Multiway ANOVA for Functional Data

J.A. Cuesta-Albertos* and M. Febrero-Bande†
Departamento de Matemáticas, Estadística y Computación,
Universidad de Cantabria, Spain
and
Departamento de Estatística e I.O.
Universidad de Santiago de Compostela, Spain

May 16, 2008

Abstract

In this paper a procedure to handle complicated ANOVA designs for functional data is proposed. The procedure is based on the analysis of randomly chosen one-dimensional projections. The paper contains some theoretical results as well as some simulations and the analysis of some real data sets. Functional data include multidimensional data, thus the paper includes a comparison between the propoesed procedure and some usual MANOVA tests.

Key words and phrases: Two-way ANOVA, ANCOVA, Functional Data, Random Projections.

1 Introduction

The popularization of the computers has allowed that many processes are continuously monitored, thus providing data which are functions, the so-called functional data. Those include the changes in value of a particular share or index in the stocks market, the temperatures in a given localization, the levels of a pollutant in the atmosphere,... but it is also possible to include in this setting data which are not functions, but that are highly multidimensional like those handled in genomics. Classical references on functional data are [18] and [19].

A natural statistical problem is to decide on the existence or not of differences in the process of interest when some conditions which may affect it vary. We refer to, for example, to the existence of differences in the behavior of the temperatures between different localizations, between the values of the stocks in American, English, German and Japanese markets,...

This problem is usually handled by employing a model which assumes the existence of an underlying function describing the typical evolution of the process under consideration,

*This author has been partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant MTM2005-08519-C02-02 and the Consejería de Educación y Cultura de la Junta de Castilla y León, grant PAPIJCL VA102/06.

†This author has been partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant MTM2005-00820

assuming that the data we have at hand have been obtained by adding random fluctuations to this typical function. Thus, the problem becomes a kind of functional ANOVA problem. A basic reference on the functional ANOVA is [18].

Let us state the problem more precisely. Let us assume that we have a real-valued random process $\mathbf{X} = \{\mathbf{X}(t), t \in [a,b]\}$, measured along the closed and bounded interval $[a,b]$. Assume that there are R different conditions which may affect the process and that it have been measured n_r times under each of those conditions, giving the sample $\mathbf{X}_i^r(t), i = 1, ..., n_r,\ r = 1, ..., R$. Assume also that for every $r = 1, ..., R$, there exists a (non random) function $f^r : [a,b] \to I\!R$ such that

$$\mathbf{X}_i^r(t) = f^r(t) + \epsilon_i^r(t),\ t \in [a,b], i = 1, ..., n_r, \tag{1}$$

where ϵ_i^r are independent and identically distributed random functions centered in mean (thus, $E[\epsilon_i^r(t)] = 0$ for every i, r, t).

In [18], as well as in many other papers on the subject, it is taken advantage of the fact that the measurements are usually only made in a finite set of values $\{t_1, ..., t_m\} \subset [a,b]$. The idea is to apply a real ANOVA analysis for every value t_j. Thus, we would obtain m values for the F-test, one for each t_j. If all F-values are below the rejection value, we can, safely, say that there is no evidence of differences between the functions f^r. The problems begin when we obtain some F-values inside the rejection area because the usual procedure to handle multiple comparisons (Bonferroni's inequality) is quite conservative.

Moreover, even in the case that $R = 1$, and that the distribution of $(\epsilon_i^1(t_1), ..., \epsilon_i^1(t_m))$ is standard Gaussian, it is shown in [12] that if we want to test the null hypothesis $f^1 = 0$, then the powers of the likelihood ratio test converge to the level of the test if $m \to \infty$, and $\| (f^1(t_1), ..., f^1(t_m)) \| \to \infty$, with $\| (f^1(t_1), ..., f^1(t_m)) \| = o(m^{1/2})$. The asymptotic distribution of the likelihood ratio test for the MANOVA when $m/(n-R) \to \gamma \in (0, \infty)$ was obtained in [25] (here we denote $n = n_1 + \ldots + n_R$).

To solve those problems, a kind of bootstrap procedure is proposed in [18]. Alternatively, in [13] a criteria to select a finite subset of t_j's is proposed. Both procedures are employed, for instance, in [24].

Other approaches are based on the application of a dimension-reduction and/or smoothing technique (see, for instance, [1], [2], [3], [4], [14], [16]). Different points of view are handled in [10], [15] and [21]. In [10] an F-test based on the squares of the norms of the between samples and within samples differences is proposed. In [15] and [21] the authors analyze the behavior of the traces of the between and within covariance matrices.

A common drawback of previous solutions is that they are intended to solve only one-way problems. Thus, in spite that the design handled in [24] is three-ways, no interaction between factors is analyzed. The same happens in [2] or [4] where no interaction between the fixed and random factors is considered.

In this paper we propose a simple technique which allows to test complicated designs, including interactions, covariables,...

This technique is based on Theorem 4.1 in [9] which, roughly speaking, states that if we have two different distributions, one of them verifying a condition on their moments,

and we randomly choose a marginal of them, those marginals are almost surely different. We state here this result for further reference.

Theorem 1.1 *Let $\mathcal{H}$ be a separable Hilbert space with norm $\|-\|$. Let μ be a Gaussian distribution on $\mathcal{H}$ such that that each of its one-dimensional projections is non-degenerate.*

Let P and Q be two probability distributions on $\mathcal{H}$ such that the absolute moments $m_k := \int \|x\|^k dP(x)$ are finite and satisfy $\sum_{k\geq 1} m_k^{-1/k} = \infty$. If $P \neq Q$, then

$$\mu\left\{v \in \mathcal{H} : \ P \circ \Pi_v^{-1} = Q \circ \Pi_v^{-1}\right\} = 0,$$

where Π_v is the orthogonal projection on the one-dimensional subspace generated by v.

The assumption $\sum_{k\geq 1} m_k^{-1/k} = \infty$ is known as Carleman's condition and is a sufficient condition to that the distribution P to be determined by its moments (see [23]). It is satisfied, for instance, for all Gaussian or with bounded support distributions

Thus, according to Theorem 1.1, if we want to test the null hypothesis

$$H_0 : f^1 = ... = f^R$$

in (1), we only need to randomly take a one-dimensional projection and analyze the associated one-dimensional projections of the random functions in the sample. This analysis can be carried out using a one-dimensional ANOVA to test if the means of those projections are the same.

We have presented this reasoning in the one-way ANOVA just to simplify the exposition, but, obviously, the procedure can be extended to cover more complicated linear models.

The test that we propose here is conditional on the obtained one-dimensional projection which has to be chosen independently of the sample at hand. Conditional tests are not new in Statistics. In fact, most of bootstrap procedures are conditional (on the obtained sample), but we can trace conditional procedures, at least, to the exact Fisher test for contingency tables.

We also want to remark that since Theorem 1.1 holds if we take $\mathcal{H} = I\!R^d$, then the proposed procedure can be used as an alternative to the usual MANOVA.

The paper follows the following scheme: In Section 2 we develop the procedure. In Section 3 we present some simulations. In Section 4 we analyze a real data set related to temperatures in different locations in Spain; this is a two-way ANOVA model plus a continuos covariable. Section 5 is devoted to compare the proposed procedure with some usual MANOVA tests. The paper ends with a Discussion where we comment on the possibility to apply the proposed procedure to more general abstract spaces.

2 The procedure

In this section $\mathcal{H}$ will denote a separable Hilbert space endowed with norm $\|-\|$ and scalar product $\langle\cdot,\cdot\rangle$. All the random variables will be assumed to be defined on the same,

rich enough, probability space. Moreover, without loss of generality we will assume that the closed interval in which we measure the process is $[0,1]$.

2.1 The model. Assumptions

We begin generalizing the model (1) to a two-ways ANOVA model with a covariable. Obviously, other models can be treated similarly.

Let $R, S \in I\!N$ and assume that for every $r = 1, ..., R,\ s = 1, ..., S$ there exist $\mathbf{X}_i^{r,s}, i = 1, ..., n_{r,s} \in I\!N$ random functions in $\mathcal{H}$, such that

$$\mathbf{X}_i^{r,s}(t) = m(t) + f^r(t) + g^s(t) + h^{r,s}(t) + \gamma_i^{r,s} Y(t) + \epsilon_i^{r,s}(t),\ t \in [0,1], \tag{2}$$

where

1. The function m is non random and describes the overall shape of the process.
2. The (non random) functions $f^r, g^s, h^{r,s}$ belong to $\mathcal{H}$ and account, respectively, for the main effect of the first and second factors and for the interaction between them.

   In order to make the model identifiable, we will assume, as usually, that, for every $t \in [0,1]$, $r_0 = 1, ..., R$ and $s_0 = 1, ..., S$

   $$\sum_r f^r(t) = \sum_s g^s(t) = \sum_r h^{r,s_0}(t) = \sum_s h^{r_0,s}(t) = 0.$$

3. The $\gamma_i^{r,s} \in I\!R$ are random and known quantities which influence the process according to the weights given by the (non-random and unknown) function $Y \in \mathcal{H}$. They play the same role as the covariables in the one-dimensional case.
4. The random trajectories $\epsilon_i^{r,s}$ are assumed to be $\mathcal{H}$-valued, independent and centered in mean. Moreover, for each r, s, $\epsilon_i^{r,s}, i = 1, ..., n_{r,s}$ are identically distributed and satisfy Carleman's condition, i.e. we assume that $\sum_k \left(E\|\epsilon_1^{r,s}\|^k\right)^{-1/k} = \infty$, for every r, s.

We are interested into testing the following null hypotheses

$$\begin{aligned} H_0^A &: \quad f^1 = \ldots = f^R = 0 \\ H_0^B &: \quad g^1 = \ldots = g^S = 0 \\ H_0^I &: \quad h^{1,1} = \ldots = h^{R,S} = 0 \\ H_0^C &: \quad Y = 0 \end{aligned}$$

against the alternatives that theirs negations hold.

2.1.1 Homoscedasticity and Gaussian hypotheses.

It is well known that the homoscedasticity plays an important role in ANOVA techniques. Here, it is equivalent to say that, for every i, r, s, the distribution of $\epsilon_i^{r,s}$ does not depend on the remaining terms in (2).

We prefer do not make this assumption because in random functions the opposite looks quite reasonable. We mean that, very often, the oscillation of a process strongly depends on their values.

Concerning the Gaussian hypothesis, from Theorem 3.6 in [7] follows that if the distribution of an $\mathcal{H}$-valued random element X is not Gaussian then, almost surely, every one-dimensional projection will neither be Gaussian.

The only problem without those assumptions is that we will need to apply an one-dimensional ANOVA test which works under heteroscedastic and/or non-Gaussian conditions.

Anyway, since the procedure that we propose is conditional on the chosen marginal, it is possible to check which particular hypotheses are satisfied by the obtained projections and, then, to choose the one-dimensional test more suitable for them.

2.2 The procedure

The procedure that we propose is directly based on the following result.

Theorem 2.1 *Let us assume the model and assumptions stated in Subsection 2.1. Let μ be a Gaussian distribution on $\mathcal{H}$ such that each of its one-dimensional projections is non-degenerate. Then*

1. *If there exist r_1, r_2 such that $f^{r_1} \neq f^{r_2}$ then*

$$\mu\left\{v \in \mathcal{H} : \text{ such that } \langle v, f^1\rangle = \ldots = \langle v, f^R\rangle\right\} = 0.$$

2. *If there exist s_1, s_2 such that $g^{s_1} \neq g^{s_2}$ then*

$$\mu\left\{v \in \mathcal{H} : \text{ such that } \langle v, g^1\rangle = \ldots = \langle v, g^S\rangle\right\} = 0.$$

3. *If there exist $(r_1, s_1), (r_2, s_2)$ such that $h^{r_1,s_1} \neq h^{r_2,s_2}$ then*

$$\mu\left\{v \in \mathcal{H} : \text{ such that } \langle v, h^{1,1}\rangle = \ldots = \langle v, h^{R,S}\rangle\right\} = 0.$$

4. *If $Y \neq 0$ then $\mu\{v \in \mathcal{H} : \text{ such that } \langle v, Y\rangle = 0\rangle\} = 0$.*

Proof. Given $r = 1, ..., R$, let P^r be the probability distribution on $\mathcal{H}$ concentrated on f^r (i.e., P^r satisfies that $P^r[f^r] = 1$). Obviously, every probability distribution concentrated on a single point, satisfies Carleman's condition. Thus, we can apply Theorem 1.1 to every pair of probability distributions P^{r_1} and P^{r_2}. Since there is a finite number of

possible selections, the result in item 1 follows. The proofs of the remaining statements are identical. •

To explain the procedure, let us focus our attention on the testing of hypothesis H_0^A. Let $v \in \mathcal{H}$ be a randomly chosen vector employing the distribution μ. Oviously, if H_0^A holds, then for every $v \in \mathcal{H}$, the (conditional on v) null hypothesis

$$H_0^{A,v}: \qquad \langle v, f^1 \rangle = \ldots = \langle v, f^R \rangle = 0$$

also holds. And, according to 1 in Theorem 2.1, if H_0^A fails, then for μ-almost every $v \in \mathcal{H}$, $H_0^{A,v}$ also fails. Thus, a statistical test at level α to test $H_0^{A,v}$ is a statistical test at the same level to test H_0^A.

Therefore, if Φ^A is an statistical test at level α to test the null hypothesis H_0^A, under the model (2) in the case that the data are real, it happens that it is possible to test H_0^A in the $\mathcal{H}$-valued case by applying Φ^A to test $H_0^{A,v}$.

Concerning the hypotheses to apply a one-dimensional ANOVA, notice that for every $r, s,$ the one-dimensional errors $\langle v, \epsilon_i^{r,s} \rangle, i = 1, ..., n_{r,s}$ are independent and identically distributed. The only care to be taken into account is that if the original observations are not homoscedastic, then, the test Φ^A should work under heteroscedastic conditions, and that if the original observations are not Gaussian, then the test Φ^A should be able to work under non-Gaussian conditions. Some statistical procedures appear in the literature which fulfills those requirements. Here, we have chosen this one proposed in [5].

The same can be said with respect the null hypotheses H_0^B, H_0^I and H_0^C and the corresponding (random) counterparts $H_0^{A,v}$, $H_0^{B,v}$ and $H_0^{C,v}$.

2.2.1 Stability of the procedure

According to the previous reasoning, the exposed procedure is consistent in the sense that to reject any null hypothesis $H^{\cdot,v}$ is almost surely equivalent to reject the corresponding null hypothesis $H_0^{\cdot}$. However, it is obvious that this procedure has two main drawbacks.

On the first side, we are loosing some information because we are replacing a function by just one real number and this should bring some lose of power. On the other hand, if we employ a randomly chosen projection, we are accepting some random instability in the problem because it may happen that if we repeat twice the procedure a null hypothesis to be rejected once and accepted later.

In order to ease this inconveniences, it has been proposed (for instance in [7] and [8]) to take $k > 1$ random projections and, then, adjust the obtained p-values by bootstrap. Bonferroni's method is the selection in [6]. Alternatively, in descriptive studies where the goal is to provide an idea about how far from acceptable is the null hypothesis under consideration, report, for instance, the mean value of the obtained p-values could be enough.

We have employed Bonferroni's correction in Sections 3 and 4 and both methods (Bonferroni and bootstrap) in Section 5.

The only problem with Bonferroni's correction is that we have to make our selection between two opposite goals: the larger k, the lower the instability and, also, the harder to reject the null hypothesis because the Bonferroni adjustment becomes more conservative. However, if the projections are independent, then the Bonferroni adjustment is more accurate. Because of this, it seems reasonable to us that the number of random directions should be of the same order as the number of points in which the random trajectories have been measured.

We have tried several values of k in the simulations reported in Section 3. At the end we have found that a too large number is not required being the optimum around 50 when the number of points of in which the process is measured is sufficiently large. If we denote this number by N, we propose to choose k between $\min(N, 50)$ and $3 \times \min(N, 50)$, with k closest to the lower limit if N is large and closest to the upper limit if N is small.

In Sections 3 and 4 the value of N is, respectively, 101 and 31 and we have chosen $k = 50$ and $k = 31$ respectively. In Section 5 the data are in dimension three. Thus, they can be identified with functional data measured just on $N = 3$ points and we have decided to go closer to the upper limit, selecting $k = 10$.

On the other hand, if the bootstrap is chosen to fix the p-values, there is no theoretical reason to select a low k. However, we have decided to follow the same rule as for the previous method and in Section 5, when using bootstrap, we have also selected $k = 10$.

3 Simulations

A Monte Carlo study has been carried out in order to evaluate the perfomance of the procedure proposed in Section 2. To this, we have simulated several examples following the model (2). We have taken $R = S = 2$ and $n_{r,s} = 100,\ r, s = 1, 2$ in all of them. The underlying distributions depend on two parameters α, β as follows:

- $m(t) := 30(1-t)t$
- $f^r(t) := \alpha(-1)^r |\sin(4\pi t)|, r \in \{1, 2\}, \alpha \in \{0.0, 0.05, \ldots, 0.45, 0.5\}$
- $g^s(t) := \beta(-1)^s I_{\{t>0.5\}}, s \in \{1, 2\}, \beta \in \{0.0, 0.05, \ldots, 0.45, 0.5\}$
- $h^{r,s}(t) := -f^r(t)g^s(t)I_{\{\alpha \geq 0.25\}}$
- $Y(t) := (t - 0.5)$ and the distribution of the real random variable γ is $U[-0.5, 0.5]$
- $\epsilon_i^{r,s}(t)$ is a Gaussian process with mean equals to zero and covariance function

$$Cov(\epsilon(t_1), \epsilon(t_2)) := \sigma^2 \exp(-|t_1 - t_2|/0.3), \sigma^2 = 0.4.$$

where $t \in [0, 1]$ was discretized in $N = 101$ points, i.e. we assume that the processes have been observed at the points $t_l = l/100, l = 0, \ldots, 100$.

The parameter α (resp. β) controls the strenght of the factor $f^r(t)$ (resp. $g^s(t)$). Indeed depending on α the interaction is present or not.

Every cell in the following tables is the result of $K = 500$ replications each with $k = 50$ random projections showing the proportion of rejections at level 0.05 after correcting the p-values with Bonferroni's method. i.e., since we have $k = 50$ one-dimensional tests, we only declare as rejected a null hypothesis if the minimum obtained p-value is less or equal than 0.001 (=0.05/50).

	β										
α	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.00	0.030	0.038	0.036	0.022	0.032	0.040	0.032	0.040	0.026	0.040	0.036
0.05	0.358	0.362	0.362	0.322	0.396	0.372	0.350	0.330	0.378	0.362	0.342
0.10	0.964	0.970	0.974	0.976	0.978	0.980	0.980	0.972	0.968	0.964	0.972
0.15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.35	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.40	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.45	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 1: Proportion of rejections for H_0^A

	β										
α	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.00	0.040	0.280	0.936	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.05	0.038	0.250	0.944	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.10	0.046	0.292	0.930	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.15	0.034	0.276	0.944	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.20	0.014	0.300	0.944	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.25	0.040	0.300	0.950	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.30	0.026	0.352	0.954	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.35	0.032	0.354	0.948	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.40	0.042	0.366	0.962	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.45	0.036	0.350	0.962	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.50	0.024	0.354	0.956	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2: Proportion of rejections for H_0^B

Concerning the one-dimensional test that we have employed in this section, we have used the usual ANOVA test based on the gaussianity and homoscedasticity assumptions (which we know, by construction, that are satisfied). Anyway, we have checked the assumption of gaussianity of the projections and, as expected under the normality hypothesis, this assumption was rejected around 5% of the cases. This test was done using the package `nortest` from R ([20]).

The results are quite satisfactory. The first row in Table 1 corresponds to the situation of no effect of the factor A and the behaviour is nice although is a quite conservative (probably by the use of Bonferroni's approximation). The same can be told about the first column in Table 2. As α is growing in Table 1 (resp. β in Table 2) the proportion of rejections quickly increases. Table 3 correspond to effect of the covariate and not surprisingly all the proportions are near one.

	β										
α	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.00	0.998	0.998	0.998	0.998	0.996	1.000	1.000	0.996	0.992	0.996	0.996
0.05	0.996	0.992	1.000	0.998	0.996	0.996	1.000	1.000	1.000	1.000	0.996
0.10	0.996	0.994	0.998	0.996	0.996	0.996	1.000	0.994	0.998	1.000	1.000
0.15	0.994	0.992	0.998	0.996	1.000	0.996	0.988	0.998	0.996	0.994	1.000
0.20	0.998	1.000	0.998	0.996	0.992	0.996	0.996	0.998	0.998	0.994	0.998
0.25	0.998	1.000	0.996	0.996	0.996	0.994	0.996	0.998	0.998	0.984	1.000
0.30	0.990	0.996	0.992	0.996	0.988	0.996	1.000	0.998	0.998	0.994	1.000
0.35	0.998	0.998	0.998	0.996	1.000	0.992	0.998	0.992	0.998	1.000	0.996
0.40	0.998	0.992	0.998	0.994	0.998	1.000	0.996	1.000	1.000	0.998	0.998
0.45	0.994	0.994	0.992	0.994	0.992	1.000	0.998	0.998	0.992	1.000	0.998
0.50	0.998	0.996	0.996	0.994	0.992	0.994	0.994	1.000	0.998	1.000	0.998

Table 3: Proportion of rejections for H_0^C

	β										
α	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.00	0.022	0.040	0.028	0.016	0.038	0.040	0.042	0.038	0.042	0.034	0.028
0.05	0.032	0.024	0.032	0.038	0.020	0.026	0.022	0.020	0.022	0.030	0.026
0.10	0.040	0.036	0.018	0.038	0.022	0.022	0.042	0.022	0.028	0.030	0.048
0.15	0.030	0.048	0.040	0.020	0.020	0.034	0.034	0.020	0.032	0.032	0.030
0.20	0.036	0.032	0.040	0.038	0.036	0.032	0.042	0.018	0.040	0.024	0.040
0.25	0.024	0.044	0.052	0.084	0.122	0.158	0.244	0.370	0.460	0.644	0.694
0.30	0.030	0.024	0.058	0.088	0.162	0.258	0.356	0.570	0.678	0.826	0.898
0.35	0.024	0.024	0.058	0.124	0.222	0.354	0.514	0.722	0.854	0.922	0.966
0.40	0.034	0.044	0.062	0.172	0.296	0.484	0.714	0.836	0.952	0.978	0.992
0.45	0.038	0.042	0.088	0.234	0.380	0.626	0.802	0.936	0.982	0.998	1.000
0.50	0.030	0.048	0.126	0.226	0.514	0.718	0.912	0.972	0.994	1.000	1.000

Table 4: Proportion of rejections for H_0^I

Table 4 have the information about interaction which is only present for $\alpha \geq 0.25$. As the amount of interaction depends on the product $\alpha \times \beta$, the proportion of rejections goes near one only when both α and β are large.

4 A real data set

We illustrate the proposed procedure by using a real dataset. The data we have at hand were obtained from the web page http://clima.meteored.com on September, 2007. This site contains a lot of climatological data from a large amount of meteorological stations. We consider the daily mean temperature as the functional datum for certain locations in Spain and certain months in the last annual cycle. More precisally, we have downloaded the daily mean temperature in A Coruña, Avilés, Bilbao, San Sebastián, Santander, Vigo, Burgos, León, Madrid, Salamanca, Segovia, Soria, Valladolid, Vitoria and Zamora. The last nine locations are in the interior of Spain and have Continental climatological characteristics (colder winters and hotter summers) whereas the first six locations are situated at the North Coast of Spain (Atlantic Ocean) and have temperatures less extreme. Also, in order to take into account the variability of seasons, we have selected

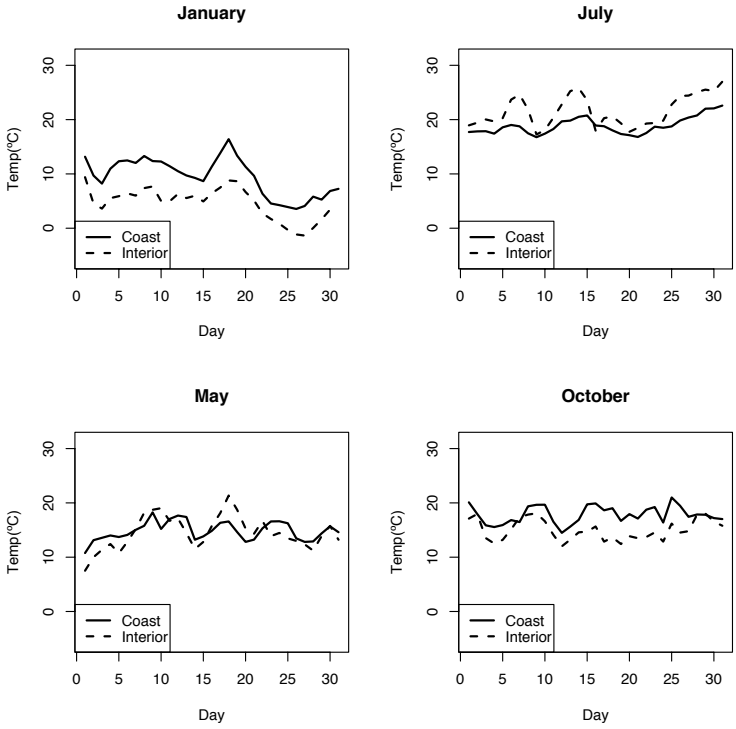

Figure 1: Mean Temperatures

the most representative available months in last twelve months: October-2006, January-2007, May-2007 and July-2007.

Indeed, we have also take into account the Monthly Total Amount of Rainfall as a covariate that can have influence in temperature. Thus, this is the real known random variable γ in (2) which multiplies the unknown and non random function Y measuring the influence of this quantity each day in the month. On the other hand, it seems us that in this particular example, Y should be constant, but this is not important in our model.

In Figure 1 we have represented the curves of the mean temperatures in each month and zone. It becomes obvious from those graphs the influence of the season factor. But, in our opinion, the differences between the corresponding curves are not so big as to make so obvious the influence of the location factor. The same happens with the existence of the interaction, in spite of the graphs suggest its existence since the mean temperature in the interior is lower than in the coastal area in winter and greater in summer.

The first row of Table 5 contains the p-values obtained for each null hypothesis with the proposed test using 31 random projections each time. However, taking into account the random nature of the test, we have repeated it 1000 times. The second row in Table 5 shows the proportions of rejections of the null hypothesis using also 31 random projections each time.

Notice that the p-values were corrected using the Bonferroni method. This means the following. For instance, in the first factor, the minimum of the obtained p-values was

	Location	Month	Interaction	Monthly Rainfall	Gaussian
p-value	$1.29 \cdot 10^{-05}$	$3.33 \cdot 10^{-37}$	0.00157	$3.53 \cdot 10^{-10}$	0.387
Proportion of rejection	1.000	1.000	1.000	0.804	0.114

Table 5: p-values obtained with the random ANOVA and Bonferroni's correction and proportions of null hypothesis rejected for climatological data

$4.169491 \cdot 10^{-07}$. Since, we are employing $k = 31$ random projections, the reported p-value in Table 5 was $1.292542 \cdot 10^{-05} = 31 \times 4.169491 \cdot 10^{-07}$

Thus, all times, the proposed procedure detects the existence of influence of both factors and, also, the existence of interaction between them. The influence of the monthly rainfall is not so strong but its influence is declared statistically significant more than eighty percent of times. The Gaussian hypothesis is rejected just about 11.4% thus, making doubtful when a procedure requiring gaussianity is appropriate. However, in this section we have also employed the test proposed in [5] which does not rely on this assumption and admits heteroscedasticity obtaining similar results.

5 A comparison with MANOVA

In this section we compare our procedure with some usual MANOVA tests. We make the comparison analyzing the two examples proposed in `R` as illustration of the `manova` procedure. First one is taken from [17, pag. 81]. Second one was proposed in [22].

As stated, those examples were selected to illustrate the `manova` procedure which is based on gaussianity assumptions. Thus, we have decided to assume also this hypothesis and employ the usual one-dimensional ANOVA test to analyze the projections.

In this section we will assume the model (2) with $\gamma_i^{r,s} = 0$, with t restricted to belong to the finite set $\{1, ..., d\}$ and, thus, $\mathcal{H}$ being the d-dimensional euclidean space $\mathbb{R}^d$ with the usual norm and scalar product. Those assumptions allow us to apply Theorem 2.1 in this setting since this result holds in every separable Hilbert space.

There are some practical points worth to be considered. First one is related to the distribution to be employed to select the random directions. Stationarity is often a reasonable assumption in random processes and, in consequence there is no need to distinguish between values of t. However, in this setting there is no reason to assume that the marginals of the underlying distribution are the same. Thus, it may be not too wise to select, for instance, μ in Theorem 2.1 as the standard Gaussian distribution.

Second problem is related to the fact that if d is low as compared with k the dependences between marginals may translate in a useless Bonferroni's correction.

Concerning the first question, our proposal is, first, to apply an affine transformation to the data in order to make them centered in mean and its covariance matrix to be the identity. Then, we choose as μ the standard d-dimensional Gaussian distribution.

Since the affine transformations do not affect the usual MANOVA tests, this transfor-

mation has no influence on them. Thus, the proposed affine transformation can be seen as an easy way to choose easily the random projections taking into account the dependences of the marginal distributions. In other words, this transformation plus μ equal to the standard Gaussian distribution is equivalent to no transformation on the data and μ being a Gaussian distribution with a covariance matrix depending on the dependences between marginals.

With respect to the dependences between marginals, as stated in Section 2.2.1, if Bonferroni's correction is going to be chosen, a not too high value for k may be needed. This fact is shown in Table 7 where we show how the proportion of rejections of the null hypothesis "the additive has no effect" (a more detailed description of the data appears in Example 5.1), decreases smoothly when k goes from $k = 5$ to $k = 15$ and decrease abruptly if we take $k = 30$.

However, this fact does not happens with the factor rate. We think that the difference between those factors is that the effect of the first factor is so intense that it is noticeable in every direction and the bigger number of directions, the better the chance to find a very small p-value. However, the effect of the second factor is not so big and, then, the p-values are not so small and, from some point on (around $k = 15$) the better chance to find a small p-value by increasing k is overridden by the Bonferroni correction.

This is the reason that made us to propose to choose k not too big when compared with d.

However, a way to overcome this difficulty is to employ bootstrap as follows. Let us assume that we have a d-dimensional sample and that we are interested in testing the null hypothesis H_0 (related to main factors, interactions, covariables,...).

After standardizing the sample as described, take k directions at random $v_1^0, ..., v_k^0$ (which remain fixed along the process). Compute the projections of the data on those directions, and apply to each set of projections a 1-dimensional ANOVA test. Thus, we have the p-values $p_i, i = 1, ..., k$ and take $p = \min_i p_i$.

In order to evaluate the significance of those values, let us, first, to center in mean by cells, the data in the sample. Then, let us take B bootstrap samples from the centered in mean data. For each bootstrap sample, compute the k projections of their points along the same directions, $v_1^0, ..., v_k^0$, that we employed before. Apply to each set of projections the same 1-dimensional ANOVA test as before and take the minimum of the obtained p-values.

Thus, we have the p-values $p^1, ..., p^B$. Since the bootstrap samples verify the null hypothesis for every factor (remember that the bootstrap samples were obtained from a sample centered in mean by cells), with this procedure we obtain a sample of the values to be expected under the null hypothesis in the previous process, which can be employed as reference to evaluate the significance of the value p obtained in the first step.

It can be checked in Table 7 that this procedure does not suffer from the previous effect and that increasing the number of random directions increases the power of the test. We have chosen $B = 500$ when applying bootstrap.

In Tables 7 and 9 the rows labeled "p-value" contain the p-values obtained in the first

application of the random ANOVA under the corresponding conditions. The rows labeled "Proportions of rejections" show the proportion of times in which the corresponding null hypothesis was rejected along 500 repetitions of the the random ANOVA test under the corresponding conditions. Thus, if the correction method is Bonferroni, the obtained p-values were modified as described in Section 4. If the correction method is Bootstrap, the correction method is the onedescribed in this section.

On the other hand, in Table 9 we only show the results obtained with $k = 15$ randomly chosen directions. This is due to that in Example 5.2 there is no so big variation of the results. However, more complete tables are available from the authors upon request.

Example 5.1 In this example we analyze some data on the production of plastic film which appear in [17, pag. 81]. This is the example used for the documentation of the function `summary.manova` in R [20] . These data are three-dimensional. On them it is analyzed how some characteristics of the plastic film ("tear", "gloss" and "opacity") vary depending on two factors ("rate" and "additive"') with two levels each ("Low","High"). Five measurements were taken under each set of production conditions.

Thus, we are in a 3-dimensional, 2-ways MANOVA problem with 2 levels in each factor and 5 observations in each cell.

First, in Table 6, we present the result to apply the MANOVA procedure with the Pillai test. Here we see that both factors (specially the rate) are found to be significative and also that there is no statistical evidence on the existence of interaction between them.

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
rate	1	0.6181	7.5543	3	14	0.003034 **
additive	1	0.4770	4.2556	3	14	0.024745 *
rate:additive	1	0.2229	1.3385	3	14	0.301782

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Table 6: Usual MANOVA for Krzanowski's data

Table 7 contains the results to apply our procedure under several conditions in order to show the differences of behavior between Bonferroni's and bootstrap corrections of the significance level. The explanation of the observed differences was already made. Here we only want to add that the p-values obtained by bootstrap procedure are closer to those in Table 6 independently on the number of projections and providing always good results, whereas with the Bonferroni correction the choice of the number of projections could be critical if the dimension of data is short and the factors are not so strong. Taking into account that bootstrap procedures are high time consuming, our recomendation is to use bootstrap procedures when d is short ($d \leq 30$) or a fine precision is needed. In functional data, where d is usually large ($d \geq 50$), Bonferroni's correction is fast and accurate enough.

Example 5.2 The second example corresponds with Example 9.6 in [22, pag. 461] as an example on testing interactions. The response variable is a vector of three components

Correction method		Number of projections	Rate	Additive	Interaction
Bonferroni	p-value	$k = 5$	0.002	0.037	0.339
		$k = 15$	0.003	0.043	0.770
		$k = 30$	0.009	0.069	0.837
Bootstrap	p-value	$k = 5$	0.012	0.006	0.242
		$k = 15$	0.000	0.022	0.380
		$k = 30$	0.006	0.024	0.294
Bonferroni	Proportions of rejections	$k = 5$	0.956	0.790	0.000
		$k = 15$	0.998	0.604	0.000
		$k = 30$	1.000	0.194	0.000
Bootstrap	Proportions of rejections	$k = 5$	1.000	0.968	0.000
		$k = 15$	1.000	0.968	0.000
		$k = 30$	1.000	0.992	0.000

Table 7: p-values obtained with the random ANOVA and proportions of null hypothesis rejected for Krzanowski's data

related with growth of paspalum grass (the fresh weight of roots (gm), the maximum root length (mm) and the fresh weight of tops (gm)) under two factors: Temperature and Treatment. Temperature has four levels (14, 18, 22, 26 ^{o}C) and Treatment has two levels ("Control" and "Inoculated"). The goal is to investigate the effect on growth of inoculating these plants with a fungal infection.

As in Example 5.1 we begin with the results obtained when applying an usual MANOVA procedure which finds evidence of significance of both factors (very strong in the case of the factor Temp). It neither finds evidence of interaction.

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
Treat	1	0.2053	3.2722	3	38	0.03145 *
Temp	3	1.5009	13.3489	9	120	$1.286 \cdot 10^{-14}$ ***
Treat:Temp	3	0.3206	1.5953	9	120	0.12410

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Table 8: Usual MANOVA for Seber's data

The results obtained by the proposed method are similar to those shown in Table 8. Again, $k = 15$ projections seems too much for a three dimensional data using the Bonferroni correction. Although the p-values in both tables give more or less the same information, the proportion of rejections for the factor Treatment with the Bonferroni correction is quite low. This does not occur with the Bootstrap method that provides a result closer to those in Table 8 but with a high computational cost. In this case, the computational cost of a single run (CC_{sr}) with $k = 15$ projections was about 0.13 sec whereas to obtain the p-value by Bootstrap ($B = 500$) the computational cost (CC_{Boot}) was 96.46 seconds ($CC_{Boot} \approx B \times (CC_{sr} + CC_{gen})$, where CC_{gen} is the time needed to select a bootstrap sample).

Correction method	Number of projections		Treat	Temp	Interaction
Bonferroni	$k = 15$	p-value	0.042	$6.67 \cdot 10^{-12}$	1
		Proportions of rejections	0.776	1.000	0.000
Bootstrap	$k = 15$	p-value	0.008	0.000	0.220
		Proportions of rejections	0.998	1.000	0.000

Table 9: p-values obtained with the random ANOVA and proportions of null hypothesis rejected for Seber's data

6 Discussion

In this paper a procedure to test complicated linear designs in the functional setting is proposed. The procedure is random since it is based on the replacement of the given functional data by randomly chosen one-dimensional projections.

The consideration of a number of projections is proposed as a way to stabilize the results, using Bonferroni's method to correct the obtained p-values.

The results obtained in the simulations carried out in Section 3 as well as the analysis of the real data set in Section 4 make us to be optimistic about its capability to discover the existence of influence of factors, interactions, covariables,...

It is also noticeable that the method proposed produces similar results to those obtained with usual MANOVA tests. However, we want to stress the fact that the flexibility of the procedure, makes easy to deal with those situations in which the assumptions are not fulfilled because, through the projections, it is possible to apply a 1-dimensional ANOVA test (like the proposed in [5]) in which the assumptions are less stringent.

Moreover, the proposed method can be applied to every random process in which we have a Hilbert space to which the trajectories belong to. Thus, it is not required that the random functions are defined on an interval or even on a one-dimensional subset.

On the other hand, in [11] the authors generalize Theorem 1.1 to general functional spaces without a Hilbert space structure. Thus, providing the way to also eliminate this assumption in our results.

References

[1] Abramovich, F., Antoniadis, A., Sapatinas, T. and Vidakovic, B. (2004). Optimal Testing in a Fixed-effects Functional Analysis of Variance Model. *Inter. J. Wavelets, Multiresolution Informat. Processing* **2**, 323–349.

[2] Abramovich, F. and Angelini, C. (2005). Testing in Mixed-Effects FANOVA Models. *J. Statist. Planning Infer.* **136**, 4326–4348.

[3] Angelini, C. and Vidakovic, B. (2002). Some Novel Methods in Wavelet Data Analysis: Wavelet ANOVA, F-test Shrinkage and Γ-Minimax Wavelet Shrinkage. Istituto per le Applicazioni del Calcolo "Mauro Picone". Rapporti Tecnici RT 256/02.

[4] Antoniadis, A. and Sapatinas, T. (2007). Estimation and Inference in Functional Mixed-effects Models. *Computat. Statist. and Data Anal.* **51**, 4793–4813.

[5] Brunner, E., Dette, H. and Munk, A. (1997). Box-Type Approximations in Nonparametric Factorial Designs. *J. Amer. Statist. Assoc.* **92**, 1494–1502.

[6] Cuesta-Albertos, J., Cuevas, A. and Fraiman, R. (2008) On projection-based tests for spherical and compositional data. *Preprint.*

[7] Cuesta-Albertos, J., del Barrio, T., Fraiman, R. and Matrán, C. (2007). The Random Projection Method in Goodness of Fit for Functional Data. *Computat. Statist. and Data Anal.* **51**, 4814–4831.

[8] Cuesta-Albertos, J., Fraiman, R., Galves, A., García, J. and Svarc, M. (2007). Identifying Rhythmic Classes of Languages Using their Sonority: a Kolmogorov-Smirnov Approach *J. Appl. Statist.* **34**, 749–761.

[9] Cuesta-Albertos, J., Fraiman, R. and Ransford, T. (2007). A Sharp Form of the Cramér-Wold Theorem. *J. Theoret. Probab.* **20**, 201–209.

[10] Cuevas, A., Febrero, M. and Fraiman, R. (2004). An ANOVA Test for Functional Data. *Computat. Statist. and Data Anal.* **47**, 111–122.

[11] Cuevas, A. and Fraiman, R. (2007). On Depth Measures and Dual Statistics: A Methodology for Dealing with General Data. *Preprint.*

[12] Fan, J. (1996). Tests of Significance Based on Waveleet Tresholding and Neyman's Truncation. *J. Amer. Statist. Assoc.* **91**, 674–688.

[13] Fan, J. and Lin, S.K. (1998). Test of Significance When Data Are Curves. *J. Amer. Statist. Assoc.* **93**, 254–261.

[14] Faraway J.J. (1997). Regression Analysis for a Functional Response. *J. R. Statist. Soc. B* **64**, 887–898.

[15] Fujikoshi, Y., Himeno, T. and Wakaki, H. (2004). Asymptotic Results of a High Dimensional MANOVA and Power Comparisons when the Dimension is Large Compared to the Sample Size. *J. Japan Statist. Soc.* **34**, 19–26.

[16] Guo, W. (2002). Inference in Smoothing Spline Analysis of Variance. *Technometrics* **93**, 1007–1021.

[17] Krzanowski, W. J. (1988) *Principles of Multivariate Analysis. A User's Perspective.* Oxford.

[18] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*. Springer Verlag, New York.

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

[20] R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.

[21] Schott, J.R. (2007). Some High-dimensional Tests for a One-way ANOVA. *J. Multivar. Anal.* **98**, 1825–1839.

[22] Seber, G.A.F. (1984) *Multivariate observations* Wiley, New York.

[23] Shohat , J.A. and Tamarkin, J.D. (1963). *The Problem of Moments*. American Mathematical Society, Providence.

[24] Spitzner, D.J., Marron, J.S. and Essick, G.K. (2003). Mixed-Model Functional ANOVA for Studying Human Tactile Perception. *J. Amer. Statist. Assoc.* **98**, 263–272.

[25] Tonda, T. and Fujikoshi, Y. (2004). Asymptotic Expansion of the Null Distribution of LR Statistic for Multivariate Linear Hyptohesis when the Dimension is Large. *Comm. Statist. Theory Methods* **33**, 1205–1220.