UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE
ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

Bootstrap independence test for functional linear models

W. González-Manteiga, G. González-Rodríguez, A. Martínez-Calvo, E. García-Portugués

Report 12-02

Reports in Statistics and Operations Research

Bootstrap independence test for functional linear models

Wenceslao González–Manteiga[1], Gil González–Rodríguez[2],

Adela Martínez–Calvo[1,3] and Eduardo García–Portugués[1]

Abstract

Functional data have been the subject of many research works over the last years. Functional regression is one of the most discussed issues. Specifically, significant advances have been made for functional linear regression models with scalar response. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a separable Hilbert space. We focus on the model $Y = \langle \Theta, X \rangle + b + \varepsilon$, where Y and ε are real random variables, X is an $\mathcal{H}$–valued random element, and the model parameters b and Θ are in $\mathbb{R}$ and $\mathcal{H}$, respectively. Furthermore, the error satisfies that $E(\varepsilon|X) = 0$ and $E(\varepsilon^2|X) = \sigma^2 < \infty$. A consistent bootstrap method to calibrate the distribution of statistics for testing $H_0 : \Theta = 0$ versus $H_1 : \Theta \neq 0$ is developed. The asymptotic theory, as well as a simulation study and a real data application illustrating the usefulness of our proposed bootstrap in practice, is presented.

Keywords: Bootstrap; bootstrap consistency; functional linear regression; functional principal components analysis; hypothesis test.

1 Introduction

Nowadays, *Functional Data Analysis* (FDA) has turned into one of the most interesting statistical fields. Particularly, functional regression models have been studied from a parametric point of view (see Ramsay and Silverman (2002, 2005)), and from a non–parametric one (see Ferraty and Vieu (2006)), being the most recent advances compiled on Ferraty and Romain (2011). This work focuses on the parametric approach, specifically, on the *functional linear regression model with scalar response* that is described below.

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a separable Hilbert space, and let $\|\cdot\|$ be the norm associated with its inner product. Moreover, let (Ω, σ, P) be a probability space and let us consider (X, Y) a measurable mapping from Ω to $\mathcal{H} \times \mathbb{R}$, that is, X is an $\mathcal{H}$–valued random element whereas Y is a real random variable. In this situation, let us assume that (X, Y) verifies the following linear model with scalar response,

$$Y = \langle \Theta, X \rangle + b + \varepsilon \tag{1}$$

where $\Theta \in \mathcal{H}$ is a fixed functional model parameter, $b \in \mathbb{R}$ is the intercept term, and ε is a real random variable such that $E(\varepsilon|X) = 0$ and $E(\varepsilon^2|X) = \sigma^2 < \infty$. Many authors have dealt with model (1), being the methods based on *Functional Principal Components Analysis* (FPCA) amongst the most popular ones to estimate the model parameters (see Cardot, Ferraty, and Sarda (1999, 2003), Cai and Hall (2006), Hall and Hosseini-Nasab (2006), and Hall and Horowitz (2007)).

The main aim of this work is to develop a consistent general bootstrap resampling approach to calibrate the distribution of statistics for testing the significance of the relationship between X and

---

[1]Department of Statistics and Operations Research. University of Santiago de Compostela (Spain).

[2]Department of Statistics and Operations Research and Mathematics Didactics. University of Oviedo (Spain).

[3]Corresponding author. e–mail: adela.martinez@usc.es.

Y, that is, for testing $H_0 : \Theta = 0$ versus $H_1 : \Theta \neq 0$, on the basis of a simple random sample $\{(X_i, Y_i)\}_{i=1}^{n}$ drawn from (X, Y). The bootstrap techniques will become an alternative useful tool when the asymptotics of test statistics are unknown or when they are inaccurate due to small sample size.

Since its introduction by Efron (1979), it is well–known that the bootstrap method results in a new distribution approximation which can be applied to a large number of situations, such as the calibration of pivotal quantities in the finite dimensional context (see Bickel and Freedman (1981), and Singh (1981)). As far as multivariate regression models are concerned, bootstrap validity for linear and non–parametric models was also stated in literature (see Freedman (1981), and Cao-Abad (1991)). Currently, the application of bootstrap to the functional field has been successfully started. For instance, Cuevas, Febrero, and Fraiman (2006) have proposed bootstrap confidence bands for several functional estimators such as the sample and the trimmed functional means. In the regression context, Ferraty, Van Keilegom, and Vieu (2010), and González-Manteiga and Martínez-Calvo (2011) have shown the validity of the bootstrap in the estimation of non–parametric functional regression and functional linear model, respectively, when the response is scalar. They have also proposed pointwise confidence intervals for the regression operator involved in each case. In addition, the asymptotic validity of a componentwise bootstrap procedure has been proved by Ferraty, Van Keilegom, and Vieu (2012) when a non–parametric regression is considered and both response and regressor are functional.

Bootstrap techniques can also be very helpful for testing purposes, since they can be used in order to approximate the distribution of the statistic under the null hypothesis H_0. For example, Cuevas, Febrero, and Fraiman (2004) have developed a sort of parametric bootstrap to obtain quantiles for an ANOVA test, and González-Rodríguez, Colubi, and Gil (2012) have proved the validity of a residual bootstrap in that context. Hall and Vial (2006) and, more recently, Bathia, Yao, and Ziegelmann (2010) have studied the finite dimensionality of functional data using a bootstrap approximation for independent and dependent data, respectively.

As was indicated previously, testing the lack of dependence between X and Y is our goal. This issue has stirred up a great interest during the last years due to its practical applications in the functional context. For instance, Kokoszka, Maslova, Sojka, and Zhu (2008) proposed a test for lack of dependence in the functional linear model with functional response which was applied to magnetometer curves consisting of minute–by–minute records of the horizontal intensity of the magnetic field measured at observatories located at different latitude. The aim was to analyse if the high–latitude records had a linear effect on the mid– or low–latitude records. On the other hand, Cardot, Prchal, and Sarda (2007) presented a statistical procedure to check if a real–valued covariate has an effect on a functional response in a nonparametric regression context, using this methodology for a study of atmospheric radiation. In this case, the dataset were radiation profiles curves measured at a random time and the authors tested if the radiation profiles changed along the time.

Regarding the regression model (1), testing the significance of the relationship between a functional covariate and a scalar response has been the subject of recent contributions, and asymptotic approaches for this problem can be found in Cardot, Ferraty, Mas, and Sarda (2003) or Kokoszka, Maslova, Sojka, and Zhu (2008). The methods presented in these two works are mainly based on the calibration of the statistics distribution by using asymptotic distribution approximations. In contrast, we propose a consistent bootstrap calibration in order to approximate the statistics distribution. For that, we firstly introduce in Section 2 some notation and basic concepts about the regression model (1), the asymptotic theory for the testing procedure, and the consistency of the bootstrap techniques that we propose. In Section 3, the bootstrap calibration is presented as an alternative to the asymptotic theory previously exposed. Then, Section 4 is devoted to the empirical

results. A simulation study and a real data application allow us to show the performance of our bootstrap methodology in comparison with the asymptotic approach. Finally, some conclusions are summarized in Section 5.

2 Asymptotic theory and bootstrap

Let us consider the model (1) given in the previous Section 1. In this framework, the regression function, denoted by m, is given by

$$m(x) = E(Y|X = x) = \langle \Theta, x \rangle + b \text{ for all } x \in \mathcal{H}.$$

The aim is to develop correct and consistent bootstrap techniques for testing

$$\begin{cases} H_0: & \Theta = 0 \\ H_1: & \Theta \neq 0 \end{cases} \tag{2}$$

on the basis of a random sample $\{(X_i, Y_i)\}_{i=1}^n$ of independent and identically distributed random elements with the same distribution as (X, Y). That is, our objective is to check whether X and Y are linearly independent (H_0) or not (H_1).

Next, we expose briefly some technical background required to develop the theoretical results presented throughout the section.

2.1 Some background

Riesz Representation Theorem ensures that the functional linear model with scalar response can be handled theoretically within the considered framework. Specifically, let $\mathcal{H}$ be the separable Hilbert space of square Lebesgue integrable functions on a given compact set $C \subset \mathbb{R}$, denoted by $\mathcal{L}^2(C, \lambda)$, with the usual inner product and the associated norm $\|\cdot\|$. The functional linear model with scalar response between a random function X and a real random variable Y is defined as

$$Y = \Phi(X) + \epsilon, \tag{3}$$

where Φ is a continuous linear operator (that is, $\Phi \in \mathcal{H}'$, being $\mathcal{H}'$ the dual space of $\mathcal{H}$ with norm $\|\cdot\|'$), and ϵ is a real random variable with finite variance and independent of X. In virtue of Riesz Representation Theorem $\mathcal{H}$ and $\mathcal{H}'$ are isometrically identified, in such a way that for any $\Phi \in \mathcal{H}'$ there exists a unique $\Theta \in \mathcal{H}$ so that $\|\Theta\| = \|\Phi\|'$ and $\Phi(h) = \langle \Theta, h \rangle$ for all $h \in \mathcal{H}$. Consequently, the model presented in equation (3) is just a particular case of the one considered in (1).

Previous works regarding functional linear models assume $b = 0$ (see Cardot, Ferraty, Mas, and Sarda (2003), and Kokoszka, Maslova, Sojka, and Zhu (2008)). Of course, the intercept term can be embedded in the variable counterpart of the model as in the multivariate case as follows. Let $\mathcal{H}_e$ be the product space $\mathcal{H} \times \mathbb{R}$ with the corresponding inner product $\langle \cdot, \cdot \rangle_e$, and define $X' = (X, 1)$ and $\Theta' = (\Theta, b) \in \mathcal{H}_e$. Then the model considered in (1) can be rewritten as $Y = \langle \Theta', X' \rangle_e + \varepsilon$ (and consequently X' cannot be assumed to be centered). Nevertheless, in the context of the linear independence test, the aim is to check if $\Theta = 0$ or not, and this is not equivalent to checking whether $\Theta' = 0$ or not. In addition, in practice the intercept term b cannot be assumed to be equal to 0. Thus, in order to avoid any kind of confusion, in this paper the intercept term b has been written explicitly.

In the same way, in the above mentioned papers, the random element X is assumed to be centered. Although, in many cases, the asymptotic distribution of the proposed statistics does not change if $\{X_i\}_{i=1}^n$ is replaced by the dependent sample $\{X_i - \overline{X}\}_{i=1}^n$, the situation with the bootstrap version of the statistics could be quite different. In fact, as it will be shown afterwards, different bootstrap statistics could be considered when this replacement is done. Hence, for the developments in this section, it will not be assumed that the X variable is centered.

2.2 Linear independence test

Given a generic $\mathcal{H}$–valued random element H such that $E(\|H\|^2) < \infty$, its associated covariance operator Γ_H is defined as the operator $\Gamma_H : \mathcal{H} \to \mathcal{H}$

$$\Gamma_H(h) = E\left(\langle H - \mu_H, h \rangle (H - \mu_H)\right) = E\left(\langle H, h \rangle H\right) - \langle \mu_H, h \rangle \mu_H,$$

for all $h \in \mathcal{H}$, where $\mu_H \in \mathcal{H}$ denotes the expected value of H. From now on, it will be assumed that $E(\|X\|^2) < \infty$, and thus, as a consequence of Hölder's inequality, $E(Y^2) < \infty$. Whenever there is no possible confusion, Γ_X will be abbreviated as Γ. It is well–known that Γ is a nuclear and self–adjoint operator. In particular, it is a compact operator of trace class and thus, in virtue of the Spectral Theorem Decomposition, there is an orthonormal basis of H, $\{v_j\}_{j \in \mathbb{N}}$, consisting on eigenvectors of Γ with corresponding eigenvalues $\{\lambda_j\}_{j \in \mathbb{N}}$, that is, $\Gamma(v_j) = \lambda_j v_j$ for all $j \in \mathbb{N}$. As usual, the eigenvalues are assumed to be arranged in decreasing order ($\lambda_1 \geq \lambda_2 \geq \ldots$). Since the operator Γ is symmetric and non–negative definite, then the eigenvalues are non–negative.

In a similar way, let us consider the cross–covariance operator $\Delta : \mathcal{H} \to \mathbb{R}$ between X and Y given by

$$\Delta(h) = E\left(\langle X - \mu_X, h \rangle (Y - \mu_Y)\right) = E\left(\langle X, h \rangle Y\right) - \langle \mu_X, h \rangle \mu_Y,$$

for all $h \in \mathcal{H}$, where $\mu_Y \in \mathbb{R}$ denotes the expected value of Y. Of course, $\Delta \in \mathcal{H}'$ and the following relation between the considered operators and the regression parameter Θ is satisfied

$$\Delta(\cdot) = \langle \Gamma(\cdot), \Theta \rangle. \tag{4}$$

The Hilbert space $\mathcal{H}$ can be expressed as the direct sum of the two orthogonal subspaces induced by the self–adjoint operator Γ: the kernel or null space of Γ, $\mathcal{N}(\Gamma)$, and the closure of the image or range of Γ, $\overline{\mathcal{R}(\Gamma)}$. Thus, Θ is determined uniquely by $\Theta = \Theta_1 + \Theta_2$ with $\Theta_1 \in \mathcal{N}(\Gamma)$ and $\Theta_2 \in \overline{\mathcal{R}(\Gamma)}$. As $\Theta_1 \in \mathcal{N}(\Gamma)$, it is easy to check that $Var(\langle X, \Theta_1 \rangle) = 0$ and, consequently, the model introduced in (1) can be expressed as

$$Y = \langle \Theta_2, X \rangle + \langle \Theta_1, \mu_X \rangle + b + \varepsilon.$$

Therefore, it is not possible to distinguish between the term $\langle \Theta_1, \mu_X \rangle$ and the intercept term b, and consequently it is not possible to check whether $\Theta_1 = 0$ or not. Taking this into account, the hypothesis test will be restricted to check

$$\begin{cases} H_0 : & \Theta_2 = 0 \\ H_1 : & \Theta_2 \neq 0 \end{cases} \tag{5}$$

on the basis of the available sample information.

Note that in this case, according to the relation between the operators and the regression parameter shown in (4), $\Theta_2 = 0$ if, and only if, $\Delta(h) = 0$ for all $h \in \mathcal{H}$. Consequently, the hypothesis test in (5) is equivalent to

$$\begin{cases} H_0 : & \|\Delta\|' = 0 \\ H_1 : & \|\Delta\|' \neq 0 \end{cases} \tag{6}$$

Remark 1. It should be recalled that, in previous works μ_X is assumed to be equal 0. Thus, the preceding reasoning leads to the fact that Θ_1 cannot be estimated based on the information provided by X (see, for instance, Cardot, Ferraty, Mas, and Sarda (2003)). Consequently the hypothesis testing is also restricted to the one in the preceding equations. In addition in Cardot, Ferraty, Mas, and Sarda (2003), it is also assumed for technical reasons that $\overline{\mathcal{R}(\Gamma)}$ is an infinite–dimensional space. On the contrary, this restriction is not imposed in the study here developed.

Remark 2. Note that another usual assumption is that the intercept term vanishes. Although this is not common in most of situations, it should be noted that if $b = 0$ and X is not assumed to be centered (as in this work), then an interesting possibility appears: to check whether $\Theta_1 = 0$ or not by checking the nullity of the intercept term of the model, and thus to check the original hypothesis testing in (2). This open problem cannot be solved with the methodology employed in the current paper (or in the previous ones) because the idea is based on checking (6), which is equivalent to the restricted test (5) but not to the unrestricted one in (2).

2.3 Testing procedure and asymptotic theory

According to the relation between $\|\cdot\|'$ and $\|\cdot\|$, the dual norm of $\Delta \in \mathcal{H}'$ can be expressed equivalently in terms of the $\mathcal{H}$–valued random element $(X - \mu_X)(Y - \mu_Y)$ as follows

$$\|\Delta\|' = \|\langle E\left((X - \mu_X)(Y - \mu_Y)\right), \cdot\rangle\|' = \|E\left((X - \mu_X)(Y - \mu_Y)\right)\|.$$

Thus, based on an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$ drawn from (X, Y),

$$D = \|E\left((X - \mu_X)(Y - \mu_Y)\right)\| = \|T\|$$

can be estimated in a natural way by means of its empirical counterpart $D_n = \|T_n\|$, where T_n is the $\mathcal{H}$–valued random element given by

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

where $\overline{X}$ and $\overline{Y}$ denote as usual the corresponding sample means. The next theorem establishes some basic properties of T_n.

Theorem 1. *Assuming that* (1) *holds with* $E(\varepsilon) = 0$, $E(\varepsilon^2) = \sigma^2 < \infty$ *and* $E(\|X\|^4) < \infty$, *then*

1. $E(T_n) = E\left((X - \mu_X)(Y - \mu_Y)\right)(n-1)/n$

2. T_n *converges a.s.–P to* $E\left((X - \mu_X)(Y - \mu_Y)\right)$ *as* $n \to \infty$

3. $\sqrt{n}\left(T_n - E\left((X - \mu_X)(Y - \mu_Y)\right)\right)$ *converges in law, as* $n \to \infty$, *to a centered Gaussian element* Z *in* $\mathcal{H}$ *with covariance operator*

$$\Gamma_Z(\cdot) = \sigma^2 \Gamma(\cdot) + E\left((X - \mu_X)\langle X - \mu_X, \cdot\rangle\langle X - \mu_X, \Theta\rangle^2\right).$$

Proof. Since T_n can be equivalently expressed as

$$T_n = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y) + (\overline{X} - \mu_X)(\overline{Y} - \mu_Y),$$

it is straightforward to check item 1. The a.s.–P convergence is a direct application of the SLLN for separable Hilbert–valued random elements.

On the other hand, given that $E(\|(X - \mu_X)(Y - \mu_Y)\|^2) < \infty$, the convergence in law can be deduced by applying the CLT for separable Hilbert–valued random elements (see, for instance, Laha and Rohatgi (1979)) together with Slutsky's Theorem. The concrete expression of the operator Γ_Z, that is, $\Gamma_Z = \Gamma_{(X-\mu_X)(Y-\mu_Y)} = \Gamma_{(X-\mu_X)\varepsilon} + \Gamma_{(X-\mu_X)\langle X-\mu_X, \Theta\rangle}$, can be obtained by simple computations. $\qquad\square$

In order to simplify the notation, from now on, given any $\mathcal{H}$–valued random element H with $E(\|H\|^2) < \infty$, Z_H will denote a centered Gaussian element in $\mathcal{H}$ with covariance operator Γ_H.

Corollary 1. *Under the conditions of Theorem 1, if the null hypothesis $H_0 : \|\Delta\|' = 0$ is satisfied, then $\sqrt{n}T_n$ converges in law to $Z_{(X-\mu_X)\varepsilon}$ (with covariance operator $\sigma^2\Gamma$), and consequently, $\|\sqrt{n}T_n\|$ converges in law to $\|Z_{(X-\mu_X)\varepsilon}\|$.*

In contrast to Theorem 1 in Cardot, Ferraty, Mas, and Sarda (2003), the result in Corollary 1 is established directly on the Hilbert space $\mathcal{H}$ instead of on its dual space. In addition, no assumption of centered X random elements or null intercept term is necessary. Nevertheless these two assumptions could be easily removed in that paper in order to establish a dual result of Corollary 1.

Furthermore, in view of Corollary 1, the asymptotic null distribution of $\|\sqrt{n}T_n\|$ is not explicitly known. This is the reason why no further research on how to use in practice this statistic (or its dual one) for checking if Θ_2 equals 0 is carried out in Cardot, Ferraty, Mas, and Sarda (2003). Instead, an alternative statistic that is used in the simulation section for comparative purposes is considered. Nevertheless, it is still possible to use $\|\sqrt{n}T_n\|$ as a core statistic in order to solve this test in practice by means of bootstrap techniques.

One natural way of using the asymptotic result of Corollary 1 for solving the test under study is as follows. Consider a consistent (at least under H_0) estimator σ_n^2 of σ^2 (for instance, the sample variance of Y could be used, or perhaps the one introduced by Cardot, Ferraty, Mas, and Sarda (2003), provided that its theoretical behavior is analyzed). Then, according to Slutsky's Theorem $\|\sqrt{n}T_n\|/\sigma_n$ converges in law under H_0 to the norm of Z_X. As its covariance operator Γ is unknown, it is approximated by the empirical one Γ_n. And thus, $\|Z_X\|$ can be approximated by $\|Z_n\|$, being Z_n a centered Gaussian element in $\mathcal{H}$ with covariance operator Γ_n. Of course, the distribution of $\|Z_n\|$ is still difficult to compute directly, nevertheless one can make use of the CLT and approximate its distribution by Monte Carlo method by the distribution of

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (X_i^* - \overline{X}) \right\|$$

for a large value of m, being $\{X_i^*\}_{i=1}^{m}$ i.i.d. random elements chosen at random from the fixed population $(X_1, \ldots, X_n)$. Obviously, this method is a precursor of the bootstrap procedures.

In order to complete the asymptotic study of the statistic $\|\sqrt{n}T_n\|$, its behavior under local alternatives is going to be analyzed. To this purpose, let us consider $\Theta \in \mathcal{H}$ so that $\|\Theta_2\| > 0$, and given $\delta_n > 0$ consider the modified random sample

$$Y_i^n = \langle X_i, \frac{\delta_n}{\sqrt{n}}\Theta \rangle + b + \varepsilon_i,$$

for all $i \in \{1, \ldots, n\}$. Then, the null hypothesis is not verified. However, if $\delta_n/\sqrt{n} \to 0$, then $\|(\delta_n/\sqrt{n})\Theta\| \to 0$, that is, H_0 is approached with "speed" $\delta_n/\sqrt{n}$. In these conditions,

$$E\left((X_i - \mu_{X_i})(Y_i^n - \mu_{Y_i^n})\right) = \frac{\delta_n}{\sqrt{n}}\Delta(\Theta),$$

and thus the following theorem that establishes the behavior of the statistic under the considered local alternatives can be easily deduced.

Theorem 2. *Under the conditions of Theorem 1 and with the above notation, if $\delta_n \to \infty$ and $\delta_n/\sqrt{n} \to 0$ as $n \to \infty$ then*

$$P\left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i^n - \overline{Y^n}) \right\| \leq t \right) \to 0$$

as $n \to \infty$, for all $t \in \mathbb{R}$.

2.4 Bootstrap procedures

The difficulty of using the previously proposed statistic to solve the hypothesis test by means of asymptotic procedures suggests the development of appropriated bootstrap techniques. The asymptotic consistency of a bootstrap approach is guaranteed if the associated bootstrap statistic converges in law to a non–degenerated distribution irrespectively of H_0 being satisfied or not. In addition, in order to ensure its asymptotic correctness, this limit distribution must coincide with the asymptotic one of the testing statistic provided that H_0 holds.

Consequently, the asymptotic limit established in Corollary 1 plays a fundamental role for defining appropriate bootstrap statistics. In this way, recall that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right) - E \left((X - \mu_X)(Y - \mu_Y) \right) \right)$$

converges in law to $Z_{(X-\mu_X)(Y-\mu_Y)}$, irrespectively of H_0 being satisfied or not and, in addition, if H_0 is satisfied then $\Gamma_{(X-\mu_X)(Y-\mu_Y)} = \sigma^2 \Gamma$. Thus, this is a natural statistic to be mimicked by a bootstrap one. Note that,

$$\left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_X) \right), \tag{7}$$

converges in law to $(\sigma^2 + E(\langle X - \mu_X, \Theta \rangle^2)) Z_X$, whose covariance operator is $(\sigma^2 + E(\langle X - \mu_X, \Theta \rangle^2)) \Gamma$. Particularly, when H_0 is satisfied, this operator reduces again to $\sigma^2 \Gamma$. Consequently, another possibility consists in mimicking this second statistic by means of a bootstrap one, improving the approximation suggested in the previous subsection. Note that the left term in the product in equation (7) could be substituted by any other estimator under H_0 of σ^2 that converges to a finite constant if H_0 does not hold. Anyway, this second approximation could lead to worst results under the null hypothesis, because the possible dependency between X and ε is lost (as the resample would focus only on the X information).

Two possibilities for mimicking the statistics which were above–mentioned are going to be explored, namely a "naive" paired bootstrap and a "wild" bootstrap approach. In order to achieve this goal, let $\{(X_i^*, Y_i^*)\}_{i=1}^n$ be a collection of i.i.d. random elements drawn at random from $(X_1, Y_1), \ldots, (X_n, Y_n)$, and let us consider the following "naive" paired bootstrap statistic

$$T_n^{N*} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left(X_i^* - \overline{X^*} \right) \left(Y_i^* - \overline{Y^*} \right) - \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right) \right).$$

In addition, let us consider $\sigma_n^2 = (1/n) \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ and $\sigma_n^{*2} = (1/n) \sum_{i=1}^{n} (Y_i^* - \overline{Y^*})^2$, the empirical estimator of σ_Y^2 under H_0 and its corresponding bootstrap version.

The asymptotic behavior of the "naive" bootstrap statistic will be analyzed through some results on bootstrapping general empirical measures obtained by Giné and Zinn (1990). It should be noted that the bootstrap results in that paper refer to empirical process indexed by a class of functions $\mathcal{F}$, that particularly extend to the bootstrap about the mean in separable Banach (and thus Hilbert) spaces. In order to establish this connection, it is enough to choose

$$\mathcal{F} = \{ f \in \mathcal{H}' | \|f\|' \leq 1 \}$$

(see Giné (1997) and Kosorok (2008), for a general overview of indexed empirical process). $\mathcal{F}$ is image admissible Suslin (considering the weak topology). In addition, $F(h) = \sup_{f \in \mathcal{F}} |f(h)| = \|h\|$

for all $h \in \mathcal{H}$ and thus $E(F^2(X)) = E(\|X\|^2) < \infty$.

Consider the bounded and linear (so continuous) operator δ from $\mathcal{H}$ to $l^\infty(\mathcal{F})$ given by $\delta(h)(f) = \delta_h(f) = f(h)$ for all $h \in \mathcal{H}$ and all $f \in \mathcal{F}$ and denote by $R(\delta) \subset l^\infty(\mathcal{F})$ its range. As $\|\delta(h)\| = \|h\|$ for all $h \in \mathcal{H}$ then, there exists $\delta^{-1} : R(\delta) \to \mathcal{H}$, so that δ^{-1} is continuous. In addition, as $R(\delta)$ is closed, Dugundji Theorem allows us to consider a continuous extension $\delta^{-1} : l^\infty(\mathcal{F}) \to \mathcal{H}$ (see for instance Kosorok (2008), Lemma 6.16 and Theorem 10.9). Thus, following the typical empirical process notation, the empirical process $(1/\sqrt{n}) \sum_{i=1}^n (\delta_{X_i} - \mathbb{P})$ indexed in $\mathcal{F}$ is directly connected with $(1/\sqrt{n}) \sum_{i=1}^n (X_i - E(X))$ by means of the continuous mapping δ^{-1} and vice–versa.

Some consequences of this formulation applied to the work developed by Giné and Zinn (1990) lead to the results collected in following lemma.

Lemma 1. *Let ξ be a measurable mapping from a probabilistic space denoted by (Ω, σ, P) to a separable Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ with corresponding norm $\| \cdot \|$ so that $E(\|\xi\|^2) < \infty$. Let $\{\xi_i\}_{i=1}^n$ be a sequence of i.i.d. random elements with the same distribution as ξ, and let $\{\xi_i^*\}_{i=1}^n$ be i.i.d. from $\{\xi_i\}_{i=1}^n$. Then*

1. *$\sqrt{n}(\overline{\xi^*} - \overline{\xi})$ converges in law to Z_ξ a.s.–P*

2. *$\overline{\xi^*}$ converges in probability to $E(\xi)$ a.s.–P*

3. *$\overline{\|\xi^*\|^2}$ converges in probability to $E(\|\xi\|^2)$ a.s.–P*

Proof. To prove item 1 note that the CLT for separable Hilbert–valued random elements (see, for instance, Laha and Rohatgi (1979)) together with the Continuous Mapping Theorem applied to δ guarantees that $\mathcal{F} \in \mathrm{CTL}(P)$. Thus, Theorem 2.4 of Giné and Zinn (1990) ensures that $n^{1/2}(\hat{P}_n(w) - P_n(w))$ converges in law to a Gaussian process on $\mathcal{F}$, $G = \delta(Z_\xi)$ a.s.–P. Consequently, by applying again the Continuous Mapping Theorem $\sqrt{n}(\overline{\xi^*} - \overline{\xi}) = \delta^{-1}(n^{1/2}(\hat{P}_n(w) - P_n(w)))$ converges in law to $Z_\xi = \delta^{-1}(G)$.

Items 2 and 3 can be checked in a similar way by applying Theorem 2.6 of Giné and Zinn (1990). Note that item 1 is also a direct consequence of Remark 2.5 of Giné and Zinn (1990); nevertheless it was proven based on Theorem 2.4 to illustrate the technique. $\square$

The following theorem establishes the asymptotic consistency and correctness of the "naive" bootstrap approach.

Theorem 3. *Under the conditions of Theorem 1, we have that $\sqrt{n}T_n^{N*}$ converges in law to $Z_{(X-\mu_X)(Y-\mu_Y)}$ a.s.–P. In addition, σ_n^{*2} converges in probability to $\sigma_Y^2 = \sigma^2 + E(\langle X - \mu_X, \Theta \rangle^2)$ a.s.–P.*

Proof. First of all consider the bootstrap statistic

$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\left(X_i^* - \mu_X\right)\left(Y_i^* - \mu_Y\right) - \overline{\left(X - \mu_X\right)\left(Y - \mu_Y\right)} \right)$$

and note that $\{\left(X_i^* - \mu_X\right)\left(Y_i^* - \mu_Y\right)\}_{i=1}^n$ are i.i.d. $\mathcal{H}$–valued random elements chosen at random from the "bootstrap population" $\{\left(X_i - \mu_X\right)\left(Y_i - \mu_Y\right)\}_{i=1}^n$. Then, item 1 in Lemma 1 guarantees that S_n^* converges in law to $Z_{(X-\mu_X)(Y-\mu_Y)}$ a.s.–P.

On the other hand, S_n^* equals $\sqrt{n}T_n^{N*}$ plus the following terms

$$\frac{1}{\sqrt{n}}\sqrt{n}(\overline{X^*} - \overline{X})\sqrt{n}(\overline{Y^*} - \overline{Y}) + \sqrt{n}(\overline{X^*} - \overline{X})(\overline{Y^*} - \mu_Y) + (\overline{X^*} - \mu_X)\sqrt{n}(\overline{Y^*} - \overline{Y}).$$

Items 1 and 2 in Lemma 1, together with Slutsky's Theorem, ensure that these three terms converge in probability to 0 a.s.–P, and consequently the convergence in law stated in the theorem is proven.

Finally, the convergence of σ_n^{*2} holds in virtue of items 2 and 3 in Lemma 1. $\qquad\square$

The "naive" bootstrap approach is described in the following algorithm.

Algorithm 1 (Naive Bootstrap).

Step 1. Compute the value of the statistic T_n (or the value T_n/σ_n).

Step 2. Draw $\{(X_i^*, Y_i^*)\}_{i=1}^n$, a sequence of i.i.d. random elements chosen at random from the initial sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, and compute $a_n = \|T_n^{N*}\|$ (or $b_n = \|T_n^{N*}\|/\sigma_n^*$).

Step 3. Repeat Step 2 a large number of times $B \in \mathbb{N}$ in order to obtain a sequence of values $\{a_n^l\}_{l=1}^B$ (or $\{b_n^l\}_{l=1}^B$).

Step 4. Approximate the p–value of the test by the proportion of values in $\{a_n^l\}_{l=1}^B$ greater than or equal to $\|T_n\|$ (or by the proportion of values in $\{b_n^l\}_{l=1}^B$ greater than or equal to $\|T_n\|/\sigma_n$)

Analogously, let $\{\varepsilon_i^*\}_{i=1}^n$ be i.i.d. centered real random variables so that $E\big((\varepsilon_i^*)^2\big) = 1$ and $\int_0^\infty (P(|\varepsilon_1| > t)^{1/2}) < \infty$ (to guarantee this last assumption, it is enough that $E\big((\varepsilon_i^*)^d\big) < \infty$ for certain $d > 2$), and consider the "wild" bootstrap statistic

$$T_n^{W*} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(X_i - \overline{X}\big)\big(Y_i - \overline{Y}\big)\varepsilon_i^*.$$

In order to analyze the asymptotic behavior of the "wild" bootstrap statistic, the following lemma will be fundamental. It is a particularization of a result due to Ledoux, Talagrand and Zinn (cf. Giné and Zinn (1990), and Ledoux and Talagrand (1988)). See also the Multiplier Central Limit Theorem in Kosorok (2008) for the empirical process indexed by a class of measurable functions counterpart.

Lemma 2. *Let ξ be a measurable mapping from a probabilistic space denoted by (Ω, σ, P) to a separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ with corresponding norm $\|\cdot\|$ so that $E(\|\xi\|^2) < \infty$. Let $\{\xi_i\}_{i=1}^n$ be a sequence of i.i.d. random elements with the same distribution as ξ, and let $\{W_i\}_{i=1}^n$ be a sequence of i.i.d. random variables (in the same probability space and independent of $\{\xi_i\}_{i=1}^n$) with $E(W_i) = 0$ and $\int_0^\infty (P(|W_1| > t)^{1/2}) < \infty$, then the following are equivalent*

1. $E(\|\xi\|^2) < \infty$ (and consequently $\sqrt{n}(\overline{\xi} - E(\xi))$ converges in law to Z_ξ).

2. For almost all $\omega \in \Omega$, $(1/\sqrt{n})\sum_{i=1}^n W_i\xi_i(\omega)$ converges in law to Z_ξ.

As a consequence, the asymptotic consistency and correctness of the "wild" bootstrap approach is guaranteed by the following theorem.

Theorem 4. *Under the conditions of Theorem 1, we get that $\sqrt{n}T_n^{W*}$ converges in law to $Z_{(X-\mu_X)(Y-\mu_Y)}$ a.s.–P.*

Proof. According to Lemma 2, for almost all $\omega \in \Omega$,

$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(X_i^w - \mu_X\big)\big(Y_i^w - \mu_Y\big)\varepsilon_i^*$$

converges in law to $Z_{(X-\mu_X)(Y-\mu_Y)}$. Moreover $(\overline{Y^w} - \mu_Y)$ and $(\overline{X^w} - \mu_X)$ converges to 0 (by SLLN).

Finally note that, for almost all $\omega \in \Omega$,

$$S_n^* = \sqrt{n}T_n^{W*} + (\overline{Y^w} - \mu_Y)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i^w - \mu_X)\varepsilon_i^*$$

$$+ (\overline{X^w} - \mu_X)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i^w - \mu_Y)\varepsilon_i^* + (\overline{X^w} - \mu_X)(\overline{Y^w} - \mu_Y)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i^*.$$

Lemma 2, together with the SLLN above–mentioned, guarantees the convergence in probability to 0 of the last three summands, and thus the result is reached in virtue of Slutsky's Theorem. $\square$

The "wild" bootstrap approach proposed can be applied by means of the following algorithm.

Algorithm 2 (Wild Bootstrap).

 Step 1. Compute the value of the statistic T_n (or the value T_n/σ_n).

 Step 2. Draw $\{\varepsilon_i^*\}_{i=1}^{n}$ a sequence of i.i.d. random elements ε, and compute $a_n = \|T_n^{W*}\|$ (or $b_n = \|T_n^{W*}\|/\sigma_n^*$, in this case σ_n^* is computed like in Step 2 of the Naive Bootstrap algorithm).

 Step 3. Repeat Step 2 a large number of times $B \in \mathbb{N}$ in order to obtain a sequence of values $\{a_n^l\}_{l=1}^{B}$ (or $\{b_n^l\}_{l=1}^{B}$).

 Step 4. Approximate the p–value of the test by the proportion of values in $\{a_n^l\}_{l=1}^{B}$ greater than or equal to $\|T_n\|$ (or by the proportion of values in $\{b_n^l\}_{l=1}^{B}$ greater than or equal to $\|T_n\|/\sigma_n$).

3   Bootstrap calibration vs. asymptotic theory

For simplicity, suppose from now on that $b = 0$ and X of zero–mean in (1), that is, suppose that the regression model is given by

$$Y = \langle \Theta, X \rangle + \varepsilon.$$

Furthermore, $\Delta(h) = E(\langle X, h \rangle Y)$ and, analogously, $\Gamma(h) = E(\langle X, h \rangle X)$. In such case, if we assume that $\sum_{j=1}^{\infty}(\Delta(v_j)/\lambda_j)^2 < +\infty$ and $Ker(\Gamma) = \{0\}$, then

$$\Theta = \sum_{j=1}^{\infty}\frac{\Delta(v_j)}{\lambda_j}v_j,$$

being $\{(\lambda_j, v_j)\}_{j\in\mathbb{N}}$ the eigenvalues and eigenfunctions of Γ (see Cardot, Ferraty, and Sarda (2003)).

A natural estimator for Θ is the FPCA estimator based on k_n functional principal components given by

$$\hat{\Theta}_{k_n} = \sum_{j=1}^{k_n}\frac{\Delta_n(\hat{v}_j)}{\hat{\lambda}_j}\hat{v}_j,$$

where Δ_n is the empirical estimation of Δ, that is, $\Delta_n(h) = (1/n)\sum_{i=1}^{n}\langle X_i, h \rangle Y_i$, and $\{(\hat{\lambda}_j, \hat{v}_j)\}_{j\in\mathbb{N}}$ are the eigenvalues and the eigenfunctions of Γ_n, the empirical estimator of Γ: $\Gamma_n(h) = (1/n)\sum_{i=1}^{n}\langle X_i, h \rangle X_i$.

Different statistics can be used for testing the lack of dependence between X and Y. Bearing in mind the expression (5), one can think about using an estimator of $\|\Theta\|^2 = \sum_{j=1}^{\infty}(\Delta(v_j)/\lambda_j)^2$ in order to test these hypotheses. In an alternative way, the expression (6) can be a motivation for different class of statistics based on the estimation of $\|\Delta\|'$.

One asymptotic distribution free based on the latter approach was given by Cardot, Ferraty, Mas, and Sarda (2003). They proposed as test statistic

$$T_{1,n} = k_n^{-1/2} \left(\hat{\sigma}^{-2} \| \sqrt{n} \Delta_n \hat{A}_n \|^2 - k_n \right),$$ (8)

where $\hat{A}_n(\cdot) = \sum_{j=1}^{k_n} \hat{\lambda}_j^{-1/2} \langle \cdot, \hat{v}_j \rangle \hat{v}_j$ and $\hat{\sigma}^2$ is a consistent estimator of σ^2. Cardot, Ferraty, Mas, and Sarda (2003) showed that, under H_0, $T_{1,n}$ converges in distribution to a centered Gaussian variable with variance equal to 2. Hence, H_0 is rejected if $|T_{1,n}| > \sqrt{2} z_\alpha$ (z_α the α–quantile of a $\mathcal{N}(0,1)$), and accepted otherwise. Besides, Cardot, Ferraty, Mas, and Sarda (2003) also proposed another calibration of the statistic distribution based on a permutation mechanism.

On the other hand, taking into account that $\|\Theta\|^2 = \sum_{j=1}^{\infty} (\Delta(v_j)/\lambda_j)^2$, one can use the statistic

$$T_{2,n} = \sum_{j=1}^{k_n} \left(\frac{\Delta_n(\hat{v}_j)}{\hat{\lambda}_j} \right)^2,$$ (9)

which limit distribution is not known.

Finally, a natural competitive statistic is the one proposed throughout Section 2.3

$$T_{3,n} = \left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \right\|,$$ (10)

which we will denote by "F–test" from now on since it is the natural generalization of the well–known F–test in the finite–dimensional context. Another possibility is to consider the studentized version of (10)

$$T_{3s,n} = \frac{1}{\hat{\sigma}} \left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \right\|,$$ (11)

where $\hat{\sigma}^2$ is the empirical estimation of σ^2.

In general, for the statistics such as (8), (9), (10) and (11), the calibration of the distribution can be obtained by using bootstrap. Furthermore, in the previous section, "naive" and "wild" bootstrap were shown to be consistent for the F–test, that is, the distribution of $T_{3,n}$ and $T_{3s,n}$ can be approximated by their corresponding bootstrap distribution, and H_0 can be rejected when the statistic value does not belong to the interval defined for the bootstrap acceptation region of confidence $1 - \alpha$. The same calibration bootstrap can be applied to the tests based on $T_{1,n}$ and $T_{2,n}$, although the consistence of the bootstrap procedure in this cases have not been proved in this work.

4 Simulation and real data applications

In this section a simulation study and an application to a real dataset illustrate the performance of the asymptotic approach and the bootstrap calibration from a practical point of view.

4.1 Simulation study

We have simulated $ns = 500$ samples, each being composed of $n \in \{50, 100\}$ observations from the functional linear model $Y = \langle \Theta, X \rangle + \varepsilon$, being X a Brownian motion and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with signal–to–noise ratio $r = \sigma / \sqrt{\mathbb{E}(\langle X, \Theta \rangle^2)} \in \{0.5, 1, 2\}$.

Under H_0, we have considered the model parameter $\Theta_0(t) = 0$, $t \in [0,1]$, whereas under H_1, the selected model parameter was $\Theta_1(t) = \sin(2\pi t^3)^3$, $t \in [0,1]$. Furthermore, under H_0 we have chosen $\sigma = 1$, while in the alternative H_1 we assigned the three different values that were commented before. Let us remark that both X and Θ were discretized to 100 equidistant design points.

We have selected the statistical tests which were introduced in the previous section: (8), (9), (10) and (11). For (8), three distribution approximations were considered: the asymptotic approach ($\mathcal{N}(0,2)$) and the following two bootstrap calibrations

$$T_{1,n}^{*(a)} = \frac{1}{\sqrt{k_n}}\left(\frac{n}{\hat{\sigma}^2}\sum_{j=1}^{k_n}\frac{(\Delta_n^*(\hat{v}_j))^2}{\hat{\lambda}_j} - k_n\right),$$

$$T_{1,n}^{*(b)} = \frac{1}{\sqrt{k_n}}\left(\frac{n}{(\hat{\sigma}^*)^2}\sum_{j=1}^{k_n}\frac{(\Delta_n^*(\hat{v}_j))^2}{\hat{\lambda}_j} - k_n\right).$$

The difference between the two proposed bootstrap approximations is that in the latter the estimation of σ^2 is also bootstrapped in each iteration. On the other hand, for (9), (10) and (11), only the bootstrap approaches were computed

$$T_{2,n}^* = \sum_{j=1}^{k_n}\left(\frac{\Delta_n^*(\hat{v}_j)}{\hat{\lambda}_j}\right)^2,$$

$$T_{3,n}^* = \left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\varepsilon_i^*\right\|,$$

$$T_{3s,n}^* = \frac{1}{\hat{\sigma}^*}\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\varepsilon_i^*\right\|.$$

For this simulation study, we have used the "wild" bootstrap algorithm introduced in Section 2.4 for the F–test and its studentized version, and the following adaptation of this consistent "wild" bootstrap for $T_{1,n}$ and $T_{2,n}$.

Algorithm 3 (Wild Bootstrap)**.**

Step 1. Compute the value of the statistic $T_{1,n}$ (or the value $T_{2,n}$).

Step 2. Draw $\{\varepsilon_i^*\}_{i=1}^n$ a sequence of i.i.d. random elements ε, and define $Y_i^* = Y_i\varepsilon_i^*$ for all $i = 1,\ldots,n$.

Step 3. Build $\Delta_n^*(\cdot) = n^{-1}\sum_{i=1}^n \langle X_i, \cdot\rangle Y_i^*$ for all $i = 1,\ldots,n$, and compute $a_n = |T_{1,n}^*|$ (or $b_n = |T_{2,n}^*|$).

Step 4. Repeat Steps 2 and 3 a large number of times $B \in \mathbb{N}$ in order to obtain a sequence of values $\{a_n^l\}_{l=1}^B$ (or $\{b_n^l\}_{l=1}^B$).

Step 5. Approximate the p–value of the test by the proportion of values in $\{a_n^l\}_{l=1}^B$ greater than or equal to $|T_{1,n}|$ (or by the proportion of values in $\{b_n^l\}_{l=1}^B$ greater than or equal to $|T_{2,n}|$).

Let us indicate that $1,000$ bootstrap iterations were done in each simulation.

Due to k_n and α must be fixed to run the procedure, the study was repeated with different numbers of principal components involved ($k_n \in \{1,\ldots,20\}$) and confidence levels ($\alpha \in \{0.2, 0.1, 0.05, 0.01\}$).

Nevertheless, in order to simplify the reading, the information collected in the following tables corresponds to only three of the values of k_n which were analyzed: $k_n = 5$, $k_n = 10$ and $k_n = 20$.

n	α	$\mathcal{N}(0,2)$ k_n			$T_{1,n}^{*(a)}$ k_n			$T_{1,n}^{*(b)}$ k_n			$T_{2,n}^{*}$ k_n			$T_{3,n}^{*}$	$T_{3s,n}^{*}$
		5	10	20	5	10	20	5	10	20	5	10	20		
	20%	19.4	17.6	16.0	21.4	21.6	20.0	21.6	19.0	15.2	19.8	20.8	18.4	21.6	20.8
50	10%	10.8	10.4	8.2	9.0	10.8	10.6	8.0	7.2	3.2	8.6	7.2	7.2	11.8	11.2
	5%	8.2	7.0	4.4	5.0	4.0	4.6	5.0	2.4	0.0	4.0	3.2	3.0	6.0	6.2
	1%	4.8	4.2	2.2	1.2	0.4	0.0	0.6	0.0	0.0	0.2	0.6	0.4	0.6	1.2
	20%	15.0	19.4	20.0	20.8	21.0	19.0	21.0	20.8	18.0	21.4	19.4	17.6	21.6	21.2
100	10%	8.6	9.6	9.0	11.8	10.8	10.4	10.4	9.6	6.2	9.8	8.8	7.0	11.6	11.8
	5%	5.6	5.2	4.0	4.4	4.6	3.6	3.6	3.4	2.2	4.6	5.2	2.8	5.6	5.6
	1%	2.6	2.4	1.2	1.4	1.2	0.8	1.2	0.6	0.2	1.0	0.6	0.8	0.4	0.4

Table 1: Comparison of the estimated levels for $T_{1,n}$ (using the asymptotic distribution $\mathcal{N}(0,2)$ and the bootstrap distributions of $T_{1,n}^{*(a)}$ and $T_{1,n}^{*(b)}$), $T_{2,n}$ (using the bootstrap distribution of $T_{2,n}^{*}$), $T_{3,n}$ (using the bootstrap distribution of $T_{3,n}^{*}$) and its studentized version, $T_{3s,n}$ (using the bootstrap distribution of $T_{3s,n}^{*}$).

n	α	$\mathcal{N}(0,2)$ k_n			$T_{1,n}^{*(a)}$ k_n			$T_{1,n}^{*(b)}$ k_n			$T_{2,n}^{*}$ k_n			$T_{3,n}^{*}$	$T_{3s,n}^{*}$
		5	10	20	5	10	20	5	10	20	5	10	20		
	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.8	0.0	0.0	100.0	100.0
50	10%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	60.8	0.0	0.0	100.0	100.0
	5%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0	32.2	0.0	0.0	100.0	100.0
	1%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.4	51.4	3.4	0.0	0.0	99.4	100.0
	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1.0	0.0	100.0	100.0
100	10%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.0	100.0	100.0
	5%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.4	0.0	0.0	100.0	100.0
	1%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0	0.0	0.0	100.0	100.0

Table 2: For $r = 0.5$, comparison of the empirical power for $T_{1,n}$ (using the asymptotic distribution $\mathcal{N}(0,2)$ and the bootstrap distributions of $T_{1,n}^{*(a)}$ and $T_{1,n}^{*(b)}$), $T_{2,n}$ (using the bootstrap distribution of $T_{2,n}^{*}$), $T_{3,n}$ (using the bootstrap distribution of $T_{3,n}^{*}$) and its studentized version, $T_{3s,n}$ (using the bootstrap distribution of $T_{3s,n}^{*}$).

Table 1 on page 13 displays the sizes of the test statistics obtained in the simulation study. For $T_{1,n}$, it can be highlighted that bootstrap approaches have closer sizes to the theoretical α than the asymptotic approximation for $T_{1,n}$, mainly when k_n is small. If we compare the performance of the two bootstrap procedures proposed, it seems that if σ^2 is bootstrapped ($T_{1,n}^{*(b)}$) the results are worse than if the same estimation of the variance is considered in all the bootstrap replications ($T_{1,n}^{*(a)}$) above all when k_n is large. As far as $T_{2,n}$ is concerned, the estimated levels are quite near to the nominal ones, being $k_n = 20$ the case in which they are farther from the theoretical α. Finally, it must be remarked that the F–test and its studentized versions also get good results in terms of test levels, which are slightly closer to α when one uses the bootstrap distribution of $T_{3s,n}^{*}$ to approximate the distribution of the statistic.

On the other hand, Table 2 on page 13, Table 3 on page 14, and Table 4 on page 14 show the empirical power obtained with the different procedures for each considered signal–to–noise ratio r.

In terms of power, when $r = 0.5$ the results for all the methods are similar, except for $T_{2,n}$ for which the empirical power decreases drastically, above all when k_n increases (this effect is also observed for $r = 1$ and $r = 2$). This fact seems to be due to the construction of $T_{2,n}$ since this test statistic is the only one which does not involve the estimation of σ^2. In addition, the power of $T_{1,n}$ also falls abruptly when $T_{1,n}^{*(b)}$ is considered, n is small and k_n is very large.

A similar situation can be observed when $r = 1$ and $r = 2$. In the latter it can be seen that the empirical power is smaller for all the methods in general, being obtained an important loss of power when the sample is small ($n = 50$), and k_n increases and/or α decreases (see Table 4 on page 14). Furthermore, in this case, it can be seen that the empirical power relies heavily on the selected k_n value. Hence, the advantage of using $T_{3,n}$ or $T_{3s,n}$ is that they do not require the selection of any parameter and they are competitive in terms of power. Nevertheless, it also seems that an adequate k_n selection can make $T_{1,n}$ obtain larger empirical power than $T_{3,n}$ or $T_{3s,n}$ in some cases.

n	α	$\mathcal{N}(0,2)$ k_n			$T_{1,n}^{*(a)}$ k_n			$T_{1,n}^{*(b)}$ k_n			$T_{2,n}^{*}$ k_n			$T_{3,n}^{*}$	$T_{3s,n}^{*}$
		5	10	20	5	10	20	5	10	20	5	10	20		
	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.2	66.6	3.6	0.2	100.0	100.0
50	10%	100.0	100.0	100.0	100.0	100.0	99.8	100.0	99.8	89.6	33.6	0.8	0.0	100.0	100.0
	5%	100.0	100.0	99.8	100.0	100.0	99.6	100.0	99.0	59.6	16.6	0.2	0.0	99.2	99.2
	1%	100.0	100.0	99.6	99.6	97.6	94.6	95.2	67.6	2.6	2.2	0.0	0.0	87.8	92.4
	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.0	7.8	0.0	100.0	100.0
100	10%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	86.4	2.2	0.0	100.0	100.0
	5%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	67.8	1.0	0.0	100.0	100.0
	1%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8	21.6	0.2	0.0	100.0	100.0

Table 3: For $r = 1$, comparison of the empirical power for $T_{1,n}$ (using the asymptotic distribution $\mathcal{N}(0,2)$ and the bootstrap distributions of $T_{1,n}^{*(a)}$ and $T_{1,n}^{*(b)}$), $T_{2,n}$ (using the bootstrap distribution of $T_{2,n}^{*}$), $T_{3,n}$ (using the bootstrap distribution of $T_{3,n}^{*}$) and its studentized version, $T_{3s,n}$ (using the bootstrap distribution of $T_{3s,n}^{*}$).

n	α	$\mathcal{N}(0,2)$ k_n			$T_{1,n}^{*(a)}$ k_n			$T_{1,n}^{*(b)}$ k_n			$T_{2,n}^{*}$ k_n			$T_{3,n}^{*}$	$T_{3s,n}^{*}$
		5	10	20	5	10	20	5	10	20	5	10	20		
	20%	85.4	75.6	66.8	89.0	81.2	77.2	89.0	76.8	51.4	34.0	11.8	7.2	90.4	89.8
50	10%	80.0	68.6	56.4	79.4	68.6	59.4	76.4	57.4	20.2	16.6	4.0	2.4	79.0	79.0
	5%	74.4	62.2	48.4	67.4	51.6	43.6	60.8	37.8	6.2	10.4	1.0	0.4	67.8	67.2
	1%	67.4	51.4	35.6	40.0	26.4	20.2	25.4	6.0	0.0	0.8	0.0	0.0	34.4	39.0
	20%	99.8	98.8	94.6	100.0	99.8	98.0	100.0	99.2	94.2	60.0	14.6	7.6	99.8	99.8
100	10%	99.6	96.6	91.2	99.6	97.2	93.6	99.6	96.0	82.4	34.2	6.2	2.0	97.8	97.4
	5%	99.6	95.6	85.8	97.8	94.0	85.8	97.2	90.4	64.6	18.0	2.8	0.4	94.4	94.4
	1%	97.6	91.4	75.4	88.2	76.4	64.0	85.2	63.4	26.2	2.2	0.8	0.0	79.2	82.4

Table 4: For $r = 2$, comparison of the empirical power for $T_{1,n}$ (using the asymptotic distribution $\mathcal{N}(0,2)$ and the bootstrap distributions of $T_{1,n}^{*(a)}$ and $T_{1,n}^{*(b)}$), $T_{2,n}$ (using the bootstrap distribution of $T_{2,n}^{*}$), $T_{3,n}$ (using the bootstrap distribution of $T_{3,n}^{*}$) and its studentized version, $T_{3s,n}$ (using the bootstrap distribution of $T_{3s,n}^{*}$).

4.2 Data application

For the real data application, we have obtained concentrations of hourly averaged NO_x in the neighborhood of a power station belonging to ENDESA, located in As Pontes in the Northwest of Spain. During unfavorable meteorological conditions, NO_x levels can quickly rise and cause an air–quality episode. The aim is to forecast NO_x with half an hour horizon to allow the power plant staff to avoid NO_x concentrations reaching the limit values fixed by the current environmental legislation. This fact implies that it is necessary to estimate properly the regression model which defines the relationship between the observed NO_x concentration in the last minutes (X) and the NO_x concentration with half an hour horizon (Y). For that, a first step is to determine if there exists a linear dependence between X and Y.

Therefore, we have built a sample where each curve X corresponds to 240 consecutive minutal values of hourly averaged NO_x concentration, and the response Y corresponds to the NO_x value half an hour ahead (from Jan 2007 to Dec 2009). Applying the tests for dependence to the dataset, the null hypothesis is rejected in all cases (thus, there is a linear relationship between the variables), except for $T_{2,n}$ when k_n is large (see Table 5 on page 15). Nevertheless, as we have commented in the simulation study, this test statistic does not take into account the variance term and its power is clearly lower than the power of the other tests.

$\mathcal{N}(0,2)$			$T_{1,n}^{*(a)}$			$T_{1,n}^{*(b)}$			$T_{2,n}^{*}$			$T_{3,n}^{*}$	$T_{3s,n}^{*}$
k_n			k_n			k_n			k_n				
5	10	20	5	10	20	5	10	20	5	10	20		
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.000	0.000

Table 5: Real data application. P–values for $T_{1,n}$ (using the asymptotic distribution $\mathcal{N}(0,2)$ and the bootstrap distributions of $T_{1,n}^{*(a)}$ and $T_{1,n}^{*(b)}$), $T_{2,n}$ (using the bootstrap distribution of $T_{2,n}^{*}$), $T_{3,n}$ (using the bootstrap distribution of $T_{3,n}^{*}$) and its studentized version, $T_{3s,n}$ (using the bootstrap distribution of $T_{3s,n}^{*}$).

5 Final comments

The proposed bootstrap methods seems to give test sizes closer to the nominal ones than the tests based on the asymptotic distributions. In terms of power, the statistic tests which include a consistent estimation of the error variance σ^2 are better that the tests which do not take it into account. Furthermore, in all the cases, a suitable choice of k_n seems to be quite important, and currently it is still an open question.

Besides of the optimal k_n selection, other issues related to these dependence tests require further research, such as their extension to functional linear models with functional response. On the other hand, and in addition to the natural usefulness of this test, if would be interesting to combine it with the functional ANOVA test (see Cuevas, Febrero, and Fraiman (2004), and González-Rodríguez, Colubi, and Gil (2012)) in order to develop an ANCOVA test in this context.

Acknowledgements

References

Bathia, N., Yao, Q. and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *Ann. Statist.* **38**, 3352–3386.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–1217.

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

Cao-Abad, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Statist.* **19**, 2226–2231.

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

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional Linear Model. *Statist. Probab. Lett.* **45**, 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13**, 571–591.

Cardot, H., Prchal, L. and Sarda, P. (2007). No effect and lack–of–fit permutation tests for functional regression. *Comput. Statist.* **22**, 371–390.

Cuevas, A., Febrero, M. and Fraiman, R. (2004). An anova test for functional data. *Comput. Statist. Data Anal.* **47**, 111–122.

Cuevas, A., Febrero, M. and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Comput. Statist. Data Anal.* **51**, 1063–1074.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Ferraty, F. and Romain, Y. (eds.) (2011). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press, Oxford.

Ferraty, F., Van Keilegom, I. and Vieu, P. (2010). On the validity of the bootstrap in non–parametric functional regression. *Scand. J. Stat.* **37**, 286–306.

Ferraty, F., Van Keilegom, I. and Vieu, P. (2012). Regression when both response and predictor are functions. *J. Multivariate Anal.* **109**, 10–28.

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

Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9**, 1218–1228.

Giné, E. (1997). Lectures on some aspects of the bootstrap. In *Lectures on Probability Theory and Statistics (Saint–Flour, 1996)* (Edited by B. Pierre), 37–151. Springer, Berlin.

Giné, E. and Zinn, J. (1990). Bootstrapping General Empirical Measures. *Ann. Probab.* **18**, 851–869.

González-Manteiga, W. and Martínez-Calvo, A. (2011). Bootstrap in functional linear regression. *J. Statist. Plann. Inference* **141**, 453–461.

González-Rodríguez, G., Colubi, A. and Gil, M. Á. (2012). Fuzzy data treated as functional data: A one–way ANOVA test approach. *Comput. Statist. Data Anal.* **56**, 943–955.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 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**, 109–126.

Hall, P. and Vial, C. (2006). Assessing the finite dimensionality of functional data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 689–705.

Kokoszka, P., Maslova, I., Sojka, J. and Zhu, L. (2008). Testing for lack of dependence in the functional linear model. *Canad. J. Statist.* **36**, 207–222.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer, New York.

Laha, R. G. and Rohatgi, V. K. (1979). *Probability Theory.* Wiley, New York.

Ledoux, M. and Talagrand, M. (1988). Un critère sur les petites boules dans le théorème limite central. *Probab. Theory Related Fields* **77**, 29–47.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis. Methods and Case Studies.* Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis.* 2nd edition. Springer, New York.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187–1195.

Reports in Statistics and Operations Research

2005

05-01 SiZer Map for Evaluating a Bootstrap Local Bandwidth Selector in Nonparametric Additive Models. M. D. Martínez-Miranda, R. Raya-Miranda, W. González-Manteiga and A. González-Carmona.

05-02 The Role of Commitment in Repeated Games. I. García Jurado, Julio González Díaz.

05-03 Project Games. A. Estévez Fernández, P. Borm, H. Hamers.

05-04 Semiparametric Inference in Generalized Mixed Effects Models. M. J. Lombardía, S. Sperlich.

2006

06-01 A unifying model for contests: effort-prize games. J. González Díaz.

06-02 The Harsanyi paradox and the "right to talk" in bargaining among coalitions. J. J. Vidal Puga.

06-03 A functional analysis of NOx levels: location and scale estimation and outlier detection. M. Febrero, P. Galeano, W. González-Manteiga.

06-04 Comparing spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

06-05 On the spectral simulation of spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal.

06-06 An L_2-test for comparing spatial spectral densities. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

2007

07-01 Goodness-of-fit tests for the spatial spectral density. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

07-02 Presmothed estimation with left truncated and right censores data. M. A. Jácome, M. C. Iglesias-Pérez.

07-03 Robust nonparametric estimation with missing data. G. Boente, W. González-Manteiga, A. Pérez-González.

07-04 k-Sample test based on the common area of kernel density estimators. P. Martínez-Camblor, J. de Uña Álvarez, N. Corral-Blanco.

07-05 A bootstrap based model checking for selection-biased data. J. L. Ojeda, W. González-Manteiga, J . A. Cristobal.

07-06 The Gaussian mixture dynamic conditional correlation model: Bayesian estimation, value at risk calculation and portfolio selection. P. Galeano, M. C. Ausín.

2008

08-01 ROC curves in nonparametric location-scale regression models. W. González-Manteiga, J. C. Pardo Fernández, I. Van Keilegom.

08-02 On the estimation of α-convex sets. B. Pateiro-López, A. Rodríguez-Casal.

2009

09-01 Lasso Logistic Regression, GSoft and the Cyclyc Coordinate Descent Algorithm. Application to Gene Expression Data. M. García-Magariños, A. Antoniadis, R. Cao, W. González-Manteiga.

2010

10-01 Asymptotic behaviour of robust estimators in partially linear models with missing responses: The effect of estimating the missing probability on simplified marginal estimators. A. Bianco, G. Boente, W. González-Manteiga, A. Pérez-González.

10-02 First-Price Winner-Takes-All Contents. J. González-Díaz.

10-03 Goodness of Fit Test for Interest Rate Models: an approach based on Empirical Process. A. E. Monsalve-Cobis, W. González-Manteiga, M. Febrero-Bande.

2011

11-01 Exploring wind direction and SO_2 concentration by circular–linear density estimation. E. García–Portugués, R.M. Crujeiras, W. González–Manteiga.

11-02 Utilities for Statistical Computing in Functional Data Analysis: The R Package fda.usc. M. Oviedo de la Fuente, M. Febrero-Bande.

11-03 Multivariate uniformity tests: the distance to boundary method for the case of unknown support. J. R. Berrendero, A. Cuevas, B. Pateiro-López.

11-04 On the meaning, properties and computing of a coalitional Shapley value. J. M. Alonso-Meijide, B. V. Casas-Méndez, A. M. González-Rueda, S. M. Lorenzo-Freire.

11-05 Share functions for cooperative games with levels structure of cooperation. M. Álvarez-Mozos, R. van den Brink, G. van der Laan, O. Tejada.

2012

12-01 A goodness of fit test for the functional linear model with scalar response. E. García–Portugués, W. González–Manteiga, M. Febrero-Bande.

12-02 Bootstrap independence test for functional linear models. W. González-Manteiga, G. González-Rodríguez, A. Martínez-Calvo, E. García-Portugués

Previous issues (2001 – 2003):
http://eio.usc.es/reports.php