working papers

HOMOGENEITY TEST FOR FUNCTIONAL DATA BASED ON DEPTH MEASURES

R. Flores, R. Lillo and J. Romo

Abstract: In the context of functional data analysis, we propose a new method to test the homogeneity of families of functions. Based on some well-known depth measures, we construct four different statistics in order to measure distance between the two families. A simulation study is performed to check the efficiency of the tests when confronted with shape and magnitude perturbation. Finally, we apply these tools to measure the homogeneity in some families of real data, obtaining good results for these new methods.

Keywords: FDA, homogeneity test, functional depth.

Flores, Ramón, Departamento de Estadística, Universidad Carlos III de Madrid, Avda. de la Universidad Carlos III, 22, 28270 Colmenarejo (Madrid), e-mail: rflores@est-econ.uc3m.es.

Lillo, Rosa, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: lillo@est-econ.uc3m.es.

Romo, Juan, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: romo@est-econ.uc3m.es.

HOMOGENEITY TEST FOR FUNCTIONAL DATA BASED ON DEPTH MEASURES

RAMÓN FLORES, ROSA LILLO AND JUAN ROMO

Ramón Flores

Department of Statistics, Universidad Carlos III de Madrid, Spain

e-mail: rflores@est-econ.uc3m.es

Rosa Lillo

Department of Statistics, Universidad Carlos III de Madrid, Spain

e-mail: lillo@est-econ.uc3m.es

Juan Romo

Department of Statistics, Universidad Carlos III de Madrid, Spain

e-mail: romo@est-econ.uc3m.es

ABSTRACT. In the context of functional data analysis, we propose a new method to test the homogeneity of families of functions. Based on some well-known depth measures, we construct four different statistics in order to measure distance between the two families. A simulation study is performed to check the efficiency of the tests when confronted with shape and magnitude perturbation. Finally, we apply these tools to measure the homogeneity in some families of real data, obtaining good results for these new methods.

1. INTRODUCTION

In the last years functional data analysis (FDA) has become one of the most active domains of research in Statistics, because of its own interest and also for its applications in a number of context like medical science, biology, chemistry or social sciences. In essence, the objects of study in FDA are real functions which are assumed to be generated by means of a stochastic process. The functions are observed in a certain number of fixed points or time instants, but instead of being treated as multivariant

data, they are smoothed using appropriate tools that frequently rely on techniques of functional analysis. Nevertheless, a number of techniques of multivariate data have been adapted or generalized to the FDA context. The main references on this field are Ramsay-Silverman [RS05] and Ferraty-Vieu [FV06].

In this paper we address the problem of homogeneity between families of functions; that is, given two families of curves, we attempt to decide whether these two families have been produced or not by the same process, so they have equal probability distributions. This problem has been recently considered in the last years. Benko et al. [BHK09] present methods for testing equality of means between functional data that respectively rely on bootstrap and asymptotic procedures; and Horváth-Kokoszka [HK12], moreover, also describe tests to compare the quality of the covariance operator. López Pintado and Romo [LR09] (see also [LS93]) Romo-Pintado describe a rank test to establish the homogeneity of two functional samples, obtaining quite accurate results. Cluster algorithms have also been proposed in [ACMM03], for example. Finally, a different point of view is developed by Cuevas-Febrero-Fraiman in [CFF04], where an ANOVA F-test based in functional distances was proposed. In the same way, the approach we take in this paper is related to distances between the two functional families which are based on depth measures.

Consider an interval $I \subset \mathbb{R}$, and a finite family $\mathcal{F} = \{x_1 \dots x_n\}$ of real functions defined over the interval. We will assume that the functions lie in $C^1(I)$. The concepts of distance between families that we introduce in this paper will be based on the notion of *statistical depth*, a concept originated in the field of the statistical analysis of multivariate data, and then extended to functional data. In our context, a *depth functional* with respect to the family $\mathcal{F}$ will be a functional $d : C^1(I) \to \mathbb{R}$, whose value should depend on a certain way on the family $\mathcal{F}$ and also on a depth measure defined a priori. In this way, the value of d over the function will constitute a measure of how deep is the function f "inside" the family $\mathcal{F}$. By means of these functionals, we construct four families of statistics which are shown to be useful to decide if two families of functions are homogeneous or heterogeneous. In order to understand the behaviour of the measures with respect to differences of magnitud and shape in the families, we have tested our methods on several samples of simulated functions. Moreover, we discuss homogeneity in some real contexts, as Ramsay height data, the tecator family and the mitochondrial data MCO. See last section for details.

The structure of the paper is as follows. In Section 2 we present the concept of depth, introduce the notion of depth with regard to a family and describe our statistics. Section 3 reviews the measures of depth that are used in order to undertake the different homogeneity tests. Section 4 is devoted to the analysis of some families of simulated data, whereas in the last section we perform the test for the real data examples.

2. Distances between functional families

In the growing field of FDA, where functions are data, a crucial general goal is to define concepts which mimic or transport the usual notions in multivariate data analysis. The depth of functions was defined -in any of its versions-, in this sense, as a generalization to this context of a notion of centrality, being the deepest function of a certain family an adequate definition of the "median" of the data.

Usually, the concept of depth is quite related with some notion of distance between functions, a phenomenon which is more clearly seen in the multivariate versions. In order to introduce intuitive statistics that indicate in some sense distance between two samples, we propose the definition of depth of a family with respect to another. Given a certain measure of functional depth d, and given a family of functions $\mathcal{F}$ and another function g not necessarily in $\mathcal{F}$, we denote by $d_{\mathcal{F}}(g)$ the depth of g with regard to the family $\mathcal{F} \cup \{g\}$. We define henceforth the notion of the deepest function of a family with respect to another:

Definition 2.1. *Let $\mathcal{F}$ and $\mathcal{G}$ be two finite families of continuous functions defined in an interval I. The* deepest function *of $\mathcal{G}$ with regard to $\mathcal{F}$ is the function g of the family $\mathcal{G}$ which maximizes $d_{\mathcal{F}}(g)$ among $g \in \mathcal{G}$. We will denote this function by $\mathfrak{D}_{\mathcal{F}}(\mathcal{G})$, or simply $\mathfrak{D}(\mathcal{G})$ if the base family $\mathcal{F}$ is understood. If there are more than one function in $\mathcal{G}$ for which the depth is reached, we can choose any of them as $\mathfrak{D}_{\mathcal{F}}(\mathcal{G})$, or else we can consider the whole set as the* deepest subfamily *of $\mathcal{G}$ with regard to $\mathcal{F}$.*

Observe that the if the families $\mathcal{F}$ and $\mathcal{G}$ are large enough, the probability of finding two functions of $\mathcal{G}$ which maximize the depth gets small, so usually we could talk about the deepest function.

The definition of the deepest function of $\mathcal{G}$ with respect to $\mathcal{F}$ is addressed to propose a solution to the problem of homogeneity in the context of functional data. Given two or more families of functions, we say that the functions are homogeneous if they

come from the same experiment, and then have equal probability distributions. In our framework, to determine explicitly the distributions is usually a very difficult problem, so we are forced to design different strategies to test homogeneity. We offer hence a different approach to the problem, by using the depth measures to perform an analysis which, by nature, may include the distance between functions, their magnitude and their shape.

Our starting point are two samples of functional data, $\mathcal{F}$ and $\mathcal{G}$. The function $\mathfrak{D}_{\mathcal{F}}(\mathcal{G})$ or $\mathfrak{D}_{\mathcal{G}}(\mathcal{F})$ may supply interesting information about homogeneity. In this sense the concept of deepest function can be used in different ways. In the following, we propose several possible statistics which depend on the notion of deepest function and allow to undertake the analysis of homogeneity.

We define the first statistic $\mathbf{P_1}$ as

$$\mathbf{P_1}(\mathcal{F}, \mathcal{G}) = d_{\mathcal{F}} \mathfrak{D}_{\mathcal{G}} \mathcal{G}.$$

Probably this is the more natural approach to the homogeneity problem, since roughly speaking, the function $\mathfrak{D}_{\mathcal{G}} \mathcal{G}$ is the best estimator of the experiment which produces the sample $\mathcal{G}$. Hence, it is reasonable to compute how deep is this estimator with respect to $\mathcal{F}$. The bigger this depth, the less likely the two samples come from different experiments.

The second statistic is defined as a variation of the previous one:

$$\mathbf{P_2}(\mathcal{F}, \mathcal{G}) = |\mathbf{P_1}(\mathcal{F}, \mathcal{G}) - \mathbf{P_1}(\mathcal{F}, \mathcal{F})|.$$

This definition may be considered a kind of normalization of the previous one. It could happen that the nature of the experiment which originates the sample $\mathcal{F}$ makes impossible for any datum of the experiment to reach the value 1 (for example, if the experiment produces two well-defined "bands" of functions, or if some deep functions cross themselves in close points). In this case, $\mathbf{P_1}(\mathcal{F}, \mathcal{F})$ would give a good estimation of the maximum of these depths, and the difference $|\mathbf{P_1}(\mathcal{F}, \mathcal{G}) - \mathbf{P_1}(\mathcal{F}, \mathcal{F})|$ would be more informative than the value $\mathbf{P_1}(\mathcal{F}, \mathcal{G})$ alone; see the computations section for interesting questions about this issue. In this case, the families are likely to come from the same experiment as the statistic gets closer to zero. It would be probably equivalent to consider the quotient instead of the absolute value of the difference.

A different approach is given by the statistic

$$\mathbf{P_3}(\mathcal{F}, \mathcal{G}) = d_{\mathcal{F}}(\mathfrak{D}_{\mathcal{F}}\mathcal{G}),$$

which identifies the $\mathcal{F}$-depth of the deepest function of $\mathcal{G}$ with respect to $\mathcal{F}$. This is the function of the family $\mathcal{G}$ which is more likely to come from the experiment that generates the sample $\mathcal{F}$, and then it is relevant from the point of view of the classification. In particular, $|\mathbf{P_3}(\mathcal{F}, \mathcal{F})| = \mathbf{P_1}(\mathcal{F}, \mathcal{F})$ for any family $\mathcal{F}$.

Observe that the function $\mathfrak{D}_{\mathcal{F}}\mathcal{G}$ could not be a good estimator for the result of the experiment that generates $\mathcal{F}$. Then, if we intend to use it for the classification of experiments, it would also be interesting to produce a measure that controls simultaneously the $\mathcal{F}$-depth and the $\mathcal{G}$-depth of $\mathfrak{D}_{\mathcal{G}}\mathcal{G}$. One possible option is to define a measure in $[0, 1] \times [0, 1]$ whose values are the $\mathcal{F}$-depth and the $\mathcal{G}$-depth of $\mathfrak{D}_{\mathcal{F}}\mathcal{G}$; in this context, the first number would be the measure the depth itself, while the second would be interpreted as a control number of how sharp is the measure. However, this approach is bivariant, so we propose instead an alternative univariant version that avoid that disadvantage and catches essentially the same information:

$$\mathbf{P_4}(\mathcal{F}, \mathcal{G}) = |\mathbf{P_3}(\mathcal{F}, \mathcal{G}) - \mathbf{P_1}(\mathcal{F}, \mathcal{F})||\mathbf{P_3}(\mathcal{F}, \mathcal{G}) - \mathbf{P_1}(\mathcal{G}, \mathcal{G})|.$$

The greater this number, the less likely the two samples come from the same experiment.

Once the statistics are defined, we propose the following method for testing the null hypothesis of equality of distributions of the two functional samples. We use a bootstrap approach to take the final decision since the theoretical asymptotic distributions of these statistics are difficult to obtain.

1. Select a functional depth measure $d_{\mathcal{F}}$ and a statistic $\mathbf{P} = \mathbf{P}_i$ for some $i \in \{1, 2, 3, 4\}$, which will depend on the previous concrete choice. In this paper, to deal with $d_{\mathcal{F}}$ we will use Fraiman-Muniz depth, h-modal depth, random depth RPD, band depth BD and modified band depth mBD but there are other possibles choices for the depth measure.

2. Now consider the families $\mathcal{F}$ and $\mathcal{G}$, and propose as a null hypothesis H_0 that $\mathcal{F}$ and $\mathcal{G}$ come from the same experiment. We perform then a hypothesis test to reject (or not) H_0.

3. Define the family $\mathcal{H}$ as the union $\mathcal{F} \cup \mathcal{G}$, and obtain N bootstrap samples of $\mathcal{H}$ of size $|\mathcal{H}|$. For any $1 \leq j \leq N$, let $\mathcal{S}_j$ be the corresponding sample, denote

by $\mathcal{S}_j^1$ the family of the first $|\mathcal{F}|$ functions and by $\mathcal{S}_j^2$ the family of the last $|\mathcal{G}|$. Then compute $P_j = \mathbf{P}(\mathcal{S}_j^1, \mathcal{S}_j^2)$.

4. For an appropriate size α, compute a confidence interval I for the values P_j.

5. The null hypothesis will be rejected if and only if the functional $\mathbf{P}(\mathcal{F}, \mathcal{G})$ does not belong to the interval I, and in this case we will assume that the groups are not homogeneous.

The nature of the computations suggests to use unilateral confidence intervals. To obtain the critical value at 95% of confidence, we trim the five percent of the data in the appropriate side of the interval: in measures $\mathbf{P}_1$ and $\mathbf{P}_3$ should be the smaller values, while in the normalized values will be the bigger ones.

Below we present our results with both simulated data and real data, but we first review the functional depths we use.

3. FUNCTIONAL DEPTHS

The concept of depth in the context of functional data analysis generalizes the same notion for multivariate data, which in turn was an extension to the multidimensional framework of the order statistics and other univariant measures. While the multivariate measures are mainly addressed to explore a certain centrality of a point in some real vector space, the different nature of the functional data forces the statistics to consider another features of the functions involved, as the shape of the functions or the amount of time they spend in a certain range of real numbers. In this sense, we have chosen different depth measures which in turn explore different features of the functions inside the families. We start with the pioneering work of Fraiman-Muniz, whose goal is to measure how much time every function is deep inside the family.

Fraiman-Muniz depth. Consider an instant a bunch of curves $\{x_1(t), \dots x_n(t)\}$ defined on the interval $[0, 1]$. Denote by $I(-)$ the indicator function, and consider, for every $i \in 1 \dots n$, the function:

$$F_{n,t}(x_i(t)) = \frac{1}{n} \sum_{k=1}^{n} I(x_k(t) \le x_i(t)),$$

and also the univariate depth

$$D_n(x_i(t)) = 1 - |\frac{1}{2} - F_{n,t}(x_i(t))|.$$

Then, the Fraiman-Muniz depth of the function $x_i(t)$ is defined in [FM01] as the integral:

$$FM(x_i(t)) = \int_0^1 D_n(x_i(t))dt.$$

h-modal depth. This measure was first defined by Cuevas et al. [CFF06] and is addressed to identify the functional mode of the family. Consider again a family of curves $\{x_1(t), \ldots x_n(t)\}$, select a value h which should be interpreted as a bandwidth, and also consider a kernel function defined on the real positive numbers. Then the h-modal depth of the function $x_i(t)$ with respect to K and h is defined as:

$$hD_n(x_i, h) = \sum_{k=1}^n \frac{K(\|x_i - x_k\|)}{h}.$$

In this paper, as recommended by the aforementioned authors, we take the norm L^2, h as the 15th percentile of the empirical distribution of the norms $\|x_i - x_k\|$, and K a convenient truncated Gaussian kernel.

Random projection depths. These two versions of depth were proposed by Cuevas et al. [CFF07], and combine random projections of the functions of the family in different directions with a bivariate data depth which is used to order the corresponding results. More precisely, given a family of functions $\{x_1(t), \ldots x_n(t)\}$ and ν, a realization of a stochastic process whose values are random directions, we define the projection of x_i along the direction ν as

$$T_{i,\nu} = \int_0^1 \nu(t)x_i(t)dt,$$

and analogously,

$$T'_{i,\nu} = \int_0^1 \nu(t)x'_i(t)dt$$

considering the derivatives instead of the trajectories of the function. If we select a bivariate data depth D and assume P realizations of V, we may define the following two versions of the random projection depth:

$$RPD1(x_i) = 1/P \sum_{p=1}^P D(T_{i,\nu}, T_{i,\nu}),$$

which takes into account only the trajectories of the functions, and

$$RPD2(x_i) = 1/P \sum_{p=1}^{P} D(T_{i,\nu}, T'_{i,\nu})$$

which considers the functions and their derivatives.

In this note, the role of D to compute depths will be played by the bivariate version of h-modal depth. Moreover, we checked that the results obtained in our contexts using RPD1 and RPD2 were similar, and as the second one was computationally harder, in this paper we use only the first version, which we will denote simply by RPD.

Band depth. In [LR09], López-Pintado and Romo define two different versions of a new depth of essentially geometric nature. It is based on the concept of band, understood as a portion of the plane that is delimited by the family of curves. More precisely, fix the family $\mathcal{F}$, and given a function $x \in C(I)$, denote by $G(x)$ the graph of x. Then, for every j such that $2 \leq j \leq n$, the n-th band depth is defined by the formula:

$$BD_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \ldots \leq i_j \leq n} I\{G(x) \subseteq B(x_{i_1}, x_{i_2}, \ldots, x_{i_j})\}.$$

Here $x_{i_1}, x_{i_2}, \ldots, x_{i_j}$ are functions in the family and $B(x_{i_1}, x_{i_2}, \ldots, x_{i_j})$ is defined in the following way:

$$B_j(x) = B_j(x; x_{i_1}, \ldots, i_j) = \{(t,y) \in I : \min_{k=1_1,\ldots i_j} x_k(t) \leq y \leq \max_{k=1_1,\ldots i_j} x_k(t) \leq x(t)\}.$$

Here I stands, as usual, for the indicator function. Note that $BD_n^{(j)}(x)$ measures the proportion of j-uplas $(x_{i_1}, x_{i_2}, \ldots, x_{i_j})$ in $\mathcal{F}$ such that x belongs to the band determined by them.

The global band depth compiles all the previous measures:

Given a family $\mathcal{F}$ as above and a value J such that $2 \leq J \leq n$, the band depth of a function x is defined as

$$BD_{n,J}(x) = \sum_{j=2}^{J} BD_n^{(j)}(x).$$

Of course, from an analytic point of view, the most logical choice for J is n, so we collect all the posible information given by the curves in the family $\mathcal{F}$. However, if

$|\mathcal{F}|$ is big, the depth can turn to be computationally intractable. The authors prove that the value is quite stable in J, so in this paper we will use $J = 2$. In this case, the depth depends generally on non-degenerate bands.

The authors also define a modified version of the band depth, by considering bands in the interval I, instead of bands in the plane:

$$A_j(x) = A_j(x; x_{i_1}, \ldots, i_j) = \{t \in I : \min_{k=1_1,\ldots i_j} x_k(t) \leq x(t) \leq \max_{k=1_1,\ldots i_j} x_k(t) \leq x(t)\}.$$

Now the authors consider a Lebesgue measure λ on the interval (usually the standard one), and define as in the previous case:

$$mBD_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \ldots \leq i_j \leq n} \frac{\lambda(A_j(x))}{\lambda(I)},$$

again with $2 \neq j \neq n$. Now the definition of the modified band depth is analogous to the previous one:

$$mBD_{n,J}(x) = \sum_{j=2}^{J} BD_n^{(j)}(x),$$

for $2 \leq J \leq n$.

4. SIMULATION STUDY

In order to describe the characteristics and features of our procedures, we perform a simulation study using the four different statistics defined in Section 2 and the five depth measures defined in the previous section: Fraiman-Muniz, h-modal, random measure, band depth and modified band-depth. We consider six functional populations in the space $\mathcal{C}[0, 1]$, which are considered as the realizations of a stochastic process $X(-)$ which has continuous trajectories in the interval $[0, 1]$.

Sample 0. This is the reference set, generated by a Gaussian process

$$X(t) = E(t) + e(t)$$

with mean function $E(t) = E(X(t)) = 30t^{3/2}(1 - t)$, and $e(t)$ is a centered Gaussian process, whose covariance matrix is given by $Cov(e_i, e_j) = 0.3 * \exp(-\frac{|t_i - t_j|}{0.3})$.

The remaining sets are produced by perturbing the generation process in two ways. The first three suffer magnitude contamination in the mean, while the covariant matrix does not change.

Sample 1. This set is generated by the Gaussian process $X(t) = 30t^{3/2}(1-t) + 1 + e(t)$.

Sample 2. In this case the contamination is smaller than in Sample 1: $X(t) = E(X(t)) = 30t^{3/2}(1-t) + 0.5 + e(t)$.

The next samples are obtained from the reference set of Sample 0 by changing in a more drastic way the mean function, and also the covariance matrix. These changes give rise to shape contamination.

Sample 3. This set is generated by the Gaussian process $X(t) = 30t(1-t)^2 + e(t)$, where $e(t)$ is defined in the same way as above.

Sample 4. Defined as $X(t) = 30t(1-t)^2 + h(t)$, where $h(t)$ is a centered Gaussian process whose covariance matrix is given by $Cov(e_i, e_j) = 0.5 * \exp(-\frac{|t_i - t_j|}{0.2})$.

Sample 5. The last group combines the previous cases, being defined by $30t^{3/2}(1-t) + h(t)$. Hence, the perturbation here is only induced by the process $h(t)$.

The routines used to undertaken the simulations were developed in R and are available upon request. We adopt the following notation:

- For $i \in \{1, 2, 3, 4, 5\}$ The five sets of simulated functions will be denoted by S_i.
- For every $k \in \{1, 2, 3, 4\}$, the statistic $\mathcal{P}_k$ used in the hypothesis test will be as defined in Section 2.
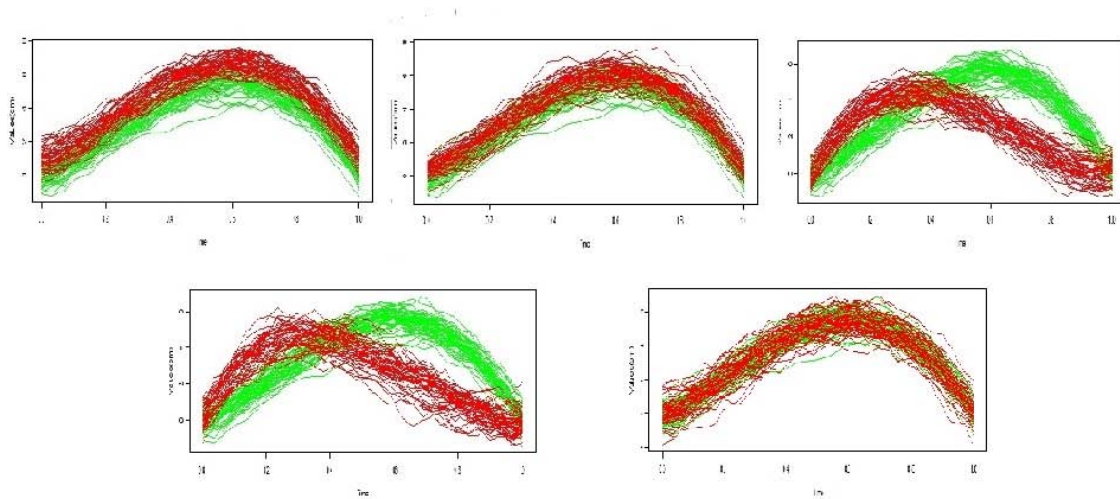

FIGURE 1. From left to right and from up down, the five samples. In green, the reference family.

We use the following method to test homogeneity. Select a depth measure $d_|$ and a statistic $\mathbf{P} = \mathbf{P}_k$ from the list above. Now generate 50 functions with the algorithm for S_0 and 50 functions with the algorithm for S_i, for a certain i. Each curve is observed in 30 equidistant points. Now compute $\mathbf{P}_i(S_0, S_i)$. Then consider 1000 standard bootstrap samples of size 100 of the family $\mathcal{H} = S_0 \cup S_i$. For any $1 \leq j \leq 1000$, let $\mathcal{S}_j$ be the corresponding sample, denote by $\mathcal{S}_j^1$ the sample of the first 50 functions and by $\mathcal{S}_j^2$ the sample of the last 50, and compute $P_j = \mathbf{P}(\mathcal{S}_j^1, \mathcal{S}_j^2)$. With this 1000 values we compute an unilateral confidence interval for a confidence of 0.05. Now the null hypothesis is that $\mathcal{F}$ and $\mathcal{G}$ come from the same experiment, and we reject if and only if $\mathbf{P}(\mathcal{F}, \mathcal{G})$ does not belong to the interval I. Finally we repeat the whole process 100 times and count the number of rejections. Our results are shown in the tables, and commented below.

The results of our computations are listed in Table 1 with the information of both the previous measures and the rank tests. There we denote respectively by FM, dmode, RPD, BD and mBD, the Fraiman-Muniz depth, h-modal depth, random projection depth, and band depth and modified band depth. We maintain the notation for the samples of functions which are already described and are the target of our study. For each statistic, the table shows the number of rejections in 100 essays, for the usual level 0.05 of confidence.

We may analyze our results from three differents point of view, focusing respectively in the classification criteria, the depth measures or the populations. Considering criteria, it is clear from the data that the most accurate is $\mathbf{P_3}$ as it always distinguishes the families, with a perfect 100% of success. Its normalized version works also quite well, being uneffective when combining it with h-modal depth, or when the magnitude contamination is too small. The measure $\mathbf{P_3}$ only presents problems when its associated depth is BD, and same phenomenon happens to $\mathbf{P_4}$.

From the point of view of the depth measures, it is clear all of them work well (at least 75 rejections in almost all the cases) except the band depth, so in case we need to use these kind of measure, the modified version is clearly preferrable. Finally, it is apparent from the simulations that the difficulties only appear if the magnitude contamination is really small (Sample 2) or we combine the two perturbations (Sample 5) and the measures are powerful when confronted with other type of contaminations. Note that, excluding BD, the measures detect always the difference for families $\mathcal{F}_1$, $\mathcal{F}_3$ and $\mathcal{F}_4$ and $\mathcal{F}_5$.

4.1. **Sensibility analysis.** We carry out a sensibility analysis of our approach with respect to several aspects that can be considered:

Size of the bootstrap. In order to test the importance of the size of the bootstrap sample, we also undertook some test cases enlarging it to 1000 and 3000. The computation time increased in a significant way, while there were not an apparent change in the conclusions of our study. So we may conclude that our statements are stable with regard to the size of the bootstrap resampling.

Confidence. We choose the usual signification level of 0.05, but in order to check the robustness of our results, we tested some of the data for a level of 0.025. We obtain the same conclusions as in the 0.05 case, so may assume that our measures are also robust in this sense.

Symmetry. We also check what happens if in each case, we take the population S_i as the reference sample in the hypothesis test, and S_0 as the test sample. Again, the results where similar to the ones that are shown in the paper. While a priori it would be a good idea to take into account this symmetric values, we check that the benefit of this strategy would be exiguous, and at the same time the computational cost would increase significantly.

Power test. In order to show the performance of the measures introduced in the paper, we have carried out a power test for a concrete model case. Consider the gaussian stochastic process $X(t) = 30t^{3/2}(1 - t) + e(t) + \eta$, which depends on the parameter $\eta > 0$, and consider the measure $\mathbf{P}_1$ referred to Fraiman-Muniz depth. We know by Table 1 that for $\eta = 1$ and $\eta = 0.5$ the measure separates this family from the reference family in 100 out of 100 replications. After generating another 100 replications for $\eta = 0.25$, we found that the measure detected heterogeneity in all cases. However, for $\eta = 0.1$ the family gets really close to Family 0 and then the measure only discriminates in 9 out of 100 cases.

Observe that in the same situation (see Table 2) the rank test produces a perfect score for samples 1, 3 and 4, but it fails to prove homogeneity when the difference of

	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_3$	$\mathbf{P}_4$
FM				
1	100	100	100	100
2	100	100	100	100
3	100	100	100	100
4	100	100	100	100
5	57	52	100	76
h-modal				
1	100	100	100	48
2	95	92	100	80
3	100	100	100	29
4	100	100	100	100
5	87	70	100	83
RPD				
1	100	100	100	98
2	100	100	100	28
3	100	100	100	100
4	100	100	100	100
5	16	19	100	36
BD				
1	100	19	100	97
2	52	21	100	25
3	100	32	100	93
4	46	30	100	92
5	65	22	100	73
mBD				
1	100	100	100	76
2	100	99	100	41
3	100	100	100	100
4	100	100	100	100
5	67	46	100	83

TABLE 1.   Simulation results

magnitude is small (Family 2) or when the shape contamination is important (Family 5). The latter was early advised in [LR09].

To prevent disfunctions caused by outliers, is is usual to define trimmed measures, considering a subsample of functions in F, for example the 95% of deeper functions. The smaller these numbers are, the greatest the probability that both series of data come from the same experiment. We have checked the trimmed measures in some of

Rank test simulated data	FM	h-modal	RPD	BD	mBD
Sample 1	100	100	100	100	100
Sample 2	65	41	57	49	44
Sample 3	100	100	100	100	100
Sample 4	100	100	100	100	100
Sample 5	61	94	6	99	78

TABLE 2. Rank test for simulated data

our previous computations, but the results were very similar to the measures without trimming, so we offer here the results of the latter.

5. REAL DATA

In this last section, we illustrate the validity of our methods with four different real data sets: a) *Ramsay growth curves dataset*, which consists of the height (in cm) of 93 people measured throught time; b) *MCO data*, where data measure calcium content in cardiac cells of mice; c) *Tecator spectrometric data set*, which consists of 215 infrared spectra of meat samplest obtained by a Tecator IFF Analyzer, and d) the second derivative of the spectrometric data. The results of the rank test are included at the end of the section.

In the tables below, CV (critical value) stands for the extreme of the unilateral confidence interval of the test. Observe that for the measures $\mathbf{P}_1$ the null-hypothesis is rejected when the value of the statistic is smaller than CV, whereas in the remaining two we reject when the value of $\mathbf{P}_n$ is larger than CV. In the corresponding columns labeled "Rej." we specify if the null-hypothesis is rejected or not in each case.

5.1. **Ramsay data.** We start our analysis of real cases with the classical growth dataset first studied by Ferraty-Vieu in [FV06], and also analyzed more recently by López Pintado and Romo in [LR09] and by Alonso, Casado and Romo in [ACR12]. The variables are the 93 growth curves for 39 boys and 54 girls, measured between 1 and 18 years, and we intend to test the homogeneity of samples by sex. The results are shown in Table 3.
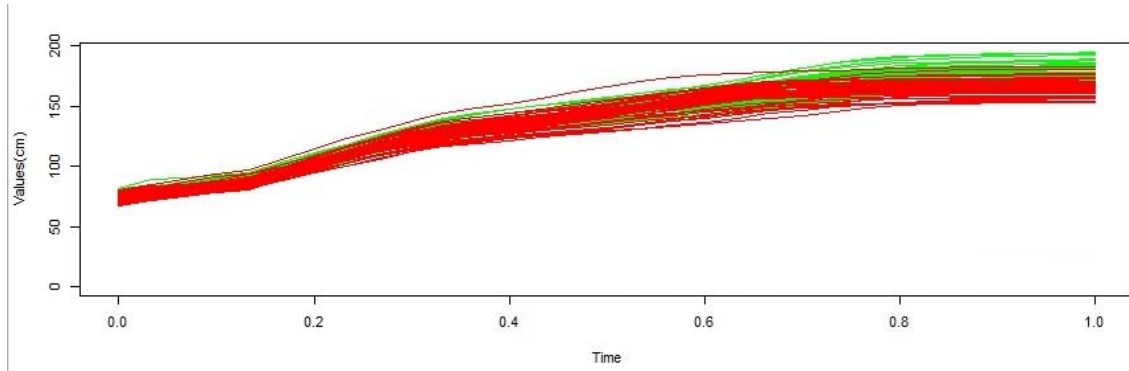
FIGURE 2. Ramsay growth data.

	$\mathbf{P}_1$	CV	Rej.	$\mathbf{P}_2$	CV	Rej.	$\mathbf{P}_3$	CV	Rej.	$\mathbf{P}_4$	CV	Rej.
FM	0.879	0.827	No	0.128	0.089	Yes	0.893	0.888	No	0.0007	0.001	No
dmode	6.685	6.153	No	3.157	0.722	Yes	4.908	6.579	Yes	4.389	1.276	Yes
RPD	0.224	0.211	No	0.088	0.03	No	0.239	0.24	Yes	0.00002	0.00004	Yes
BD	0.05	0.147	Yes	0.271	0.125	Yes	0.194	0.204	Yes	0.011	0.006	Yes
mBD	0.392	0.46	Yes	0.121	0.0572	Yes	0.497	0.499	Yes	0.00004	0.0003	No

TABLE 3.   Measures for Ramsay data

It is obtained that at 95% level of confidence, the measure $\mathbf{P}_2$ establishes a clear difference between male and female data for the four considered depths. Moreover, the four statistics separate when they are combined with band-depth and modified band depth. The "natural" measure $\mathbf{P}_3$ is effective in four out of five cases, and the remaining one (when combining with Fraiman-Muniz) is very close to being so. For these data, only $\mathbf{P}_1$ seems to be not quite powerful, as it separates only when combined with BD and mBD. Looking at the 24 outcomes of Table 3, we obtain a 70.8 percent of separations, which increases to 83.3 percent if we do not take into account the measure $\mathbf{P}_1$. Observe also that for these data the rank test only separates in half of the cases, and in particular is ineffective for mBD. It is also remarkable that both methods show weakness when combined with Fraiman-Muniz depth, which seem not quite appropriate to confront these kind of observations.

5.2. **MCO data.** Now we apply our measures to the mitochondrial calcium overload dataset ([RM03]), previously studied from a statistical point of view in [CFF06] and [BCC11]. The functional variable measures the level of mitochondrial calcium in mouse cardiac cells, as high levels of this element usually imply good protection of these cells in the event of ischemia process. The ultimate goal of the study is to test
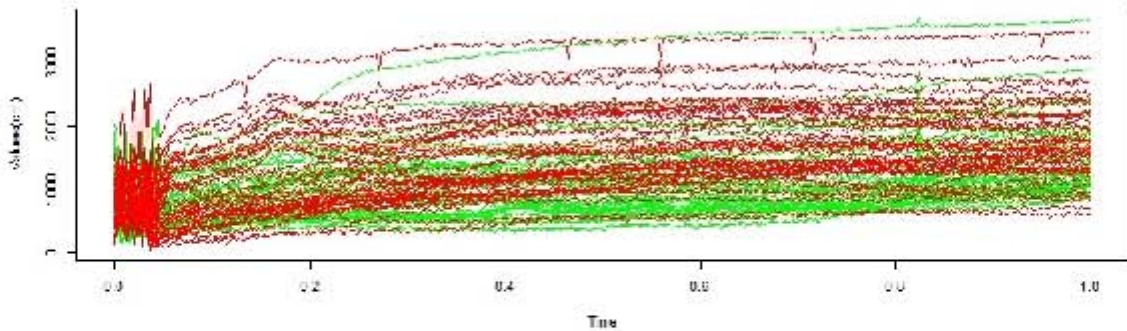
FIGURE 3. Mitochondrial calcium data.

	$\mathbf{P}_1$	CV	Rej.	$\mathbf{P}_2$	CV	Rej.	$\mathbf{P}_3$	CV	Rej.	$\mathbf{P}_4$	CV	Rej.
FM	0.689	0.814	Yes	0.263	0.127	Yes	0.922	0.894	No	0.0002	0.002	No
dmode	4.425	5.865	Yes	2.649	1.392	Yes	6.854	6.799	Yes	0.046	0.416	No
RPD	0.227	0.205	No	0.0366	0.033	Yes	0.244	0.239	No	0.000007	0.00004	No
BD	0.047	0.07	Yes	0.096	0.088	Yes	0.078	0.111	Yes	0.008	0.002	Yes
mBD	0.338	0.449	Yes	0.181	0.077	Yes	0.502	0.498	No	0.00026	0.00034	No

TABLE 4.   Measures for MCO data

the power of the drug Cariporide to increase the levels of calcium in the cells. The dataset consist in a control group of 45 observations and in a treated group of 44. The levels of MCO are measured every ten seconds during an hour, so each function is observed in principle at 360 points; however, the data which correspond to the first three minutes are eliminated from the sample, as they show a high variability which depend on factors that are hard to control.

In Table 4 we present the results of our computations for the mitochondrial data MCO.

Different to the case of Ramsay data, we do not know a priori if the data are naturally split into two families or not. Again the measure $\mathbf{P}_2$ gives the biggest evidence to the splitting hypothesis, as it shows heterogeneity in all the cases. The measure $\mathbf{P}_1$ also offers support to that hypothesis, as it only fails to make difference when combining with the random depths. Measure $\mathbf{P}_3$ only rejects homogeneity in

half of the cases, and $\mathbf{P}_4$ just one. From the point of view of the depth, band-depth shows again difference in all the families. For these data, the rank test shows heterogeneity in two cases, when it is carry out with the h-modal and band depths.
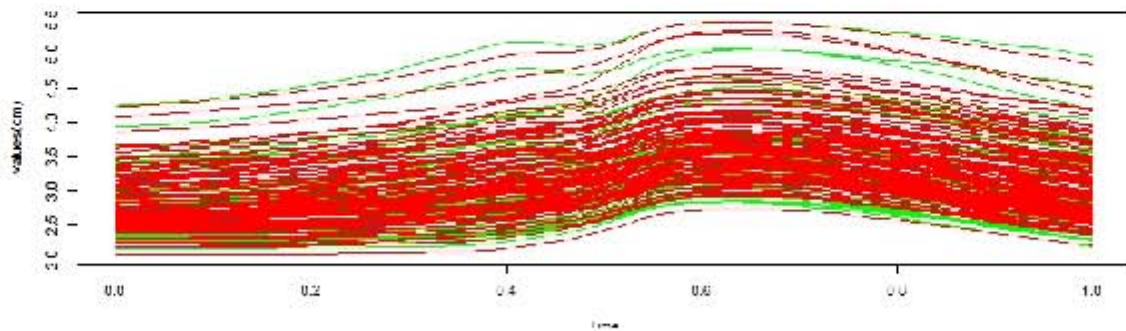

FIGURE 4. Tecator data.

5.3. **Tecator.** The tecator dataset have been intensively studied in the last years, see for example [FV06], [LY08] and [MLR13]. Tecator is a commercial name for a Infracted Food Analyzer, that in this case is used to measure the infrared absorbance spectrum of meat samples. These absorbances are given as functions of the intensity of the light measured just before and just after passing through the sample. The observations measure the contents of moisture, protein and fat in every sample of meat, and the goal is to separate two samples according to their different levels of fat. The discrete observations consist in 100 channel-absorbance spectrum for a given wavelength, which are made continuous using a B-spline basis of order 6. The dataset is divided in data with high fat content (77 observations) and data with low (the remaining 158). Following the approach of the aforementioned papers of Ferraty-Vieu and Li-Yu, we have computed our homogeneity measures also for the spectrometric data and for the second derivative of it. Recall that the discrete derivative is defined by means of the differences between subsequent points where the values for the functions are taken. The results for the first case appear in Table 5:

	$\mathbf{P}_1$	CV	Rej.	$\mathbf{P}_2$	CV	Rej.	$\mathbf{P}_3$	CV	Rej.	$\mathbf{P}_4$	CV	Rej.
FM	0.946	0.866	No	0.032	0.094	No	0.983	0.959	No	0.00001	0.0002	No
dmode	9.301	7.686	No	0.144	2.062	No	9.446	8.91	No	2.008	3.173	No
RPD	0.243	0.233	No	0.007	0.016	No	0.249	0.248	No	0.0000004	0.000003	No
BD	0.382	0.344	No	0.085	0.095	No	0.457	0.417	No	0.0002	0.001	No
mBD	0.511	0.476	No	0.007	0.037	No	0.518	0.514	No	0.003	0.00002	Yes

TABLE 5.   Measures for tecator data

Our computations support the widespread impression that the meat samples of the tecator data may proceed from the same family. As just one out of our 24 measures is able to separate the data (concretely $\mathbf{P}_4$ combined with modified band-depth), it is quite likely that this is an outlier instead of a genuine difference. Moreover, it can be seen as the critical values are usually quite far from the extremes of the corresponding interval.

More evidence is extracted from the rank test, that shows homogeneity in the five cases, and always in a quite robust way. The evidence then suggests that we cannot reject the hypothesis of equality between the two families.
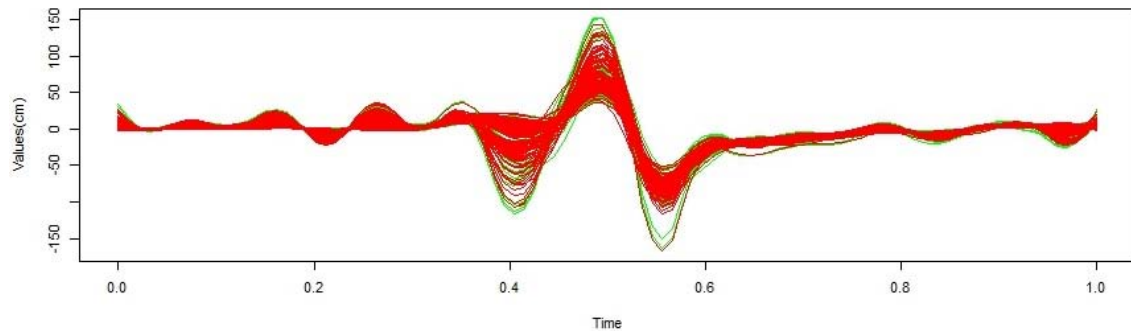

FIGURE 5. Tecator second derivatives.

5.4. **Tecator second derivatives.** There is quite more evidence of heterogeneity in the family of the second derivatives, as we may check in Table 6:

As in the previous familiy MCO, both $\mathbf{P}_1$ and $\mathbf{P}_2$ are able to separate, in this case four out of five cases, and again the other two measures seem not powerful in this case.

	$\mathbf{P}_1$	CV	Rej.	$\mathbf{P}_2$	CV	Rej.	$\mathbf{P}_3$	CV	Rej.	$\mathbf{P}_4$	CV	Rej.
FM	0.835	0.859	Yes	0.091	0.067	Yes	0.919	0.887	No	0.0001	0.0008	Yes
dmode	10.581	12.032	Yes	2′729	0.722	Yes	13′413	12.402	No	0.874	5.663	No
RPD	0.215	0.225	Yes	0.025	0.017	Yes	0.237	0.233	No	0.00003	0.00008	No
BD	0.121	0.089	No	0.056	0.1	No	0.179	0.121	No	0.00006	0.00437	No
mBD	0.448	0.461	Yes	0.054	0.036	Yes	0.498	0.482	No	0.101	0.0003	Yes

TABLE 6.   Measures for tecator data (second derivatives)

Rank test	Ramsay	MCO	Tecator	Tecator 2
FM	1733	2140	8427	7737
h-modal	1233	1625	8553	8490
RPD	1721	2051	8296	7768
BD	1159	1482	8136	6989
mBD	1703	2140	8427	7757
CV	1623.095	1781.395	7595.08	7595.08

TABLE 7.   Rank test in real data

The scheme is very similar to that case, except for the fact that band-depth gives no difference in any of the four cases. It is also remarkable that $\mathbf{P}_4$ only separates when combining with modified band-depth, just as it happens in the possibly outlier case described above.

The rank-test gives support in these case to the hypothesis of non homogeneity, as it shows it in all of the five observations.

6. Discussion

In this paper, we have defined some new measures of distances between families of functions to solve the problem of homogeneity in the context of functional data analysis. Combining these measures with the depth functions defined by Fraiman-Muniz, Cuevas-Fraiman-Muniz and López-Pintado-Romo, we propose a hypothesis test based on the bootstrap methology and apply it to a number of simulated and real functional data. Our measures shows their effectiveness in detecting differences of magnitudes and shape in some families generated by gaussian processes, and moreover are able to show heterogeneity for Ramsay data, mitochondrial data and the second derivatives tecator data. It is significant that our methods show homogeneity in the

tecator data without differentation, a phenomenon widely treated in the literature. It is also worth to comment that our method improves the rank-test in some cases.

Once the concept of depth of a function with regard to a family is defined, several generalizations appear to be possible. For example, the family of tecator data discussed above shows that there is information about homogeneity hidden in the derivatives that cannot be directly extracted from the original functions. Hence, it should be interesting to define and describe a unified way to deal with all the depth measures and statistics used in our work when applied at the same time to all the functions and all their derivatives. It is likely that such a notion would be able to show patterns in the homogeneity of the families that could not be deduced without differentiation. On the other hand, it would be also interesting to define some measures that allow to test at the same time the homogeneity of several families of functions. We plan to undertake this task in subsequent work.

References

[ACMM03]   C. Abraham, P. Cornillon, E. Matzner-Løber and N. Molinari. *Unsupervised curve clustering using B-splines. Scand. J. Stat.*, 30(3): 581–595, 2003.

[ACR12]   A. Alonso, D. Casado and J. Romo. *Supervised classification for functional data: A weighted distance approach. Comput. Stat. Data An.*, 56(7): 2334-2346, 2012.

[BCC11]   A. Baíllo, A. Cuevas and J.A. Cuesta-Albertos. *Supervised classification for a family of Gaussian functional models. Scand. J. Stat.*, 38: 480-498, 2011.

[BHK09]   M. Benko, W. Härdle and A. Kneip. *Common functional principal components. Ann. Stat.*, 37: 1–34, 2009.

[CFF04]   A. Cuevas, M. Febrero and R. Fraiman. *An anova test for functional data. Comput. Stat. Data An.*, 47: 111-122, 2004.

[CFF06]   A. Cuevas, M. Febrero and R. Fraiman. *On the use of bootstrap for estimating functions with functional data. Comput. Stat. Data An.*, 51: 1063–1074, 2006.

[CFF07]   A. Cuevas, M. Febrero and R. Fraiman. *Robust estimation and classification for functional data via projection-based depth notions. Computation Stat.*, 22: 481-496, 2007.

[FM01]   R. Fraiman and G. Muniz. *Trimmed means for functional data. Test*, 10: 419–440, 2001.

[FV06]   F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice.* Springer-Verlag, New York, 2006.

[HK12]   L. Horváth and P. Kokoszka. *Inference for functional data with applications.* Springer-Verlag, New York, 2009.

[LR09]   S. López-Pintado and J. Romo. *On the concept of depth for functional data. J. Am. Stat. Assoc.*, 104: 718–734, 2009.

[LS93]   R. Liu and K. Singh. *A quality index based on data depth and multivariate rank test. J. Am. Stat. Assoc.*, 88: 257–260, 1993.

[LY08]   B. Li and Q. Yu. *Classification of Functional Data: a segmentation approach. Comput. Stat. Data An.*, 52: 4790–4800, 2008.

[MLR13]   B. Martín-Barragán, R. Lillo and J. Romo. *Interpretable support vector machines for functional data. Eur. J. Oper. Res*, to appear.

[RS05]   J.O. Ramsay and B.W. Silverman. *Functional data analysis.* Springer-Verlag, New York, 1972.

[RM03]   M. Ruiz-Meana, D. García-Dorado, P. Pina, J. Inserte, L. Agulló and J. Soler-Soler. *Cariporide preserves mitochondrial proton gradient and delays ATP depletion in cardiomyocites during ischemic conditions. Am J. Physiol.-Heart C.*, 285: 999–1006, 2003.