator dataset (left) and the AEMET dataset (right). The number of Monte Carlo replicates for the projections is $G = 200$.