Chemometrics and Intelligent Laboratory Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Homogeneity tests for functional data based on depth-depth plots with chemical applications

Alejandro Calle-Saldarriaga [a], Henry Laniado [a], Francisco Zuluaga [a], Víctor Leiva [b,*]

[a] Department of Mathematical Sciences, Universidad EAFIT, Medellín, Colombia
[b] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso. Chile

ARTICLE INFO

Keywords:
Bootstrapping
Data science
DD plots
Hypothesis testing
Nonparametric statistics
Robustness

ABSTRACT

One of the standard problems in statistics is determining if two samples come from the same population, that is, testing homogeneity for two samples. In this paper, we propose homogeneity tests in the context of functional data, adopting an idea from multivariate analysis corresponding to the depth-depth plot. This plot is a multivariate generalization of the quantile-quantile plot. We propose some statistics based on the depth-depth plot, and use bootstrapping to approximate their null distributions. We conduct simulations to state the empirical size and power of the proposed tests, obtaining better results than other homogeneity tests considered in the literature. We detect that our test has very high power in relation to other competing tests. We employ many different depths based on what is proposed in the literature to see which is more suitable for this kind of homogeneity testing. Finally, we illustrate the obtained results with chemical heterogeneous data to show potential applications, getting consistent results.

1. Introduction

Functional data analysis (FDA) is an area of statistics of important current development, where the data are functions. The term FDA was coined by Ramsay (1982) [53], but the area is older and dates back to the 50s [27,55]. With the advance of technology, continuously recorded data have become more common, and thus interest in FDA has spiked, particularly in times of big data and data science [2,51]. In the FDA context, we consider that certain functions originated the data that we record discretely, and that those functions are the sample members, not the explicit discrete data. Pre-processing discrete data for smoothing them is a usual step, but some methods do not need this. Often, the continuum in which functional data take values is the time. However, recently certain potential applications have appeared for spatially recorded or spatially correlated functional data [14,24–26,40]. Other applications have investigated the potential of FDA characterization, comparison, and classification for chemical data [1,4]. For a complete introduction to FDA, the interested reader is referred to [19,52], as well as to [10,65] for reviews of the recent advancements in the area.

There are different two-sample tests for homogeneity with non-functional data, as for example those presented in [36,60]. Previous work on testing for homogeneity in the two-sample problem for functional data include testing for location –or mean– [11,16,41,47], for equality of covariance operators [23,32,45], or equality of functional principal components [39,49]. If two samples come from the same population, they are homogeneous, otherwise, they are heterogeneous. In FDA, we rarely consider explicit distributions and then we say that two samples are homogeneous if they come from the same parent or generator process.

Testing general equality in distribution of specific characteristics of the samples was studied in [29] for multivariate and infinite-dimensional distributions (that is, functional) by considering specific measures of distance between the data. A Cramér-von Mises type statistic to test equality of two-sample distributional functionals was considered in [28]. By using a L^2 type criterion, tests for two-sample problems in the field of oceanography were proposed in [3]. The same type of criterion was employed in [31] for the empirical characteristic function to propose a two-sample test with univariate and multivariate functional data. The kernel method was utilized in [69] to state a maximum mean discrepancy type test for the two-sample problem in a functional space.

The concept of data depth in the multivariate context was introduced in [61] and a first notion of functional depth was stated in [21]. The idea behind this notion is to measure how much time each function is deep inside the sample, that is, how surrounded the function is by other

* Corresponding author.
E-mail addresses: victorleivasanchez@gmail.com, victor.leiva@pucv.cl (V. Leiva).

https://doi.org/10.1016/j.chemolab.2021.104420
Received 30 November 2020; Received in revised form 7 August 2021; Accepted 5 September 2021
Available online 14 September 2021
0169-7439/

functions. Their idea is to measure univariate depth for each instant of time and the deepest function is those that maximizes these univariate depth functions on average.

The random-Tukey depth (RTD) was proposed in [8], which approximates the Tukey depth. Instead of considering all the possible one-directional projections of each point or curve in a sample, it only takes into account some random one-directional projections of each point or curve. This makes the depth computationally efficient while maintaining good convergence properties.

A depth based on the bands defined by curves was proposed in [37]. A band is a portion of the plane that is delimited by a given number of curves. Then, we count how much each curve is contained in each of those bands, with a curve being deeper as more bands contain it. The original measure chooses all the possible bands created by each combination of curves contained in a sample. If the sample size is large, then this is computationally intractable. Thus, an adjusted band depth (ABD) was proposed in [42], which only considers the bands generated in every pair of curves of the sample.

The notion of depth-depth (DD) plots was introduced in [35], which are a way of comparing multivariate distributions of two samples utilizing depth measures. There are several examples of successful uses of depth measures in FDA, as for example in [59], where the boxplot was extended to FDA and called it the functional boxplot; or as in [9], where the DD plot was employed for supervised functional data classification.

Other two-sample homogeneity tests can be based on functional generalizations of the concepts of ordering from low dimensional data (quantiles for univariate data and depth for multivariate data), believing that these orderings reflect the law governing the process that generates the data samples. A nonparametric test based on depth of combined samples was proposed in [20].

Alternative orders which are not based on depth can be found in the literature. For example, by using the epigraph and hypograph of a function proposed in [38], we can order from largest to smallest values or from smallest to largest values (as opposed to depth, which organizes data from the center-out). The dimension of a functional data problem was reduced by employing epigraph and hypograph in [22], generalizing quantiles of functional data derived in [38]. After transforming the functional samples, traditional bivariate two-sample tests were applied. A new order for functions based on areas under curves utilized for generalizing the concept of the Kendall coefficient for infinite-dimensional data was introduced in [62]. To the best of our knowledge, a general two-sample homogeneity test for functional data using nonparametric tools, and particularly DD plots, has not been studied until now.

To propose a test from the DD plot, one must be able to compute a statistic from the set of pairs that comprise this plot. In a multivariate setting, some statistics for testing homogeneity of scale and location based on the construction of DD plots in multivariate samples were proposed in [33]. Other statistics have also been proposed in the literature. For example, a modification of the Kurskal-Wallis test using ranks given by the DD plot construction was proposed in [7]. A test statistic based on the correlations between the depths of the combined sample between each of the individual samples was derived in [46]. Note that homogeneity based on DD plots and their corresponding test statistics (and in general for many kinds of test statistics) has been successfully and widely employed in the multivariate literature. However, observe that all of these test statistics based on the DD plot only consider specific differences between samples considering location or scale, but they do not consider differences in distributions as a whole. Therefore, also to the best of our knowledge, in the literature on the topic, there are not statistics based on DD plots that can detect whether two samples can have been generated from different distributions, similarly to the philosophy of the Kolmogorov-Smirnov test in the univariate case [6].

Therefore, our primary objective is to construct multivariate homogeneity tests based on DD plots for FDA. Note that we are not proposing a new measure of data depth for FDA but homogeneity tests that employ the existing approaches from the literature. The secondary objective is to demonstrate that the new tests are more powerful in varied scenarios than other two-sample tests proposed until now. We evaluate our tests by simulations, knowing previously whether the samples are homogeneous or not, and following the simulation scenarios proposed in [22] as well as further scenarios. Since distributions of functional data are rarely considered in an explicit way, their direct comparison is often unfeasible. Then, a good homogeneity test must compare different aspects of the samples like means, variance/covariance structure, and curve shapes simultaneously, while a two-sample test is not good if only contrast means between samples. We compare our tests with order-based tests in size for several simulation scenarios. We believe the power of these tests can be improved in some cases. Our approach is general because we test equality in law and not in specific characteristics of the samples. In this work, we evaluate different depths or measures in various simulation scenarios empirically, assessing what measures work best for two-sample tests using DD plots. We detect that our test has very high power compared to other competing tests and employ different depths based on what is found in the literature to assess which is more adequate for this kind of homogeneity testing.

The rationale of the proposed statistic is that it assesses whether two samples differ in some characteristics of the distributions that are generating the data and not just in one specific characteristic. This is the reason why we have proposed diverse simulation scenarios and shown numerically that our statistic does indeed work in these scenarios. The deficiency of the other statistics proposed in the literature till the date versus our statistic is that we propose a more general method, that is, the proposed statistic considers any kind of heterogeneity between the two samples, while other statistics consider a more specific heterogeneity (in location or in scale).

Our statistic, such as it was proposed, is not designed to identify differences between any specific parameter of the two functional distributions to be tested but whether both functional samples come from the same functional distribution or not. If the test rejects the null hypothesis, this test detects a difference in the distributions that are generating both samples to be contrasted. Nevertheless, this decision does not identify which are the parameters that may be giving rise to this difference, similarly to the philosophy of the Kolmogorov-Smirnov test in the univariate case [6], as mentioned. In the multivariate case, the differences in location parameters could generate some specific shapes in the DD plot, and analogously for the differences between the scale and asymmetry parameters [35]. However, due to the nature of the functional data, where the mean curve can even have a very small depth value, these same differences could not be stated. Then, under the alternative hypothesis and in general for any statistic, the shape of the DD plot in functional data should not be a criterion to determine if the difference between the two functional distributions that generate the data is due to changes in a location parameter or in a scale parameter. Nonetheless, although a general purpose test, as one based on the statistic proposed in the present investigation, is helpful in a broad range of settings, once the hypothesis of homogeneity is rejected, it is natural to ask us in what way the two functional distributions differ. Thus, exploring the shape of a DD plot might have some information to offer regarding the nature of the difference, since this is the reason for considering various rationales in the existing DD plot-based test statistics in a multivariate setting. Therefore, identifying in what functional parameter(s) the distributions are differing based on the test statistic proposed in this study is an interesting aspect to be further explored.

The paper is organized as follows. Section 2 presents functional depths, depth measures for functional data, and homogeneity tests proposed in the literature. In Section 3, we introduce our DD plot-based tests. In Section 4, the results of a simulation study are reported while comparing our tests with other competing tests. Section 5 applies the tests proposed to real-world chemical data. Finally, in Section 6, we provide the main insights of this work and ideas for future research.

2. Functional depth

In this section, we introduce the concepts of functional depth, depth measures for functional data, and homogeneity tests proposed in the literature used in this investigation.

The depth measures may be employed to generalize the quantile concept to the multivariate case. These measures provide an inward-out ordering of the sample elements, where a multivariate median can be defined as that deepest point. Similar ideas can be adapted to the FDA context utilizing depth measures for some statistical concepts like centrality, shape, or closeness between sample curves. In functional depth, one can interpret the deepest function on a sample as the median function.

These notions of inward-out ordering have been generalized for FDA. Thus, many depths have been proposed for this ordering with the sample being a set of curves, where the functional depth can be used to explore different statistical features of these curves. To define the different notions for functional depth employed in this work, let us introduce below the main definitions and preliminary notations.

Consider a sample of functions $\mathscr{X} \equiv \{x_1(t), \ldots, x_n(t)\}$, where each $x_i(t)$, for $i \in \{1, \ldots, n\}$, is a function observed on the same interval $\mathscr{T} \subset \mathbb{R}$. In practice, functions are observed on a discrete set, but we assume that some smoothing steps are performed in order to obtain smooth and continuous functions

The Fraiman and Muniz (FM) depth [21] is defined as follows. Fix a $t' \in \mathscr{T}$ and then evaluate each curve at t' obtaining a univariate observed sample $\{x_1(t'), \ldots, x_n(t')\}$. Now, for i-th curve evaluated at t', consider the univariate depth $D_n(x_i(t'))$, and state the FM depth for $x_i(t) \in \mathscr{X}$ as

$$\mathrm{FM}(x_i(t)) = \int_{\mathscr{T}} D_n(x_i(t))\mathrm{d}t, \quad i \in \{1, \ldots, n\}, \tag{1}$$

that is, the expression stated in (1) is the average of the depths for all $t' \in \mathscr{T}$. Note that D_n defined in (1) can be any notion of univariate depth, such as

$$D_n(x_i(t')) = 1 - |0.5 - \hat{F}_{n,t}(x_i(t'))|, \tag{2}$$

where $\hat{F}_{n,t}$ defined in (2) is the empirical distribution function on $x_1(t')$, ..., $x_n(t')$. Hence, the deepest function in the sample of curves $\mathscr{X}$ is the deepest, on average, for each $t \in \mathscr{T}$.

The random projection depth (RPD) proposed in [13] considers random univariate projections of the curves. Let $v(t)$ be a Brownian motion on the interval $\mathscr{T}$. A realization $v_j(t)$ of $v(t)$ can project each curve $x_i(t) \in \mathscr{X}$ onto $\mathbb{R}$ by using the standard inner product given by

$$\langle v_j(t), x_i(t)\rangle = \int_{\mathscr{T}} v_j(t)x_i(t)\mathrm{d}t, \quad i,j \in \{1, \ldots, n\}. \tag{3}$$

Denote the projection formulated in (3) by $r_{i,j}$. Let $v_1(t)$, ..., $v_p(t)$ be p realizations of the Brownian motion $v(t)$, and $r_{i,1}$, ..., $r_{i,p}$ be the p univariate projections for a fixed curve $x_i(t)$, with $i \in \{1, \ldots, n\}$.

Let $\hat{F}_n(r_{i,j})$ be the empirical distribution function evaluated at $r_{i,j}$, but $\hat{F}_n$ is calculated with respect to the univariate sample $r_{1,j}$, ..., $r_{n,j}$, which come from all the functions projected on the same realization $v_j(t)$. In addition, let us define

$$d_{i,j} = \min\{\hat{F}_n(r_{i,j}), 1 - \hat{F}_n(r_{i,j})\}. \tag{4}$$

Then, the RPD for any $x_i(t) \in \mathscr{X}$ is the mean of the univariate depths for each function, which is expressed as

$$\mathrm{RPD}(x_i(t)) = \frac{1}{p}\sum_{j=1}^{p} d_{i,j}, \quad i \in \{1, \ldots, n\}, \tag{5}$$

where $d_{i,j}$ is defined in (4). Note from (5) that the deepest function is the deepest one on average in its univariate projections.

Another notion of depth was proposed in [12], which generalizes the mode to the functional setting and is called the h-modal (hmode) depth. This notion assigns larger depths to curves that have a larger number of neighboring curves in the sample. To compute this depth, select a bandwidth h and a kernel function defined on the real positive numbers. Then, the $h$mode depth for a curve $x_i(t) \in \mathscr{X}$ is established as

$$D_n^h(x_i, h) = \sum_{k=1}^{n} \frac{K(\|x_i - x_k\|)}{h}, \quad i \in \{1, \ldots, n\}, \tag{6}$$

where $\|\cdot\|$ is an appropriate norm for functions, such as the L^2-norm, K is an appropriate kernel (as the Gaussian kernel), and h is a tuning parameter.

A J-th ordered integrated depth that can capture global features of the sample was introduced in [43] by means of

$$D_J^{\mathrm{F}}(x_i(t); P) = \int_{\mathscr{T}} \cdots \int_{\mathscr{T}} D((x_i(t_1), \ldots, x_i(t_J))^\top) P_{(X(t_1),\ldots,X(t_J))^\top} \mathrm{d}t_J \cdots \mathrm{d}t_1, \quad i \in \{1, \ldots, n\}, \tag{7}$$

where P is a probability measure in the functional space, D is a depth measure in a finite space of dimension J, and $P_{(X(t_1),\ldots,X(t_J))^\top}$ is the corresponding probability measure in that finite space associated with the random vector $(X(t_1), \ldots, X(t_J))^\top$. The depth of each function is the average of the corresponding J-dimensional multivariate depths. In this work, we fix $J = 2$ and use the multivariate half-space depth proposed in [61]. In order to estimate $P_{(X(t_1),\ldots,X(t_J))^\top}$, we employ a plug-in estimate as empirical measure. Connections from the J-th ordered depth to the $(J-1)$-th derivative of the data were stated in [43], which means that the second ordered depth brings us information about the shape of the curve (first derivative), whereas the third ordered depth provides information related to the convexity of the curve (second derivative).

As mentioned, a homogeneity test was defined in [20] based on depth to propose a distance measure between samples of functional data. Specifically, let $\mathscr{X}$ and $\mathscr{Y}$ be the two samples that we want to test for homogeneity. Denote by $d_{\mathscr{X}}(y)$ the depth of y in a sample $\mathscr{X} \cup y$, and define by $\mathscr{D}_{\mathscr{X}}(\mathscr{Y})$ the function that maximizes $d_{\mathscr{X}}(y)$ for $y \in \mathscr{Y}$. Flores et al. [20] proposed the statistics given by

$$\begin{aligned} P_1(\mathscr{X}, \mathscr{Y}) &= d_{\mathscr{X}}\mathscr{D}_{\mathscr{Y}}\mathscr{Y}, \\ P_2(\mathscr{X}, \mathscr{Y}) &= P_1(\mathscr{X}, \mathscr{Y}) - P_1(\mathscr{X}, \mathscr{X}), \\ P_3(\mathscr{X}, \mathscr{Y}) &= d_{\mathscr{X}}\mathscr{D}_{\mathscr{X}}\mathscr{Y}, \\ P_4(\mathscr{X}, \mathscr{Y}) &= |P_3(\mathscr{X}, \mathscr{Y}) - P_1(\mathscr{X}, \mathscr{X})||P_3(\mathscr{X}, \mathscr{Y}) - P_1(\mathscr{Y}, \mathscr{Y})|. \end{aligned}$$

The idea behind P_1 is that the function $\mathscr{D}_{\mathscr{Y}}\mathscr{Y}$ is the most representative element of $\mathscr{Y}$. Then, if its depth is large in $\mathscr{X}$, it is most likely that $\mathscr{X}$ and $\mathscr{Y}$ are in the same family, that is, they are homogeneous. The idea behind P_3 is that if the function of $\mathscr{Y}$, most likely to come from the experiment $\mathscr{Y}$, is very deep in $\mathscr{X}$, then the two experiments are likely very mixed and so both of them come from the same population. Note that P_2 and P_4 are normalizations of P_1 and P_3. Hence, Flores et al. [20] used these statistics and bootstrapping to test the null hypothesis of homogeneity. Specifically, consider the functional samples $\mathscr{X} \equiv \{x_1, \ldots, x_n\}$ and $\mathscr{Y} \equiv \{y_1, \ldots, y_m\}$, defined on the same interval $\mathscr{T} \subset \mathbb{R}$. We assume that the functions lie on $C^1(T)$, that is, the space of functions with continuous first derivatives. Therefore, we wish to test the hypotheses established as $\mathscr{H}_0$: $\mathscr{X} =_{\mathscr{L}} \mathscr{Y}$ versus $\mathscr{H}_1$: $\mathscr{X} \neq_{\mathscr{L}} \mathscr{Y}$, where $\mathscr{L}$ means equality in law. Next, we propose tests that have high power in many deviations from $\mathscr{H}_0$, while maintaining an appropriate size.

3. DD plots and their relation with homogeneity

In this section, we introduce the proposed test and state its relation to the DD plot.

Let $\mathscr{Z} \equiv \mathscr{X} \cup \mathscr{Y}$ be a combined sample. In addition, let us define the DD plot of the combined sample as

$$\mathrm{DD}(\mathscr{X}, \mathscr{Y}, \mathscr{Z}) \equiv \{(D_{\mathscr{X}}(z), D_{\mathscr{Y}}(z)), z \in \mathscr{Z}\},$$

where D is an arbitrary measure in any of the sample spaces $\mathscr{X}$ or $\mathscr{Y}$. Then, $\mathrm{DD}(\mathscr{X}, \mathscr{Y}, \mathscr{Z})$ is a set of pairs of size $|\mathscr{Z}| = n + m$.

Following this idea, since depths try to characterize the distributions of samples, the DD plot between homogeneous samples must be similar to the identity [35], that is, the scatter plot generated by the DD plot should concentrate towards the identity line. Fig. 1 shows DD plots for homogeneous and heterogeneous samples. The DD plot on the left of Fig. 1 corresponds to homogeneous samples, where both samples follow a $\mathrm{N}_2(\mathbf{0}_{2\times1}, \boldsymbol{I}_{2\times2})$ distribution (bivariate normal), with $\mathbf{0}_{2\times1}$ being the 2×1 null vector and $\boldsymbol{I}_{2\times2}$ being the 2×2 identity matrix. The DD plot on the right of Fig. 1 is associated with heterogeneous samples, where the first sample is $\mathrm{N}_2(\mathbf{0}_{2\times1}, \boldsymbol{I}_{2\times2})$ distributed, and the second sample is $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributed, with $\boldsymbol{\mu} = [0.5, 1.3]$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.5 & 0.3 \\ 0.3 & 1.5 \end{pmatrix}$$

Note that the depth function used in Fig. 1 is the simplicial depth [34]. The points in the homogeneous DD plot stay concentrated towards the (0, 0)-(1, 1) line, whereas in the heterogeneous case they do not.

The scenarios in Fig. 1 display, with two artificial examples, the main idea of the test. Note that, in the first scenario, we consider two bivariate standard normal distributions and the DD plot is quite concentrated around the diagonal line. However, in the second scenario, we also consider two bivariate normals, but one of them with a weak correlation structure and a different mean. This aspect of the increase in correlation, even the increase in the marginal variances, highly distorts the DD plot, that is, its points are not concentrated on the diagonal line. The reason for the triangular shape is really unknown to us. The aspects that affect the DD plot could be differences in the marginal distributions, differences in location and scale parameters, or different dependency structure, among others, as shown in the simulation examples. Nevertheless, note that the proposed test does not consider how a DD plot deviates from the concentration of points on the diagonal line but whether it deviates or not from such a line. Capturing the way how the DD plots deviate from the diagonal line is indeed an ideal test statistic for homogeneity, which will be studied in a future investigation following the ideas indicated in the final section of the present work. Observe that our proposed test contrasts whether the points of the DD plot are concentrated around the diagonal line or not.

We use the proposed relationship presented in [35] between homogeneity and the DD plot to propose some statistics that can capture how concentrated the DD plot is towards the diagonal line that passes from (0, 0) to (1, 1).

Note that if two multivariate distributions are the same, then the DD plot is simply a segment on the 45° line from (0, 0) [33]. Inspired by this idea, here, we propose to employ a linear model with $\beta_0 = 0$ and $\beta_1 = 1$ to test, under the null hypothesis, that both samples come from the same distribution.

It is worth noting that the DD plot can be constructed utilizing any notion of depth, so that we must consider many depths to see which of them provides us with higher power for our test. To detect if the DD plot is concentrated towards the line passing through (0, 0)-(1, 1), we assume

$$D_{\mathscr{X},i} = \beta_0 + \beta_1 D_{\mathscr{Y},i} + u_i, \quad i = 1, \ldots, n+m, \tag{8}$$

with u_i being the usual error term in a regression model. Consider the null hypothesis $\mathscr{H}_0$: $\mathscr{X} =_{\mathscr{L}} \mathscr{Y}$ to be true, which also means that $D_{\mathscr{X}} = D_{\mathscr{Y}}$. By employing the formula for finding β_1 given in (8), and noting that $\mathrm{Cov}(D_{\mathscr{X}}, D_{\mathscr{Y}}) = \mathrm{Cov}(D_{\mathscr{X}}, D_{\mathscr{X}}) = \mathrm{Var}(D_{\mathscr{X}})$, under the null hypothesis, we get

$$\beta_1 = \frac{\mathrm{Cov}(D_{\mathscr{X}}, D_{\mathscr{Y}})}{\mathrm{Var}(D_{\mathscr{X}})} = \frac{\mathrm{Var}(D_{\mathscr{X}})}{\mathrm{Var}(D_{\mathscr{X}})} = 1. \tag{9}$$

Hence, by using the formula for β_0, we obtain that $\beta_0 = \mathrm{E}(D_{\mathscr{X}}) - \beta_1 \mathrm{E}(D_{\mathscr{Y}}) = \mathrm{E}(D_{\mathscr{X}}) - \mathrm{E}(D_{\mathscr{X}}) = 0$. This means that we should test whether $\beta_1 = 1$ and $\beta_0 = 0$. However, our model must be symmetric, such that the test is invariant when $D_{\mathscr{X}}$ is the independent or dependent variable. Then, from (8), let

$$D_{\mathscr{Y},i} = \beta_0 + \beta_1 D_{\mathscr{X},i} + u_i, \quad i = 1, \ldots, n+m, \tag{10}$$

be the model with different independent variables. Thus, from (9) and (10) and with the formulas for β_1 and β_0 above stated, we reach

$$\beta_1 = \frac{\mathrm{Cov}(D_{\mathscr{Y}}, D_{\mathscr{X}})}{\mathrm{Var}(D_{\mathscr{Y}})} = \frac{\mathrm{Var}(D_{\mathscr{Y}})}{\mathrm{Var}(D_{\mathscr{Y}})} = 1, \tag{11}$$

$$\beta_0 = \mathrm{E}(D_{\mathscr{Y}}) - \beta_1 \mathrm{E}(D_{\mathscr{X}}) = \mathrm{E}(D_{\mathscr{Y}}) - \mathrm{E}(D_{\mathscr{Y}}) = 0. \tag{12}$$

Therefore, from (11) and (12), we should test that $\beta_1 = 1$ and $\beta_0 = 0$ as a novel approach for homogeneity based on DD plots. Consequently, we must now propose a statistic to test the corresponding hypotheses. Since the values of $D_{\mathscr{X},i}$ and $D_{\mathscr{Y},i}$ do not necessarily follow a normal distribution, using the traditional t-test is not reasonable as the normality assumption of u_i could not hold. Then, we employ bootstrapping-t, which has second-order convergence properties [15], to propose a test that contrasts whether $\theta = \theta_0$ or not.

Suppose that, with the sample $\mathscr{X}$, we can estimate $\hat{\theta}$ and its standard error, namely $\hat{\sigma}$. By utilizing the Wald test, we consider the statistic defined as

$$T = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}}. \tag{13}$$

We approximate the distribution of T defined in (13) as follows. First,

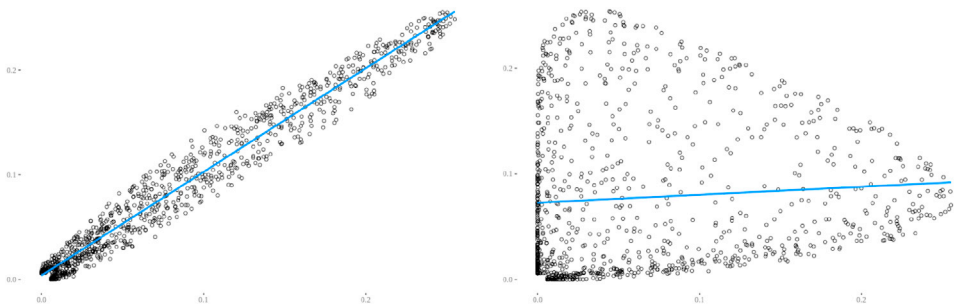

Fig. 1. DD plots for (left) homogeneous and (right) heterogeneous multivariate data.

draw bootstrap samples under $\mathscr{H}_0$ from $\mathscr{X}$. Second, call them $\mathscr{X}^*$, which yields $\hat{\theta}^*$ and $\hat{\sigma}^*$. Hence, estimate T by the bootstrap replicates as $T^* = (\hat{\theta}^* - \theta_0)/\hat{\sigma}^*$. By employing bootstrapping-t, we contrast whether $\beta_0 = 0$ and $\beta_1 = 1$ for the DD plot while resampling the original curves. Thus, compute DD($\mathscr{X}, \mathscr{Y}, \mathscr{Z}$) and get the least square (LS) estimates of the parameters stated in (8): $\hat{\beta}_1$ and $\hat{\beta}_0$. In addition, compute the standard error of the estimators for these parameters given by

$$\hat{\sigma}_{\beta_0} = \sqrt{\frac{\sum_{i=1}^{n+m}\hat{u}_i^2\sum_{i=1}^{n+m}D_{\mathscr{Y},i}}{(n+m-2)\sum_{i=1}^{n+m}(D_{\mathscr{Y},i}-D_{\mathscr{Y}})}}, \tag{14}$$

$$\hat{\sigma}_{\beta_1} = \sqrt{\frac{\sum_{i=1}^{n+m}\hat{u}_i^2}{(n+m-2)\sum_{i=1}^{n+m}(D_{\mathscr{Y},i}-D_{\mathscr{Y}})}}, \tag{15}$$

where $\hat{u}_i$ defined in (14) and (15) are the residuals when fitting the expression stated in (8) with the LS method and $D_{\mathscr{Y}}$ is the mean of the $D_{\mathscr{Y},i}$. Now, we use the statistic defined in (13) to compute

$$T_0 = \frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}, \quad T_1 = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\beta_1}}, \tag{16}$$

which are the test statistics. Then, since $\mathscr{H}_0$: $\mathscr{X} =_{\mathscr{L}} \mathscr{Y}$, we can resample $\mathscr{Z}$ as the null hypothesis states that the same model generated both data samples. This allows us to resample the sets of curves, which is equivalent to resampling each of them individually, under $\mathscr{H}_0$.

Let us call the new sample with replacement as $\mathscr{Z}^*$. Then, compute the points DD($\mathscr{X}, \mathscr{Y}, \mathscr{Z}^*$) and utilize them to obtain $T_0^{*(i)}$ and $T_1^{*(i)}$ as in (16), for $i \in \{1, \ldots, B\}$ bootstrap replicates of each test statistic. Now, establish the p-values of the test as

$$p_0 = 2\min\left\{\frac{1}{B}\sum_{i=1}^{B}\mathrm{II}(T_0^{*(i)} > T_0), \frac{1}{B}\sum_{i=1}^{B}\mathrm{II}(T_0^{*(i)} < T_0)\right\}, \tag{17}$$

$$p_1 = 2\min\left\{\frac{1}{B}\sum_{i=1}^{B}\mathrm{II}(T_1^{*(i)} > T_1), \frac{1}{B}\sum_{i=1}^{B}\mathrm{II}(T_1^{*(i)} < T_1)\right\}, \tag{18}$$

where II is the indicator function. We reject $\mathscr{H}_0$ for an appropriate level of α (for example, $\alpha = 0.05$). Since we have two p-values, we need to use a sequentially rejective multiple test to ensure a size α overall in our test. Employing the Holm-Bonferroni method proposed in [30], we order p_0 and p_1 defined in (17) and (18) increasingly. Let $p_{[1]}$ be the minimal p-value between p_0 and p_1, and $p_{[2]}$ the maximal p-value between p_0 and p_1. Therefore, reject $\mathscr{H}_0$ if $p_{[1]} < \alpha/2$ or $p_{[2]} < \alpha$, and do not reject $\mathscr{H}_0$, otherwise. The adjusted p-value of the test is then $p = \min\{2p_{[1]}, p_{[2]}\}$.

4. Simulation studies

This section reports the empirical power and size of the tests proposed in this work as well as the corresponding values for the test proposed by Flores et al. [20]. For conciseness, from now on, we refer to it as the Flores test, using P_4 with the FM depth, since power-wise is its best performer [20]. The other tests considered to compare our DD plot-based tests utilize different depth measures.

The simulations and methods mentioned in this work were implemented on `R` language [50] and the code is available from the authors upon request. We use the `R` packages named: (i) `ddalpha` [48] and `fda.usc` [18] for computing the different functional depths; (ii) `fda` [54] for handling functional data objects; (iii) `EMMIXskew` [66] for generating multivariate t-skewed distributed data; and (iv) `ggplot2` [68] and `tidyfun` [58] for producing the graphs contained in this work.

We rely our simulation procedure on the works presented in [22,49], where different data generating models are proposed, but we also add other essential scenarios. We fix a generator model and select others based on different deviations from $\mathscr{H}_0$. We determine the proportion of rejections of $\mathscr{H}_0$ when the same model generates both samples for measuring the test size, whereas for measuring its power, we compute the proportion of rejections of $\mathscr{H}_0$ when different models generate each data sample.

Table 1
Simulation scenarios.

Scenario	$\mathscr{X}$		$\mathscr{Y}$		Deviation
	μ	$\epsilon_{i,j}$	μ	$\epsilon_{i,j}$	
1	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i)$	$e_{i,j}$	none (size)
2	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i) + 0.25$	$e_{i,j}$	mean
3	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i) + 0.5$	$e_{i,j}$	mean
4	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i) + 0.75$	$e_{i,j}$	mean
5	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i) + 1$	$e_{i,j}$	mean
6	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i)$	$2e_{i,j}$	variance
7	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i)$	$4e_{i,j}$	variance
8	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i)$	$0.5e_{i,j}$	variance
9	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i)$	$0.25e_{i,j}$	variance
10	$\mu_1(t_i)$	$e_{i,j}$	$\mu_1(t_i)$	$h_{i,j}$	covariance
11	$\mu_1(t_i)$	$e_{i,j}$	$\mu_2(t_i)$	$h_{i,j}$	covariance and shape
12	$\mu_1(t_i)$	$e_{i,j}$	$\mu_2(t_i)$	$e_{i,j}$	shape
13	$\mu_3(t_i) + t_i$	$e_{i,j}$	$\mu_3(t_i) + 2t_i^3$	$e_{i,j}$	partial shape
14	$\mu_3(t_i) + t_i$	$e_{i,j}$	$\mu_3(t_i) + 4t_i^3$	$e_{i,j}$	partial shape
15	$\mu_3(t_i)$	$f_{i,j}^{(0)}$	$\mu_3(t_i)$	$f_{i,j}^{(8)}$	skewness
16	$\mu_3(t_i)$	$f_{i,j}^{(0)}$	$\mu_3(t_i)$	$f_{i,j}^{(9)}$	skewness

We consider 16 different models (scenarios) according to changes in mean, variance, covariance structure, shape, and skewness; which are defined in Table 1. This table summarizes the choices for the mean $\mu(t)$ and the multivariate process generating the error term $\epsilon_{i,j}$, where the deviation column describes which type of deviation from $\mathscr{H}_0$ is tested, or if we are determining the probability of type-I error. From the first scenario, we state the size, and from the considered deviations from $\mathscr{H}_0$, we establish the power. The size of the test should be near to the chosen significance level, and the power should be as high as possible. Fig. 2 shows the DD plots for all the different departures from the null hypothesis represented in Table 1. All the lines in these DD plots have some differences to the line passing through (0, 0)-(1, 1): some have $\beta_1 \neq 1$ and others have $\beta_0 \neq 0$.

For this simulation, we propose discretized models according to

$$x_{i,j}(t_i) = \alpha\mu(t_i) + \delta + \beta\epsilon_{i,j}, \tag{19}$$

where $\mu(t_i)$ is the mean function of the process $x(t)$ evaluated at a point t_i; δ, α, $\beta > 0$ are scalars; and $\epsilon_{i,j}$ is generated according to a zero-mean multivariate distribution in the discretized grid. We choose 30 equidistant points from the interval [0, 1], resulting in the generated grid. The model given in (19) can be expressed in a functional form as $x(t) = \alpha\mu(t) + \delta + \beta\gamma(t, s)$, and $\mu(t)$ may take forms such as (i) $\mu_1(t) = 30t^{3/2}(1 - t)$; (ii) $\mu_2(t) = 30t(1 - t)^2$; and (iii) $\mu_3(t) = \sqrt{2}\xi_1\sin(2\pi t) + \sqrt{2}\xi_2\cos(2\pi t)$, where $\xi_1 \sim$N(0, 10) and $\xi_2 \sim$N(0, 5).

To generate $\epsilon_{i,j}$ stated in (19), we consider one of the following three processes: (i) a multivariate normal distribution with covariance matrix $0.3\exp(-(|t_i - t_j|)/0.3)$, denoted by $e_{i,j}$; (ii) a multivariate normal distribution with covariance matrix $0.5\exp(-(|t_i - t_j|)/0.2)$, denoted by $h_{i,j}$; and (iii) following [67], a multivariate skew-t distribution with 4 degrees of freedom, a vector of k constant skew parameters and the same covariance matrix as $e_{i,j}$, denoted by $f_{i,j}^{(k)}$.

In Tables 2–4, we can see the empirical powers when the sample sizes are 25, 50, and 150, respectively. In each of these tables, scenario 1 refers to size simulations, while scenarios 2–16 refer to power simulations. For all sample sizes, in general, the DD plot-based test, using D_2^F (FD2) stated in (7), has very high power. For changes in the magnitude of the mean, the FM and Flores tests tend to have higher power than the other tests in all sample sizes. For changes variance/covariance, the Flores test has low

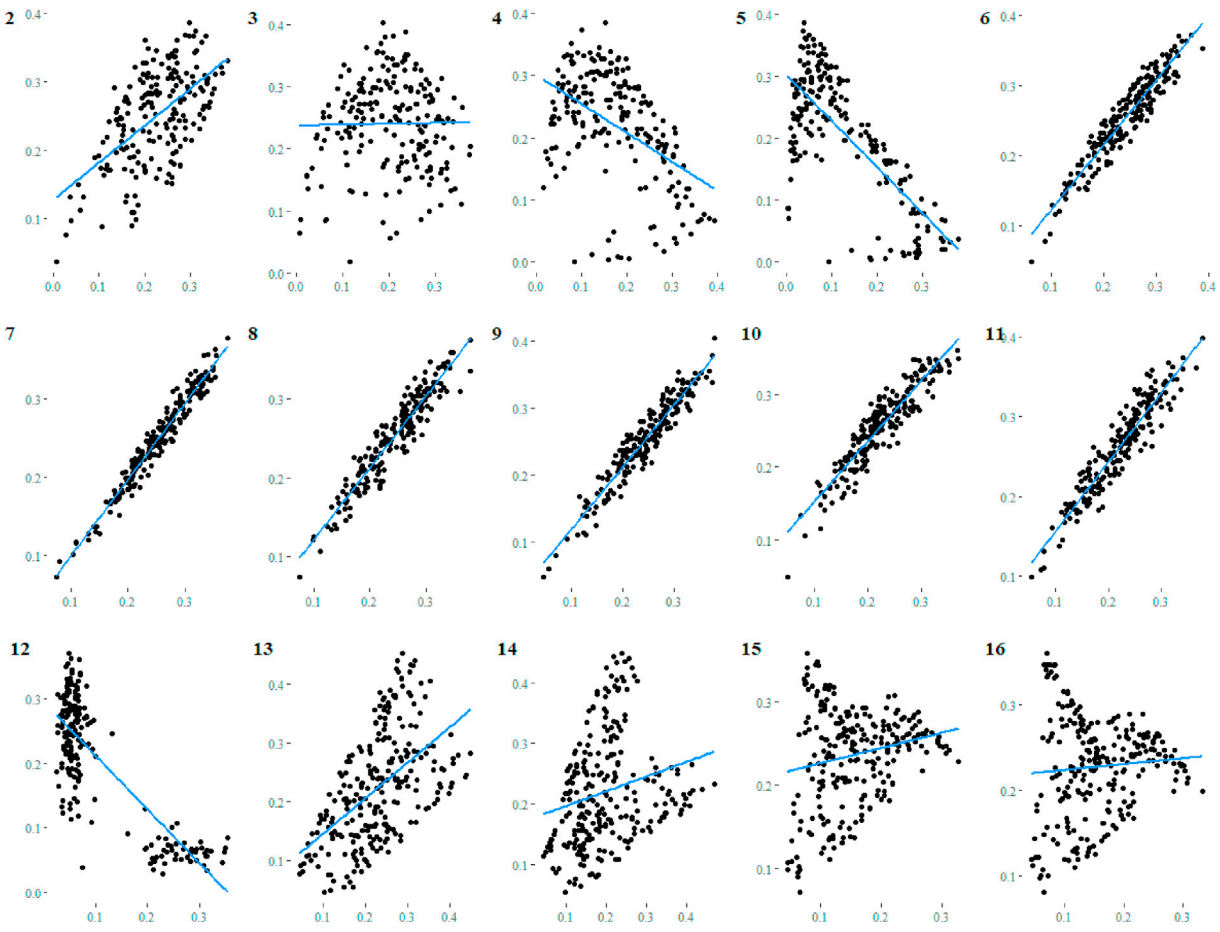

Fig. 2. DD plots for the simulation scenarios described in Table 1, when both samples have 125 curves. We use the FD2 depth to compute the DD plots. The blue lines correspond to the OLS fitted line to the points. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Empirical power and size of the indicated test for the listed scenario according to Table 1 based on simulations with samples of size 25.

Scenario	Flores	FM	hmode	RPD	RTD	FD2	ABD
1	0.11	0.02	0.00	0.01	0.00	0.01	0.00
2	0.47	0.09	0.00	0.07	0.03	0.02	0.00
3	0.99	0.87	0.03	0.74	0.08	0.52	0.18
4	1.00	1.00	0.31	1.00	0.14	0.99	0.86
5	1.00	1.00	0.76	1.00	0.27	1.00	1.00
6	0.14	0.78	0.75	0.74	0.79	0.97	0.99
7	0.43	0.90	0.59	0.87	0.99	1.00	0.95
8	0.02	1.00	0.99	0.89	0.94	1.00	1.00
9	0.03	1.00	1.00	1.00	1.00	1.00	1.00
10	0.07	0.24	0.48	0.18	0.21	0.55	0.76
11	0.54	1.00	1.00	1.00	0.14	1.00	1.00
12	0.61	1.00	1.00	1.00	0.16	1.00	1.00
13	0.29	0.15	0.78	0.54	0.62	0.22	1.00
14	0.35	0.66	1.00	0.99	0.68	0.75	1.00
15	0.96	0.57	0.03	0.57	0.52	0.75	0.99
16	0.94	0.58	0.01	0.70	0.60	0.87	1.00

Table 3
Empirical power and size of the indicated test for the listed scenario according to Table 1 based on simulations with samples of size 50.

Scenario	Flores	FM	hmode	RPD	RTD	FD2	ABD
1	0.04	0.05	0.08	0.01	0.03	0.02	0.00
2	0.56	0.43	0.00	0.00	0.00	0.18	0.03
3	0.99	1.00	0.41	0.90	0.15	0.99	0.67
4	1.00	1.00	0.55	1.00	0.66	1.00	1.00
5	1.00	1.00	0.91	1.00	0.93	1.00	1.00
6	0.11	0.91	0.67	0.92	1.00	1.00	1.00
7	0.44	0.97	0.20	0.98	1.00	1.00	1.00
8	0.02	1.00	1.00	1.00	1.00	1.00	1.00
9	0.06	1.00	1.00	1.00	1.00	1.00	1.00
10	0.05	0.60	0.69	0.26	0.61	0.86	0.99
11	0.92	1.00	1.00	1.00	0.90	1.00	1.00
12	0.89	1.00	1.00	1.00	0.92	1.00	1.00
13	0.33	0.24	0.96	0.78	0.96	0.94	1.00
14	0.53	0.79	1.00	1.00	1.00	1.00	1.00
15	0.97	0.71	0.00	0.88	0.90	0.98	1.00
16	0.99	0.89	0.00	0.97	0.93	0.99	1.00

power, but in these cases, the DD plot-based tests with FD2 and ABD have high powers. For changes in asymmetry of the data, the ABD DD plot-based test has high power across different sample sizes. The higher power of the Flores test as mean changes is explained by the deepest curves in a sample and by some statistics with relation to those curves. Nevertheless, the deepest curve in a sample is highly related to the center or the mean of the process. In that sense, the Flores test is like a mean test for functional data. This is the reason why it does not perform well when the changes between samples are not related to the mean of the stochastic processes that originated the curves.

In general, all methods presented acceptable empirical probabilities of type-I error, with most being less than the chosen significance level $\alpha = 0.05$. It is clear that our tests are overall more powerful than the Flores test. However, as mentioned, there is a more recent test in the literature, which was introduced in [22]. Fig. 6 graphically compares the empirical powers for our tests and the tests presented in [22], named the

Table 4
Empirical power and size of the indicated test for the listed scenario according to Table 1 based on simulations with samples of size 150.

Scenario	Flores	FM	hmode	RPD	RTD	FD2	ABD
1	0.04	0.00	0.03	0.01	0.03	0.07	0.01
2	0.83	0.91	0.00	0.08	0.05	0.79	1.00
3	1.00	1.00	0.06	0.99	0.66	1.00	1.00
4	1.00	1.00	0.82	1.00	0.99	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	0.15	0.99	0.66	1.00	1.00	1.00	1.00
7	0.59	0.98	0.02	1.00	1.00	1.00	1.00
8	0.02	1.00	1.00	1.00	1.00	1.00	1.00
9	0.06	1.00	1.00	1.00	1.00	1.00	1.00
10	0.00	0.94	0.95	0.39	1.00	1.00	1.00
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00	1.00	1.00	0.11
13	0.67	0.81	1.00	0.84	1.00	1.00	1.00
14	0.81	1.00	1.00	1.00	1.00	1.00	1.00
15	1.00	1.00	0.00	1.00	1.00	1.00	1.00
16	1.00	1.00	0.00	1.00	1.00	1.00	1.00

EM and MEM tests, using the same simulation scenarios considered in [22]. Observe that our tests based on FD2 and ABD are also better, as they have larger medians than the other tests. The short whiskers of the boxplots for the DD plot-based tests indicate that they are stable and do not vary significantly, while the EM and MEM tests have considerable variability. We conclude that the DD plot-based tests are more powerful than the approaches proposed in [22].

The boxplots for the empirical power of all scenarios considered in Table 1, with varying sample sizes, are displayed in Figs. 3–5. From these boxplots, note that the tests tend to be more powerful for larger sample sizes, as expected. Observe from these figures that, when both samples sizes are 25, the DD plot-based tests with ABD and FD2, in general, perform well, since that they have median powers near to one. When both sample sizes are 50 or 150, the DD plot-based tests with ABD and FD2 have large medians and are, overall, the most powerful. Nevertheless, the DD plot-based test, with *h*mode stated in (6), does not perform well so that it is not recommended.

5. Application with real chemical data

In this section, we consider four functional chemical data sets. They consist of two heterogeneous groups so that our tests should reject the null hypothesis that they are homogeneous.

The first data set consists of spectrometric curves for chopped pieces of meat, which correspond to the absorbance measured at 100 different

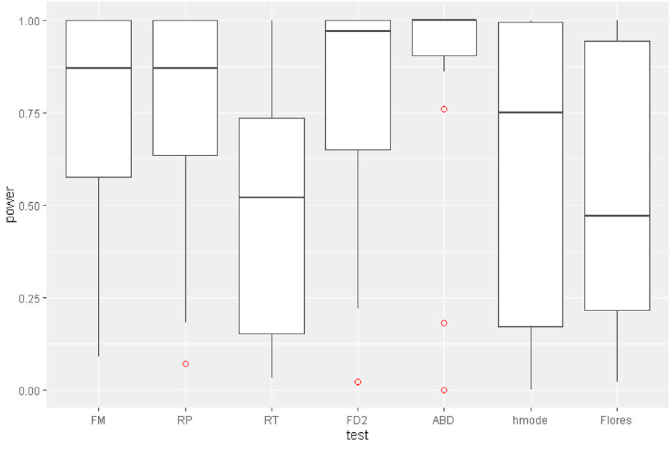

Fig. 3. Boxplots of the empirical powers based on the simulation scenario described in Table 1, when both samples have 25 curves. Points in red are outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

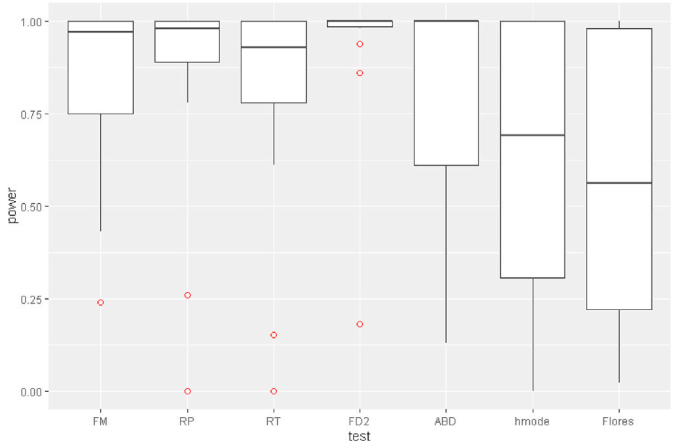

Fig. 4. Boxplots of the empirical powers based on the simulation scenario described in Table 1, when both samples have 50 curves. Points in red are outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

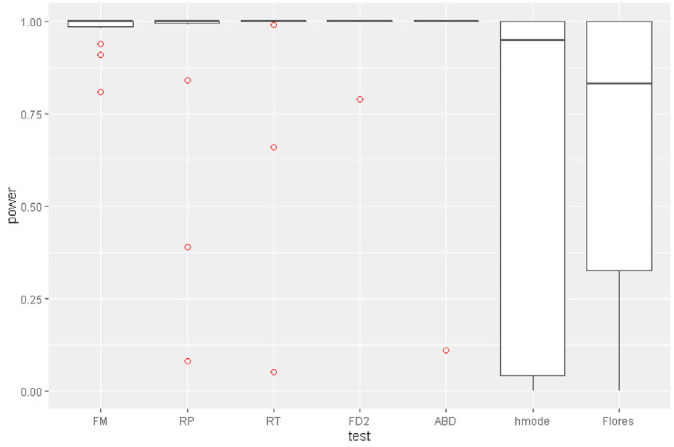

Fig. 5. Boxplots of the empirical powers based on the simulation scenario described in Table 1, when both samples have 150 curves. Points in red are outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

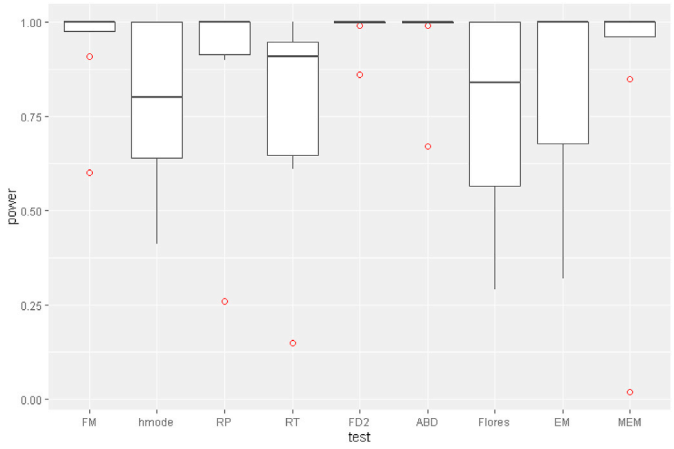

Fig. 6. Boxplots of the empirical powers of the DD plot with different depth measures compared to the EM and MEM tests for samples of 50 curves. Points in red are outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

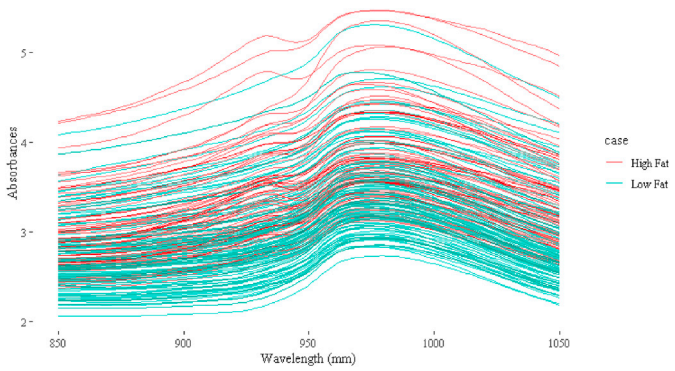

Fig. 7. Tecator spectrometry data with green for low-fat meat and red for high-fat meat. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

wavelengths [19]. We can divide these meats into two groups: the pieces with (i) small and (ii) large fat percentages (where small is less than 20% fat). The curves are in Fig. 7, whereas the DD plots using different depths of these samples are in Fig. 8. The points do not pass through the diagonal line.

The second data set corresponds to nitrogen oxide (NOx) emission levels in a control station in Poblenou, Barcelona, Spain [17]. The station measures NOx levels in $\mu g/m^3$ every hour, every day. The curves are split into two groups: (i) working days and (ii) non-working days, which are in Fig. 9, and their respective DD plots using various depths are in Fig. 10. Once again, we can see that the points do not concentrate toward the line.

The third data set is associated with mitochondrial calcium overload (MCO) [57]. During ischemic myocardia, high levels of MCO relate to better protection against ischemia. Then, it is interesting to see if some drugs can raise MCO levels in mouses. This data set consists once again of two groups: (i) one which receives no drug –control group– and (ii) another group that receives a drug –treated group– that can raise MCO levels. Every 10 s, MCO levels are measured. In Fig. 11, we can see both samples, whereas the DD plots of these samples are in Fig. 12. The points do not concentrate on the line neither.

The four set is related to Berkeley growth data, which contains the heights of 39 boys and 54 girls from ages 1 to 18. It is well known that growth dynamics differ from boys to girls, so that our test should reject the null hypothesis of homogeneity. We can see the curves in Fig. 13, whereas Fig. 14 shows the DD plots obtained from these two samples using different depths. Note the behavior typical of heterogeneous samples, that is, they are do not concentrate on the (0, 0)-(1, 1) diagonal line.

Performing a visual check for homogeneity using the DD plots is not satisfactory enough because we do not get essential metrics like the p-value. Then, we employ the tests proposed in this paper to obtain the results reported in Table 5. The FD2 DD plot-based test and the NEM test [22] were the only tests able to reject $\mathscr{H}_0$ in all the cases, getting near zero p-values. The Flores test is only able to reject $\mathscr{H}_0$ half the times. The Flores and FM DD plot-based tests are able to reject $\mathscr{H}_0$ in two cases, while not rejecting in one case. The RPD and ABD DD plot-based tests are able to reject three of the four cases correctly.

6. Conclusions, discussion and future research

In this article, we proposed a two-sample test for functional data utilizing DD plots. We adapted the idea proposed in [35] to link multivariate DD plots and homogeneity to a functional setting. Then, we formalized these notions into a test that considered a linear model for the DD plot, contrasted two linear hypotheses that relate to homogeneity between samples, used bootstraping-t to approximate the null distribution of the test statistic, and handled the multiple hypotheses employing the Holm-Bonferroni method. In summary, this paper reported the following findings:

(i) This proposal compared depth measures (or any other measure which takes a curve and returns a real number) between two samples, making it robust in many scenarios, being this the advantage of our method. Other tests, as one proposed in [20], compare only the most representative data between samples. Our tests also employed bootstrap-t procedures that are second-order accurate, better than the usual bootstrap confidence intervals, which are only first-order accurate.

(ii) We compared performance through simulation of our tests and other tests, with different sizes in both samples. Our test achieved a desirable empirical size, whereas the empirical power was greater than the corresponding power of other tests found in the literature. In particular, our proposed test had a good power when the departures from the null hypothesis are in variance/covariance structure, in shape, or asymmetry, while also achieving sufficient power when the difference between samples is in the mean magnitude.

(iii) In particular, the DD plot-based tests with FD2 and ABD tended to work well in all the different scenarios. However, the ABD DD plot-based test tended to be computationally heavy, so that we recommend to use it only for small sample sizes, where it had

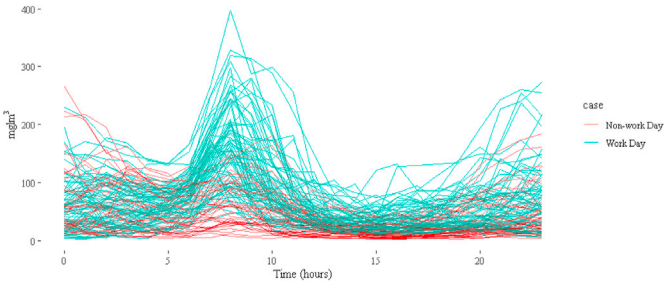

Fig. 9. Curves for NOx levels in Poblenou (Barcelona, Spain) with green for work and red for non-work days. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

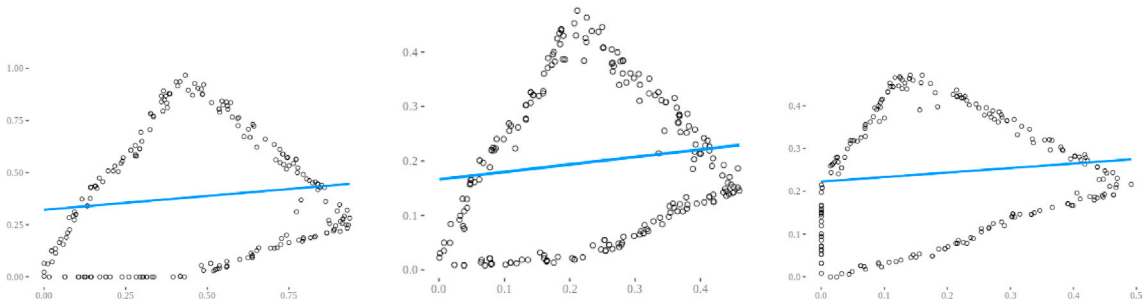

Fig. 8. DD plots for spectrometric curves of low and high fat containing meats using (left) FM, (center) RPD, and (right) FD2 depths.

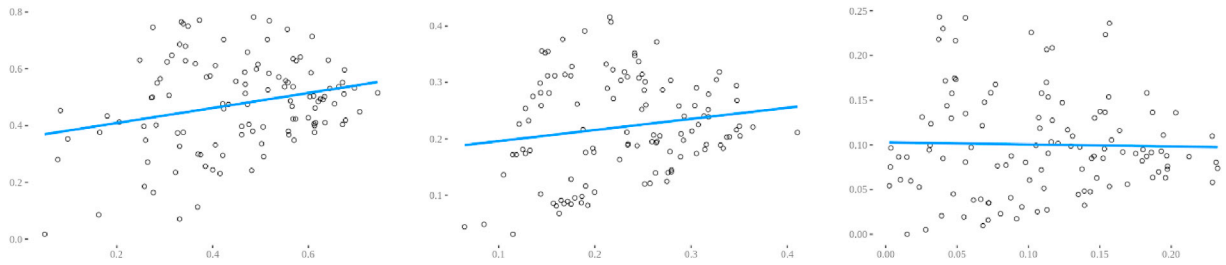

Fig. 10. DD plots for the NOx curves in Poblenou (Barcelona, Spain) for work and non-work days using (left) FM, (center) RPD, and (right) FD2 depths.

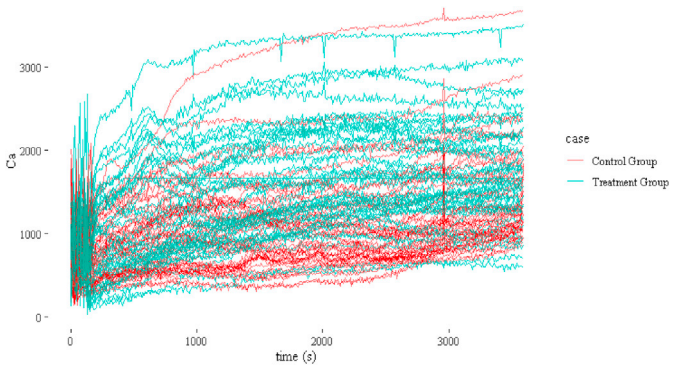

Fig. 11. Curves for MCO levels in mouses' cardiac cells for one curve as control group in red and the other in green receives a treatment. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

better power performance than the other tests. When we have medium or large samples, we should employ the FD2 DD plot-based test. This depth is quite powerful as it can detect differences not only on the samples of the curves themselves but also on their derivatives, as stated in [43]. The FD2 DD plot-based test worked well in a high variety of scenarios, having in general greater power than some other recents tests found in the literature. The hmode DD plot-based test did not perform well in different scenarios, so that its use is not recommended. We believe that this depth is not able to reflect all the different aspects of a sample, only concentrating in the mode.

(iv) The results of our tests in the four real chemical data sets analyzed were also satisfactory. In every set, the FD2 test could detect heterogeneity between real-world samples from different populations. Other recent tests could not do that, so this is another indicator that the FD2 DD plot-based test is powerful.

Thus, our study can be a knowledge addition to the tool-kit of diverse practitioners, including chemical engineers, chemists, applied statisticians, and data scientists.

The source of the performance difference between the Flores test and our test can be explained by the fact that the Flores test uses only the depths of the deepest curves in the samples, while we employ the depths of all the curves in the sample.

Some themes for future research, which arose from the present investigation, are the following:

(i) We can change the bootstrapping-t for another resampling method.
(ii) We could also propose a nonparametric version of the F-statistic utilized for multiple lineal hypothesis testing.
(iii) We might implement different depth measures that were not considered here, like in [44], to assess their empirical power.
(iv) We could consider other simulation scenarios. For example, can our test detect heterogeneity when the samples have the same means and covariance operators but have different kurtosis? Some kurtosis measures for FDA were proposed in [64]. We can create new simulation scenarios based on them.
(v) The test proposed in this work might be extended to more than two populations, as well as to design a robust version of this test by implementing in equation (8) the technique studied in [63].
(vi) The proposed test does not consider how the points of a DD plot concentrate on the diagonal line but whether they deviate or not from such a line. To contrast how concentrated this plot is around the diagonal line, it is indeed an ideal test statistic for homogeneity. Until now, we do not know how to improve our test in this sense, but an idea that could be helpful is based on the connection among goodness-of-fit tests and graphical methods as proposed in [5,6].

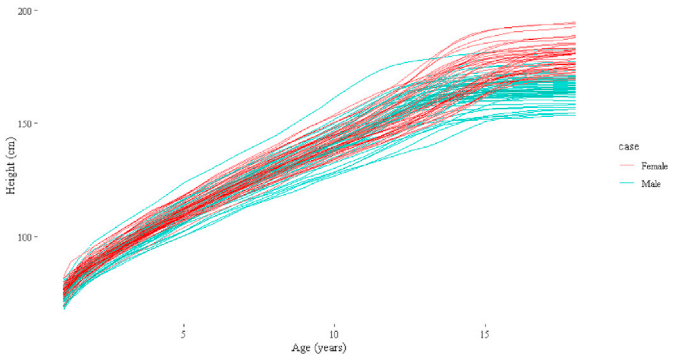

Fig. 13. Berkeley growth data with green for males and red for females. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

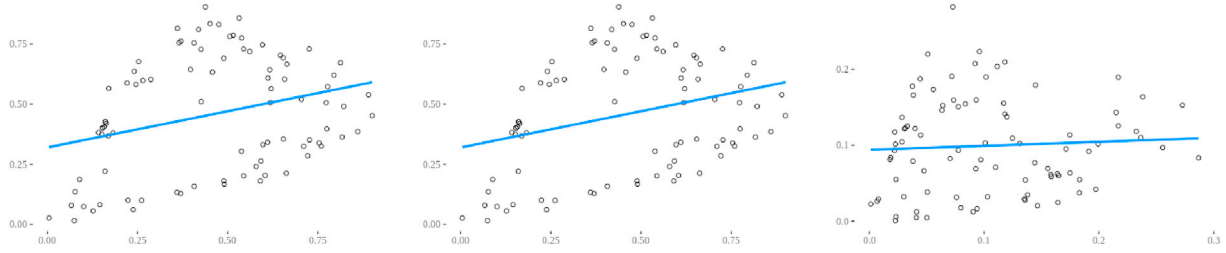

Fig. 12. DD plots for the MCO curves in the mouse's cardiac cells for control and treated groups using (left) FM, (center) RPD, and (right) FD2 depths.

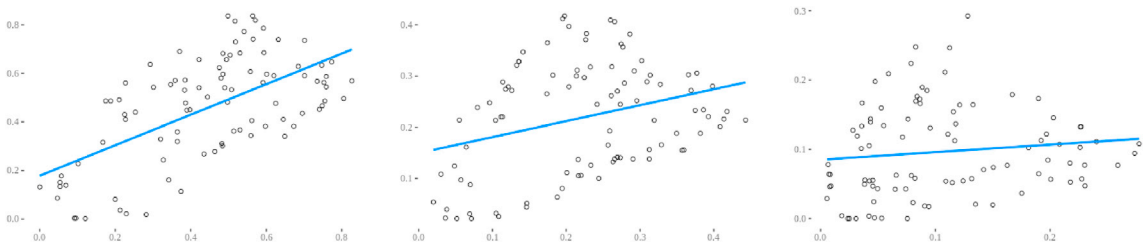

Fig. 14. DD plots for the height curves of boys and girls using (left) FM, (center) RPD, and (right) FD2 depths.

Table 5
Result of the indicated test in the listed data set, where ✓means that the test was able to reject $\mathscr{H}_0$ and × means that it does not reject $\mathscr{H}_0$ at 5% of significance.

Data set	Flores	FM	p-value	RPD	p-value	FD2	p-value	NEM	p-value	ABD	p-value
Tecator	✓	✓	0.034	✓	0.046	✓	< 0.001	✓	0.006	×	0.432
MCO	✓	×	0.072	✓	0.017	✓	< 0.001	✓	0.009	✓	< 0.001
NOx	×	✓	0.032	✓	0.028	✓	< 0.001	✓	< 0.001	✓	< 0.001
Heights	×	×	0.324	×	0.268	✓	< 0.001	✓	< 0.001	×	0.236

(vii) The statistic proposed in the present study assesses whether two samples differ in some characteristics of the distributions that are generating the functional data and not just in one specific characteristic. The deficiency of the statistics proposed in the literature till the date against our statistic is that the proposed statistic considers any kind of heterogeneity between the two samples used for testing, while other statistics consider a more specific heterogeneity (in location or in scale). Our statistic, such as it was proposed, is not designed to identify differences between any specific parameter of the two functional distributions to be tested but whether both functional samples come from the same functional distribution or not. Therefore, the proposed statistic does not identify what parameters produce this difference. While a general purpose test is helpful in a broad range of settings, once the hypothesis of homogeneity is rejected, it is natural to ask us in what way the two functional distributions differ. The shape of a DD plot could have some information to offer regarding the nature of the difference. This is the reason for considering various rationales in the existing DD plot-based test statistics in a multivariate setting. Therefore, identifying in what functional parameter(s) the distributions are differing based on the test proposed here is an interesting aspect to be explored in a future investigation.

The proposed tests in this study promote new challenges and offer open issues to be analyzed. Research on these and other issues are in progress and their findings will be reported in future articles.

Author statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript:

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Editors and Reviewers for their constructive comments on an earlier version of this manuscript which resulted in this improved version. The research of A. Calle-Saldarriaga, H. Laniado, and F. Zuluaga was funded by the Internal Project "*Statistical homogeneity test for infinite dimensional data*" with number 881-000044 from Universidad EAFIT, Medellín, Colombia. The research of V. Leiva was partially funded by FONDECYT, project grant number 1200525 from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science and Technology, Knowledge and Innovation.

References

[1] I.M. Almanjahie, M.K. Attouch, O. Fetitah, H. Louhab, Robust kernel regression estimator of the scale parameter for functional ergodic data with applications, Chilean Journal of Statistics 11 (2020) 73–94.
[2] R.G. Aykroyd, V. Leiva, F. Ruggeri, Recent developments of control charts, identification of big data sources and future trends of current research, Technol. Forecast. Soc. Change 144 (2019) 221–232.
[3] R. Bárcenas, K. Ortega, A.J. Quiroz, Quadratic forms of the empirical processes for the two-sample problem for functional data, Test 26 (2017) 503–526.
[4] R. Burfield, C. Neumann, C.P. Saunders, Review and application of functional data analysis to chemical data. The example of the comparison, classification, and database search of forensic ink chromatograms, Chemometr. Intell. Lab. Syst. 149 (2015) 97–106.
[5] C. Castro-Kuriss, D. Kelmansky, V. Leiva, E. Martinez, On a goodness-of-fit test for normality with unknown parameters and type-II censored data, J. Appl. Stat. 37 (2010) 1193–1211.
[6] C. Castro-Kuriss, M. Huerta, V. Leiva, A. Tapia, On some goodness-of-fit tests and their connection to graphical methods with uncensored and censored data, in: J. Xu, S.E. Ahmed, G. Duca, F.L. Cooke (Eds.), Management Science and Engineering Management, Springer-Verlag, Berlin, Germany, 2020, pp. 157–183.
[7] S. Chenouri, C.G. Small, A nonparametric multivariate multisample test based on data depth, Electron. J. Stat. 6 (2012) 760–782.
[8] J. Cuesta-Albertos, A. Nieto-Reyes, The random Tukey depth, Comput. Stat. Data Anal. 52 (2008) 4979–4988.
[9] J.A. Cuesta-Albertos, M. Febrero-Bande, M. Oviedo de la Fuente, The DDG-classifier in the functional setting, Test 26 (2017) 119–142.
[10] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Stat. Plann. Inference 147 (2014) 1–23.
[11] A. Cuevas, M. Febrero, R. Fraiman, An anova test for functional data, Comput. Stat. Data Anal. 47 (2004) 111–122.
[12] A. Cuevas, M. Febrero, R. Fraiman, On the use of the bootstrap for estimating functions with functional data, Comput. Stat. Data Anal. 51 (2006) 1063–1074.
[13] A. Cuevas, M. Febrero, R. Fraiman, Robust estimation and classification for functional data via projection-based depth notions, Comput. Stat. 22 (2007) 481–496.
[14] P. Delicado, R. Giraldo, C. Comas, J. Mateu, Statistics for spatial functional data: some recent contributions, Environmetrics 21 (2009) 224–239.
[15] T.J. DiCiccio, B. Efron, Bootstrap confidence intervals, Stat. Sci. 11 (1996) 189–228.
[16] J. Fan, S.K. Lin, Test of significance when data are curves, J. Am. Stat. Assoc. 93 (1998) 1007–1021.
[17] M. Febrero, P. Galeano, W. González-Manteiga, Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels, Environmetrics 19 (2008) 331–345.
[18] M. Febrero-Bande, M. Oviedo de la Fuente, Statistical computing in functional data analysis: the R package fda.usc, J. Stat. Software 51 (2012) 1–28.
[19] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, New York, US, 2006.
[20] R. Flores, R. Lillo, J. Romo, Homogeneity test for functional data, J. Appl. Stat. 45 (2018) 868–883.

[21] R. Fraiman, G. Muniz, Trimmed means for functional data, Test 10 (2001) 419–440.
[22] A. Franco-Pereira, R. Lillo, Rank tests for functional data based on the epigraph, the hypograph and associated graphical representations, Adv. Data Anal. Classif. 14 (2020) 651–676.
[23] S. Fremdt, J.G. Steinebach, L. Horváth, P. Kokozka, Testing the equality of covariance operators in functional samples, Scand. J. Stat. 40 (2013) 138–152.
[24] F. Garcia-Papani, V. Leiva, M.A. Uribe-Opazo, R.G. Aykroyd, Birnbaum-Saunders spatial regression models: diagnostics and application to chemical data, Chemometr. Intell. Lab. Syst. 177 (2018) 114–128.
[25] R. Giraldo, P. Delicado, J. Mateu, Ordinary kriging for function-valued spatial data, Environ. Ecol. Stat. 18 (2011) 411–426.
[26] R. Giraldo, L. Herrera, V. Leiva, Cokriging prediction using as secondary variable a functional random field with application in environmental pollution, Mathematics 8 (2020) 1305.
[27] U. Grenander, Stochastic processes and statistical inference, Ark. Mater. 1 (1950) 195–277.
[28] P. Hall, I.V. Keilegom, Two-sample tests in functional data analysis starting from discrete data, Stat. Sin. 7 (2007) 1511–1531.
[29] P. Hall, N. Tajvidi, Permutation tests for equality of distributions in high-dimensional settings, Biometrika 89 (2002) 359–374.
[30] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65–70.
[31] Q. Jiang, M. Husková, S.G. Meintanis, L. Zhu, Asymptotics, finite-sample comparisons and applications for two-sample tests with functional data, J. Multivariate Anal. 170 (2019) 202–220.
[32] D. Kraus, V.M. Panaretos, Dispersion operators and resistant second-order functional data analysis, Biometrika 99 (2012) 813–832.
[33] J. Li, R.Y. Liu, New nonparametric tests of multivariate locations and scales using data depth, Stat. Sci. 19 (2004) 686–696.
[34] R.Y. Liu, On a notion of data depth based on random simplices, Ann. Stat. 18 (1990) 405–414.
[35] R.Y. Liu, J.M. Parelius, K. Singh, Multivariate analysis by data depth: descriptive statistics, graphics and inference, Ann. Stat. 27 (1999) 783–840.
[36] A. Lung-Yut-Fong, C. Lévy-Leduc, O. Cappé, Homogeneity and change-point detection tests for multivariate data using rank statistics, J. Soc. Fr. Stat. 156 (2015) 133–162.
[37] S. López-Pintado, J. Romo, On the concept of depth for functional data, J. Am. Stat. Assoc. 104 (2009) 718–734.
[38] S. López-Pintado, J. Romo, A half-region depth for functional data, Comput. Stat. Data Anal. 55 (2011) 1679–1695.
[39] C. Martin-Barreiro, J.A. Ramirez-Figueroa, X. Cabezas, V. Leiva, M.P. Galindo-Villardon, Disjoint and functional principal component analysis for infected cases and deaths due to COVID-19 in South American countries with sensor-related data, Sensors 21 (2021) 4094.
[40] S. Martinez, R. Giraldo, V. Leiva, Birnbaum-Saunders functional regression models for spatial data, Stoch. Environ. Res. Risk Assess. 33 (2019) 1765–1780.
[41] A. Munk, R. Paige, J. Pang, V. Patrangenaru, F. Ruymgaart, The one- and multi-sample problem for functional data with application to projective shape analysis, J. Multivariate Anal. 99 (2008) 815–833.
[42] S. Nagy, An overview of consistency results for depth functionals, in: G. Aneiros, E.G. Bongiorno, R. Cao, P. Vieu (Eds.), Functional Statistics and Related Fields, Springer, Cham, Switzerland, 2017, pp. 189–196.
[43] S. Nagy, I. Gijbels, D. Hlubinka, Depth-based recognition of shape outlying functions, J. Comput. Graph Stat. 26 (2017) 883–893.
[44] N.N. Narisetty, V.N. Nair, Extremal depth for functional data and applications, J. Am. Stat. Assoc. 111 (2016) 1705–1714.
[45] V.M. Panaretos, D. Kraus, J.H. Maddocks, Second-order comparison of Gaussian random functions and the geometry of DNA minicircles, J. Am. Stat. Assoc. 105 (2010) 670–682.
[46] S.D. Pawar, D.T. Shirke, Nonparametric tests for multivariate locations based on data depth, Commun. Stat. Simulat. Comput. 48 (2019) 753–776.
[47] A. Pini, A. Stamm, S. Vantini, Hotelling's T2 in separable Hilbert spaces, J. Multivariate Anal. 167 (2018) 284–305.
[48] O. Pokotylo, P. Mozharovskyi, R. Dyckerhoff, Depth and depth-based classification with R package ddalpha, J. Stat. Software 91 (2019) 1–46.
[49] G.M. Pomann, A.M. Staicu, S. Ghosh, A two sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis, J. Roy. Stat. Soc. C 65 (2016) 395–414.
[50] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.
[51] J.A. Ramirez-Figueroa, C. Martin-Barreiro, A.B. Nieto, V. Leiva, M.P. Galindo-Villardon, A new principal component analysis by particle swarm optimization with an environmental application for data science, Stoch. Environ. Res. Risk Assess. 35 (2021) 969–1984.
[52] J. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, US, 2005.
[53] J.O. Ramsay, When the data are functions, Psychometrika 47 (1982) 379–396.
[54] J.O. Ramsay, H. Wickham, S. Graves, G. Hooker, Fda: Functional Data Analysis, 2018. R package version 2.4.8.
[55] C.R. Rao, Some statistical methods for comparison of growth curves, Biometrics 14 (1958) 1–17.
[57] M. Ruiz-Meana, D. Garcia-Dorado, P. Pina, J. Inserte, L. Agulló, J. Soler-Soler, Cariporide preserves mitochondrial proton gradient and delays ATP depletion in cardiomyocytes during ischemic conditions, Am. J. Physiol. Heart Circ. Physiol. 285 (2003) H999–H1006.
[58] F. Scheipl, J. Goldsmith, J. Wrobel, Tidyfun: Tools for Tidy Functional Data, 2020. R package version 0.0.82.
[59] Y. Sun, M.G. Genton, Functional boxplots, J. Comput. Graph Stat. 20 (2011) 316–334.
[60] G.J. Szkély, E-Statistics: the Energy of Statistical Samples, Technical Report., Bowling Green State University, Bowling Green, Ohio, US, 2002.
[61] J.W. Tukey, Mathematics and the picturing of data, in: Proceedings of the International Congress of Mathematicians, vol. 2, Montréal, Québec, Canada, 1975, pp. 523–531.
[62] D. Valencia, R. Lillo, J. Romo, A Kendall correlation coefficient between functional data, Adv. Data Anal. Classif. 13 (2019) 1083–1103.
[63] H. Velasco, H. Laniado, M. Toro, V. Leiva, Y. Lio, Robust three-step regression based on comedian and its performance in cell-wise and case-wise outliers, Mathematics 8 (2020) 1259.
[64] S. Walter, Defining Quantiles for Functional Data with an Application to the Reversal of Stock Price Decreases, Ph.D. thesis, The University of Melbourne, Australia, 2011.
[65] J.L. Wang, J.M. Chiou, H.G. Muller, Functional data analysis, Annu. Rev. Stat. Appl. 3 (2016) 257–295.
[66] K. Wang, A. Ng, G. McLachlan, EMMIXskew: the EM Algorithm and Skew Mixture Distribution, 2018. R package version 1.0.3.
[67] K. Wang, A. Ng, G. McLachlan, Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data, in: Proceedings of the Digital Image Computing: Techniques and Applications, 2009.
[68] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer, New York, US, 2016.
[69] G. Wynne, A.B. Duncan, A Kernel Two-Sample Test for Functional Data, 2020. https://arxiv.org/abs/2008.11095.