DOI: 10.1111/biom.12476

The Interval Testing Procedure: A General Framework for Inference in Functional Data Analysis

Alessia Pini* and Simone Vantini**

MOX-Department of Mathematics, Politecnico di Milano, Via Bonardi 9, 20133 Milano, Italy
**email:* alessia.pini@polimi.it
***email:* simone.vantini@polimi.it

Summary. We introduce in this work the Interval Testing Procedure (ITP), a novel inferential technique for functional data. The procedure can be used to test different functional hypotheses, e.g., distributional equality between two or more functional populations, equality of mean function of a functional population to a reference. ITP involves three steps: (i) the representation of data on a (possibly high-dimensional) functional basis; (ii) the test of each possible set of consecutive basis coefficients; (iii) the computation of the adjusted p-values associated to each basis component, by means of a new strategy here proposed. We define a new type of error control, the interval-wise control of the family wise error rate, particularly suited for functional data. We show that ITP is provided with such a control. A simulation study comparing ITP with other testing procedures is reported. ITP is then applied to the analysis of hemodynamical features involved with cerebral aneurysm pathology. ITP is implemented in the `fdatest` R package.

Key words: Family wise error rate; Functional data analysis; Inference; Multiple comparison; Permutation method.

1. Introduction

"Are these two groups of curves statistically different?" "If yes, which are the differences?" "What is the probability that these differences came into view by chance?" Such kinds of questions are becoming more and more urging in many research areas, due to the fast development of precise acquisition devices. Despite the recent breakthrough of functional data analysis (FDA) as a method for analyzing data sets made of curves (Ramsay and Silverman, 2002, 2005; Ferraty and Vieu, 2006), and the development of many statistical tools to answer questions similar to the former one, very few have been addressed to answer questions similar to the latter ones. In this work, we develop a new inferential procedure—in the framework of FDA—not only able to assess the equality in distribution between functional populations, but also to point out specific differences, controlling at the same time the probability of false discoveries.

Several methods dealing with inferential problems for functional data have been object of statistical investigation. The most common approach is "global testing", according to which a unique global test is performed and a unique p-value is provided. Examples of such techniques are usually based on asymptotic results and/or on strong modeling assumptions on data distribution (e.g., Fan and Lin, 1998; Spitzner et al., 2003; Cuevas et al., 2004; Shen and Faraway, 2004; Abramovich and Angelini, 2006; Antoniadis and Sapatinas, 2007; Schott, 2007; Horváth and Kokoszka, 2012; Staicu et al. 2014; Zhang and Liang, 2014). Other examples have been proposed in the framework of permutation tests, not relying on strong distributional assumptions (e.g., Cardot et al., 2007; Hall and Tajvidi, 2002). The commonality among all these procedures is that they are meant to state whether there is evidence to reject the assumption of equality in distribution; however, as they cannot impute the rejection to specific features of the data, they may not be suitable for some applications.

Although being theoretically infinite-dimensional objects, in practice, statisticians deal with functional data by projecting them on a finite dimensional space spanned by a suitable truncated basis (Ramsay and Silverman, 2005). Following this line, we here suggest to base inference directly on the set of coefficients representing data. In such a perspective, the functional test can be replaced with a family of tests pertaining the components of the basis expansion. A selection of the significant basis components may provide insights on the reason of the rejection; in particular, for the specific case of a functional basis located in space (such as B-splines) the selection turns out to be a selection of intervals of the domain presenting significant differences between the populations of curves.

A natural approach for providing such a selection might be to perform a test on each component, and then to correct the test results in order to control the probability of wrongly rejecting each possible set of true null hypotheses (strong control of the Family Wise Error Rate, FWER). The correction could be made, for instance, using Bonferroni–Holm procedure (Holm, 1979), and an example of the application of this technique to FDA can be found in Spitzner et al. (2003). The resulting procedures enable the selection of a subset of significant components; nevertheless, they are generally not suited for cases where the number of basis components is large: the computational cost of might explode and/or the power can become very low, and this is often the case in FDA.

The Interval Testing Procedure (ITP), that we present in this work, addresses this issue: similarly to Bonferroni-like inferential techniques, it is able, in case of rejection, to highlight which components have led to the rejection itself. Furthermore, its power remains comparable with the one provided by global inference techniques even when the number of components is very large. However, since "there is no such thing as a free lunch," ITP lacks the strong control of the FWER: indeed, we will prove that it is just provided with an "interval-wise" control of the FWER. This control is stronger than the weak control provided by global tests, but it is weaker than the strong control provided by component-wise procedures. In the FDA framework, this is a minor drawback since such kind of control might be sufficient in practice. For example, as we will show in the following sections, when testing the difference between two functional populations relying on the B-spline representation, the "interval-wise" control implies the control of the FWER on intervals of the domain. This means that the probability of wrongly detecting a significant difference between the two populations in any interval where there is none is controlled at the desired level.

The selection of significant intervals of the domain with respect to a given null hypothesis has been the topic of a recent study by Vsevolozhskaya et al. (2014). Differently from our current proposal (based on function discretization), they suggest an approach based on domain discretization. In detail, authors develop a procedure based on an a priori chosen partition of the domain in subintervals. Those subintervals are tested, and the corresponding p-values are adjusted in order to control the probability of wrongly rejecting any set of subintervals (i.e., strong control of the FWER between intervals). Within each interval, however, the procedure is only provided with a weak control of the FWER: the probability of wrongly rejecting the subinterval is controlled only when there is no difference between the two samples in the whole interval. Clearly, conclusions of such a test are dependent on the initial choice of the partition.

This article is outlined as follows: in Section 2 we describe the ITP for the two population test, and discuss its theoretical properties. In Section 3, we show how to extend ITP in other frameworks (e.g., test of the centrality parameter of a functional population). In Section 4, we present the results of a simulation study comparing the performances of ITP with other component-wise techniques based on the Bonferroni–Holm or Benjamini–Hockberg (Benjamini and Hochberg, 1995) corrections, and with the test proposed by Vsevolozhskaya et al. (2014). In Section 5, ITP is applied to a case study devoted to the analysis of the Aneurisk data set (Sangalli et al., 2009) concerning the comparison between geometric features of the internal carotid artery in two groups of patients associated to different levels of severity of the cerebral aneurysm pathology. Proofs of all Theorems are reported in the Appendix (available online). The R-package `fdatest` implementing the ITP is available on CRAN (Pini and Vantini, 2015). All computations and images presented in this article have been created using R (R Core Team, 2015).

2. ITP in the Two-Population Framework

2.1. *ITP Algorithm*

Let us suppose to observe two independent samples of sizes n_1 and n_2 of independent random functions on a separable Hilbert space. We aim at testing the null hypothesis of equality in distribution of the two populations which the two samples have been drawn from.

The testing procedure we present is composed of the following steps: (i) **Basis Expansion**: functional data are projected on a functional basis; (ii) **Interval-Wise Testing**: statistical tests are performed on each interval of basis coefficients; (iii) **Multiple Correction**: for each component of the basis expansion, an adjusted p-value is computed from the p-values of the tests performed in the previous step.

First step: basis expansion

Theoretically, any function can be uniquely represented through a countable sequence of coefficients associated to a basis of the functional space (i.e., B-splines, Fourier harmonics, and so on). In practice, however, very rarely functional data come with an analytic expression; more often, just some point-wise evaluations of a function are available: thus, just a limited number of components can be estimated. We need to represent data by means of a finite-dimensional representation $y_{ij}(t)$ obtained through an expansion on a reduced basis $\{\phi^{(k)}\}_{k=1,\dots,p} : y_{ij}(t) = \sum_{k=1}^{p} c_{ij}^{(k)} \phi^{(k)}(t)$, where i is the unit index, j the population index, and k the basis component index. This projection constitutes the first step in most FDA procedures. Integer p represents the finite dimension of the functional space in which data are represented. Even though ITP can deal with any functional basis and with any dimension p independently on the sample size, only the adoption of functional basis located in space (such as B-splines) allows a selection of significant intervals of the domain. The choice of (i) the functional basis, (ii) the truncation p, and (iii) the method used to estimate the coefficients is widely discussed in FDA literature. We refer to Ramsay and Silverman (2005) for an overview of this topic.

Eventually, we can represent each of the $n = n_1 + n_2$ units by means of the corresponding p coefficients $\{c_{ij}^{(k)}\}_{k=1,\dots,p}$, $i = 1, \dots, n_j$, $j = 1, 2$. The assumptions made for the functional populations can be restated in terms of coefficients: we have that for each k, $c_{11}^{(k)}, \dots, c_{n_1 1}^{(k)}, c_{12}^{(k)}, \dots, c_{n_2 2}^{(k)}$ are independent, and $c_{11}^{(k)}, \dots, c_{n_1 1}^{(k)} \sim C_1^{(k)}$, $c_{12}^{(k)}, \dots, c_{n_2 2}^{(k)} \sim C_2^{(k)}$, where $C_1^{(k)}$ and $C_2^{(k)}$ denote the (unknown) distributions of the kth basis coefficient in the two populations. It is worth noticing that we do not assume either independence between basis coefficients pertaining to different components, or joint or marginal normality of basis coefficients, or orthogonality of the basis. We just assume independence between sample units.

Second step: interval-wise testing

In the second step of ITP, each basis component k is marginally tested ($H_0^{(k)} : C_1^{(k)} \overset{\mathrm{d}}{=} C_2^{(k)}$); then, a bivariate test is performed on each couple of successive basis components ($H_0^{(k,k+1)} : H_0^{(k)} \cap H_0^{(k+1)}$); then, a three-variate test is performed on each triple of successive basis components ($H_0^{(k,k+1,k+2)} : H_0^{(k)} \cap H_0^{(k+1)} \cap H_0^{(k+2)}$), up to the global p-variate test ($H_0^{(1,\dots,p)} : H_0^{(k)} \bigcap_{k=1}^{p} H_0^{(k)}$). We obtain a family of tests with their associated p-values (e.g., Figure 1a). We

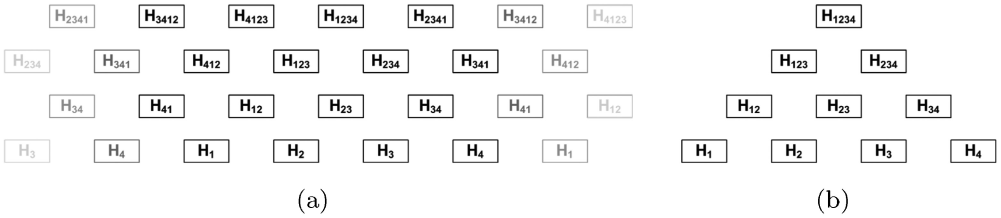

H2341 H3412 H4123 H1234 H2341 H3412 H4123
H234 H341 H412 H123 H234 H341 H412
H34 H41 H12 H23 H34 H41 H12
H3 H4 H1 H2 H3 H4 H1
(a)
H1234
H123 H234
H12 H23 H34
H1 H2 H3 H4
(b)

Figure 1. Example (with $p = 4$) of the family of multivariate tests explored by the ITP (on the left). Version focusing only on intervals, excluding the complementary sets (on the right).

denote with $\lambda^{(\mathbf{k})}$ the p-value of the multivariate test for $H_0^{(\mathbf{k})} = \bigcap_{k \in \mathbf{k}} H_0^{(k)}$ (where $\mathbf{k}$ is a vector of successive indexes in $\{1, ..., p\}$). In addition to all possible tests on intervals, we add the multivariate tests on the complementary sets of each interval, i.e., we do also test each hypothesis $H_0^{(\mathbf{k}^c)} = \bigcap_{k \notin \mathbf{k}} H_0^{(k)}$ (as shown in Figure 1a). The advantage of considering the complementary sets of intervals pertains to the multiple correction phase, as it will be discussed in detail in the next paragraph.

Interval-wise tests can be performed in different ways, depending on the sample size and on the assumptions on the distributions of $C_j^{(k)}$. The best-case scenario is the one where coefficients are jointly normal and $n > p$. In this case, we can use Hotelling's T^2 tests. In a more realistic scenario, however, we cannot assess the normality of coefficients and $n \leq p$. A possible approach to deal with this issue is to exploit the Nonparametric Combination Procedure (NPC), presented by Pesarin and Salmaso (2010). This procedure constructs multivariate permutation tests by means of combining univariate-synchronized permutation tests. Resulting tests are exact for any n and p, even in presence of dependence among components.

Third step: multiple correction

We obtain the adjusted p-value for the kth component $\lambda_{\text{ITP}}^{(k)}$ by computing the maximum over all p-values of interval-wise tests whose null hypothesis implies $H_0^{(k)}$:

$$\lambda_{\text{ITP}}^{(k)} = \max\left(\max_{\mathbf{k}\text{s.t.}\mathbf{k} \ni k} \lambda^{(\mathbf{k})}, \max_{\mathbf{k}^c\text{s.t.}\mathbf{k}^c \ni k} \lambda^{(\mathbf{k}^c)} \right).$$

In the next section we will prove that, if we reject $H_0^{(k)}$ when the kth adjusted p-value $\lambda_{\text{ITP}}^{(k)} \leq \alpha$, then, for any interval $\mathbf{k}$ s.t. $H_0^{(k)}$ is true $\forall k \in \mathbf{k}$, the probability of rejecting any $H_0^{(k)}$ is lower or equal to α. This property reads interval-wise control of the FWER.

It is worth noticing that in a general multivariate setting where the order of the variables is arbitrary, such a control has poor interest. On the contrary, in a setting where the component k has a natural order (like the one used in this article), this control becomes of immediate interest from the practitioners' point of view. For instance, when using B-splines the control holds on any subinterval of the domain, with edges in correspondence of knots. We thus can control the probability of wrongly detecting subintervals of the domain.

As a final remark, let us note that we might have based the procedure only on the tests on intervals of components, excluding the ones on complementary sets, as shown in Figure 1b. Nevertheless, according to this combination strategy the hypotheses "in the middle" would be tested more times than the ones "at the edges"; for instance, in the example of Figure 1b with $p = 4$, $H_0^{(2)}$ and $H_0^{(3)}$ are included in six tests, whereas $H_0^{(1)}$ and $H_0^{(4)}$ are only tested four times. This asymmetry may favor the rejection of the hypotheses at the edges, since they are tested fewer times. For this reason, we introduced the tests on complementary intervals. The resulting procedure has two major advantages: (i) each components is tested the same number of times $(p(p+1)/2)$, and (ii) the FWER is controlled also on complementary sets.

2.2. *Theoretical Properties of ITP*

We present here the theoretical results regarding the control of FWER and the power of ITP. All proofs of the theorems are reported in the online Appendix. In order not to overload the notation, throughout this theoretical section we will indicate with "intervals" both intervals and complementary sets of intervals. We begin by formally defining the interval-wise control of the FWER.

Definition 1. *Given a p-dimensional expansion of a functional data set, an inferential procedure is provided with an interval-wise control of the FWER if, for any interval* $\mathbf{k}$ *of components and* $\forall \alpha \in [0, 1]$, *the probability of rejecting at least one of the null hypotheses pertaining the components of* $\mathbf{K}$ *is less than* α, *when all these hypotheses are true:*

$$\forall \text{ interval } \mathbf{k} \subseteq \{1, ..., p\}: \quad \mathbb{P}_{H_0^{(\mathbf{k})}\text{true}}\left[\exists k \in \mathbf{k} : H_0^{(k)} \text{is rejected}\right] \leq \alpha.$$

The following result characterizes the control of the FWER provided by ITP.

Theorem 1. *ITP based on the p components of any basis expansion is provided with an interval-wise control of the FWER:*

$$\forall \text{ interval } \mathbf{k} \subseteq \{1, \ldots, p\}: \quad \mathbb{P}_{H_0^{(\mathbf{k})}\text{true}}\left[\exists k \in \mathbf{k} : \lambda_{\text{ITP}}^{(k)} < \alpha\right] \leq \alpha.$$

In other words, interval-wise control of the FWER means that, given any interval of components associated to true null hypotheses, the probability that at least one of the null hypotheses associated to the interval is wrongly rejected is controlled. This kind of control guarantees—among the others—

the control on the entire set of components and on single components, as extreme cases of intervals. We have indeed the following.

Corollary 1. *ITP based on the p components of any basis expansion is provided with a weak control of the FWER, i.e., the probability of rejecting at least one null hypothesis when all null hypotheses are true is controlled:*

$$\mathbb{P}_{H_0^{(\{1,\ldots,p\})}\,true}\left[\exists k \in \{1,\ldots,p\} : \lambda_{ITP}^{(k)} < \alpha\right] \leq \alpha.$$

Corollary 2. *The ITP based on the p components of any basis expansion is provided with a control of the Comparison-Wise Error Rate, i.e., for each component the probability that the null hypothesis pertaining the component is rejected when true is controlled:*

$$\forall k \in \{1,\ldots,p\} : \quad \mathbb{P}_{H_0^{(k)}\,true}\left[\lambda_{ITP}^{(k)} < \alpha\right] \leq \alpha.$$

In order to characterize further the inferential properties of ITP, it is useful to introduce two other testing procedures, derived from the adoption of different families of multivariate tests in the second step of the procedure described in Subsection 2.1. The Global Testing Procedure (GTP), which is associated to a degenerate family made of the global test only, and the Closed Testing Procedure (CTP), which is associated to the family made of all $2^p - 1$ possible multivariate tests. It is worth noticing that CTP, even though theoretically sound, becomes unfeasible in practice when the dimension p is high, due to the high number of tests it is based on. The number of tests indeed grows exponentially in p for CTP, quadratically for ITP, and is equal to one for GTP. GTP provides a weak control of the FWER, while CTP provides a strong control of the FWER. The following results depict a comparison between ITP, GTP, and CTP in terms of power and error rate.

Theorem 2: Global properties. *Let us consider a CTP, an ITP, and a GTP based on the p components of any basis expansion. The actual global levels of the CTP, of the ITP, and of the GTP (i.e., the probability of rejecting at least one $H_0^{(k)}$ when all null hypotheses are true) satisfy:*

$$\alpha_{CTP} \leq \alpha_{ITP} \leq \alpha_{GTP} = \alpha \ .$$

The powers of the CTP, of the ITP, and of the GTP (i.e., the probability of rejecting at least one $H_0^{(k)}$ when at least one of the null hypotheses is false) satisfy:

$$\pi_{CTP} \leq \pi_{ITP} \leq \pi_{GTP} \ .$$

Theorem 3: Component-wise properties. *Let us consider a CTP, an ITP, and a GTP based on the p components of any basis expansion. The Comparison-Wise Error Rates of the CTP and of the ITP on each component (i.e., the probability of rejecting $H_0^{(k)}$ when true) satisfy:*

$$CWER_{CTP}^{(k)} \leq CWER_{ITP}^{(k)} \leq \alpha \ .$$

The component-wise powers of the CTP and of the ITP on each component (i.e., the probability of rejecting $H_0^{(k)}$ when false) and the power of the GTP satisfy:

$$\pi_{CTP}^{(k)} \leq \pi_{ITP}^{(k)} \leq \pi_{GTP} \ ,$$

where π_{GTP} is the power of the global test.

Previous theorems explicit the tradeoff between the control of the FWER and the power both globally (Theorem 2) and component-wise (Theorem 3). Indeed the weaker control of the FWER of ITP with respect to CTP is counterbalanced by the fact that ITP is less conservative and more powerful (globally and component-wise) than CTP. On the contrary, the stronger control of the FWER of ITP with respect to GTP is counterbalanced by the fact that ITP is more conservative and less powerful than GTP. This power loss is anyway countered by a big gain in interpretability of the test results with respect to GTP. Indeed, differently from GTP, ITP is able to impute the rejection to specific basis components.

In conclusion, when dealing with functional data, ITP provides a good compromise between CTP and GTP, gathering the best of both procedures. Indeed, like CTP, ITP performs a selection of the significant components; and, like GTP, its computational costs remain affordable even for large values of p; furthermore, its control of the FWER and its power are intermediate between the ones provided by CTP and GTP.

We underline that all results hold for any implementation of ITP. Indeed, the corresponding proofs exclusively rely on the exactness of the family of interval-wise tests described in Subsection 2.1, and not on their nature. Thus, the theoretical results depend neither on the type and the dimension of the basis used, nor on the type of tests.

3. Extending ITP to Different Frameworks

The idea of performing a family of multivariate tests on each interval of consecutive basis components of a functional basis along the line described in Section 2 for the comparison between two independent functional populations can be applied to more complex situations. In order to apply ITP to other functional tests, we need only to define a way to construct exact multivariate tests on intervals of components of the basis expansion. For instance:

- For testing the difference in distribution of two paired functional populations, one can rely on multivariate NPC tests, plugged-in with the p-values provided by permutation-based paired t-tests on components. The permutation scheme associated to this framework would restrict possible permutations to the ones preserving the correct matching of sample units across the two samples.
- For testing the center of symmetry of a functional population, one can rely on multivariate NPC tests plugged-in with the p-values provided by permutation-based t-tests

on components for the center of symmetry of a symmetric one-dimensional distribution. The permutation scheme associated to this framework would consist in the joint reflection (for one or more sample units) of the coefficients, with respect to the hypothesized center of symmetry.

By changing the type of tests performed in the second phase of ITP, one can deal with even more complex situations, such as tests for variance, ANOVA, ANCOVA, and linear models. In detail, when all tests performed in the second phase of the algorithm are exact, the resulting ITP will provide an interval-wise control of the FWER.

4. Simulation Study

We compare here the performances of ITP against one of the most used approaches to multiple testing, that is the Benjamini–Hochberg procedure (BH, Benjamini and Hochberg, 1995) as the number of false hypotheses increases. BH deals with the problem of extremely large families of tests: when applied to a basis expansion, it controls the False Discovery Rate (FDR) over components, i.e., the expected proportion of falsely rejected components among those being rejected. The control of FDR is weaker than the strong control of FWER (i.e., BH is only provided with weak control of FWER); however, it leads to more powerful procedures compared to the ones provided with strong control of FWER.

We compared also ITP with the Bonferroni–Holm procedure (Holm, 1979), which is provided with a strong control of the FWER. Given the large number of components, the power of the Bonferroni–Holm method was always significantly lower than the one of the other testing procedures in all simulated scenarios; hence, the results for the performances of this procedure are not reported here, being not informative.

Eventually, we compare ITP with the test proposed by Vsevolozhskaya et al. (2014) (VGH test), that is based on an initial partition of the domain into q subintervals, and performs a CTP between the subintervals; hence, it is based on $2^q - 1$ tests. VGH is provided with a strong control of the FWER between subintervals, and a weak control of the FWER within each subinterval. Since the first step of the VGH test is an a priori chosen partition of the domain into subintervals, its performances can vary significantly depending on the initial choice of the partition, and on the number of subintervals q. The procedure span between these two extreme situations: (i) when $q = 1$, the VGH is a GTP. Its power is maximal, but it is only provided with a weak control of the FWER; (ii) when q is maximum (e.g., it is equal to the number of point-wise evaluations of the functional data), VGH is essentially a CTP. Thus, it is essentially provided with a strong control of the FWER, but it is both computationally unfeasible and highly conservative. In this study, as suggested by Vsevolozhskaya et al. (2014), we chose to base the VGH test on a reduced number of subintervals; in particular, we set $q = 5$. Since no a priori information is available for the choice of intervals, we decided to choose the optimal situation for VGH test, by placing subintervals in correspondence of the zones of the domain presenting differences between the two populations.

4.1. *Simulation Setting*

The simulation study is divided into two parts: in the first part, for each point of the domain, we compare the probability that the point is included in a significant interval; in the second part, we compare the probability of rejecting at least one portion of the domain where the two populations are known to be identically distributed (false discoveries) and the probability of rejecting at least one portion of the domain where the two populations are known to be differently distributed (true discoveries).

In the entire study, we consider a standard scenario for data generation. We test the differences between two independent populations of functional data on $L^2[0, 1]$, generated via cubic B-spline coefficients in a 50-dimensional space. Let $c_1^{(1)}, .., c_1^{(50)}$ be the random B-spline coefficients associated to units of the first population, and $c_2^{(1)}, .., c_2^{(50)}$ the random B-spline coefficients associated to units of the second population; let us also indicate with $\mu_1^{(k)}$ and $\mu_2^{(k)}$ the means of the coefficients, respectively, of the first and of the second population. We generate the coefficients $c_j^{(k)}$ from a normal distribution, with means $\mu_1^{(k)} = 0$, $\mu_2^{(k)} \in [0, 1]$. We suppose that the functional means of the two populations differ on a closed interval having length $h \in [0, 1]$. In this interval, the mean of the differences between coefficients is equal to a constant $v \in [0, 1]$.

Two different scenarios of alternative hypotheses are explored, by varying the parameters h and v. In the first scenario, we set a constant $v = 0.5$, and vary h between the extremes values of 0 (i.e., no difference along the entire domain) and 1 (i.e., difference along the entire domain). In the second scenario, we set a constant $h = 0.5$, and vary v between 0 and 1.

The components are generated independently: the covariance matrix of the 50-dimensional vector of differences is $\Sigma = \sigma^2 I$, with $\sigma^2 = 0.5$. Other simulations have been performed using different choices of Σ, showing that results do not change considering a more complex covariance structure. Lastly, we simulate $n_1 = n_2 = 10$ different realizations from the two populations. An instance of the simulated data for $h = 0.5$, $v = 1$ is reported in Figure 2. In all scenarios, 5000 different functional data sets are simulated.

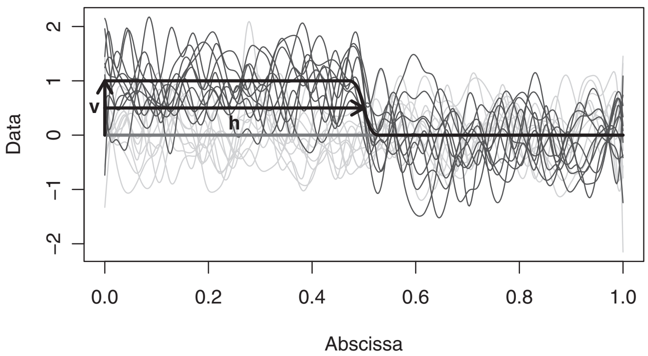


Figure 2. Functional data of the first and second group simulated for the study (light and dark gray, respectively) and functional means (bold curves), in the example $h = 0.5$, $v = 1$.

4.2. *Point-Wise Probability of Rejection*

Figure 3 reports the point-wise probability of rejection when all inferential procedures are performed at a 5% level. In particular, on each panel we depict a different scenario for h and v, and for each point of the domain we report on the vertical-axis its probability to be included in a significant interval. Also, the shaded gray portion of each panel indicates the interval where the two means are actually different, according to the simulation setting. Thus, in the white part of the plot the graph represents the probability of rejecting a point that should not be rejected (i.e., point-wise error rate), while in the gray part the probability of rejecting a point that should be rejected (i.e., point-wise powers). The top panels of Figure 3 show the results obtained varying the parameter h when $v = 0.5$; the lower panels of the same Figure show the results obtained varying the parameter v when $h = 0.5$.

The simulation shows that both ITP and BH provide the control of the point-wise error in each scenario. We notice also that ITP maximizes the point-wise power at the center of the intervals where the difference occurs, while BH point-wise power has a flat shape. This is due to the fact that ITP exploits the ordered structure of the components, gaining power in the center of the false hypotheses interval. BH procedure adjusts instead the p-values without considering the fact that coefficients are ordered. Hence, ITP appears to be more powerful in detecting the presence of a significant interval, while being more conservative with respect to the amplitude of the interval (i.e., its power at the boundaries is lower). BH procedure is less powerful in revealing the presence of a significant interval, but once detected, it is able to detect its actual amplitude.

VGH test controls the point-wise error rate only in the scenarios where the a priori chosen subintervals can identify exactly the “true” and “false” zones of the domain, that is, when $h = 0, 0.2, \dots, 1$. In all other cases, the point-wise error rate is not controlled in the subintervals intercepting both “true” and “false” zones. Because of the reduced number of tests, the power of VGH test is significantly higher than both BH and ITP ones in all scenarios.

4.3. *Global Probability of Rejection*

In top panels of Figure 4 we report, for each testing procedure, the estimated probability that at least one portion of the domain where the two populations are known to be identically distributed is rejected. The estimated probability is obtained by varying the parameters h (left) and v (right). We notice that ITP controls this probability for all scenarios, while this control is in general not guaranteed by BH procedure. In the case of VGH procedure, the control of the FWER is guaranteed when the partition on subintervals expresses exactly the partition of the domain into true and false hypotheses ($h = 0, 0.2, \dots, 0.8$), but it is not controlled in all other cases ($h = 0.1, 0.3, \dots, 0.9$). This latter finding is consistent with the fact that VGH procedure is only provided with a weak control of the FWER within intervals, implying a control exclusively when the two populations do not differ on the entire subinterval.

In lower panels of Figure 4, we report for each testing procedures the estimated probability that at least one portion of the domain where the two populations are known to be differently distributed is rejected. In all explored cases, ITP outperforms BH procedure. This confirms what was noticed for the point-wise results, i.e., that ITP is more powerful in detecting the presence of an interval with differences between the two populations. VGH test outperforms both ITP and BH. This gain presents the drawback (in some cases) of a dramatic loss of control of the probability that at least one portion of the domain where the two populations are identically distributed is rejected.

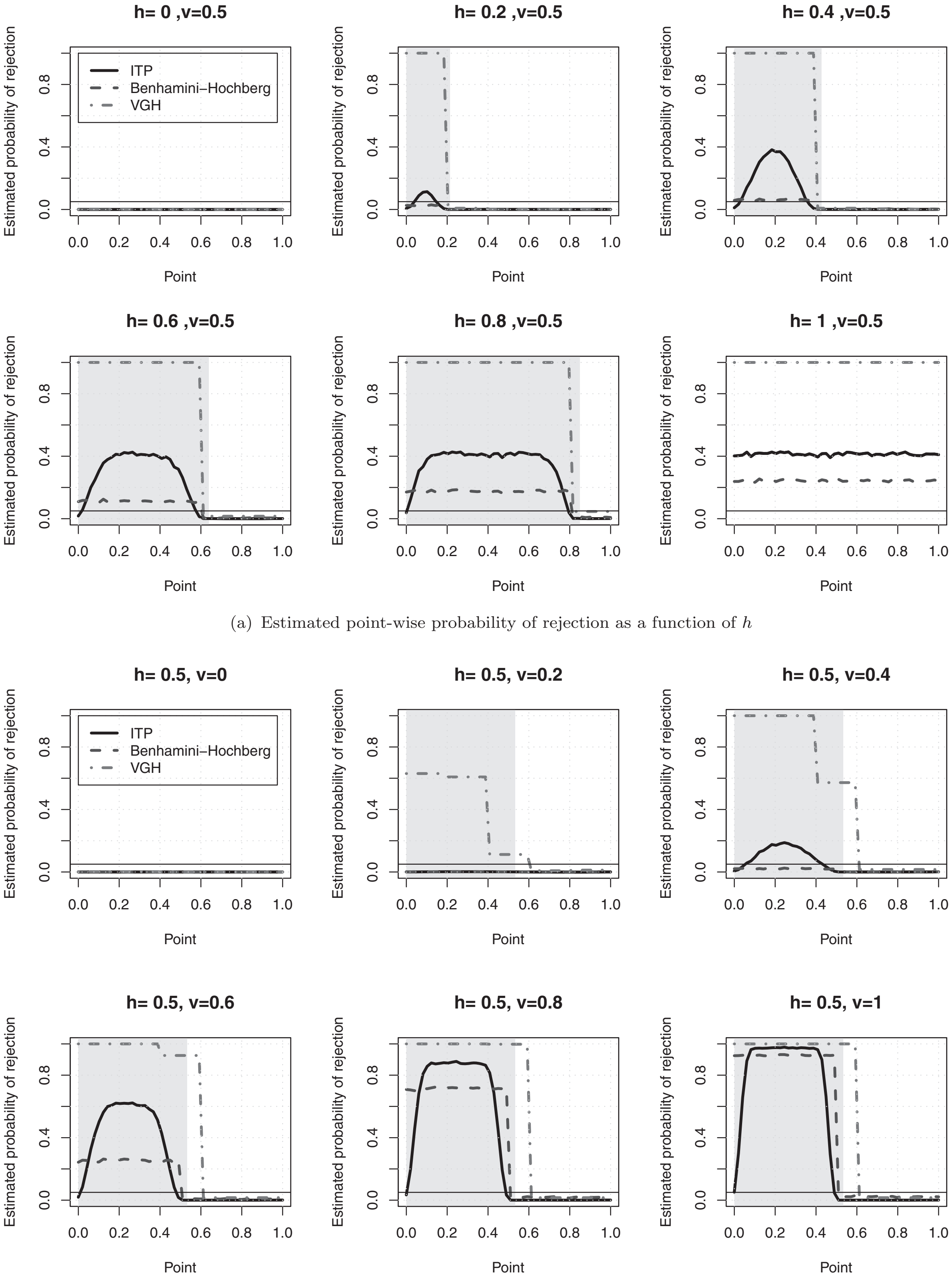
5. Analysis of the Aneurisk Data Set

We present in this Section the analysis of the Aneurisk data set (Passerini et al., 2012), which deals with the geometrical and hemodynamical features of the internal carotid arteries (ICA) of subjects affected by a cerebral aneurysm. The aim of this analysis is to assess whether the geometry and/or the hemodynamics of the ICA is related to the type and severity of the pathology. In particular, we look for possible differences in the distributions of vessel-radius, centerline-curvature, and wall-shear-stress (WSS)—as functions of the arch-length along the carotid centerline—between two groups: a first group composed of 25 subjects affected by a severe form of the pathology (patients with an aneurysm in the upper part of the brain within the skull), namely upper group; and a second group composed of 25 subjects affected by a minor form of the pathology (patients with an aneurysm in the lower part of the head outside the skull) or healthy (without any aneurysm), namely lower group. Data were obtained from angiographic exams, where an overall number of points ranging from 350 to 1380 were acquired for each subject. A detailed description of data gathering and processing may be found in Passerini et al. (2012). We show in the bottom panels of Figure 5 the projection of data on $p = 100$ B-splines of order $m = 3$, having uniformly spaced knots for radius, curvature, and WSS; upper and lower group functions are reported in light and dark gray, respectively.

We perform three separated analyses for radius, curvature, and WSS functions, respectively, by implementing three ITPs for detecting the differences between two independent functional populations. The tests of the second step are obtained using the NPC procedure based on Fisher combination function (Pesarin and Salmaso, 2010).

The adjusted p-values of the three ITPs are reported in the central panels of Figure 5. At level $\alpha = 5\%$, we do not detect any statistical difference between upper and lower groups, neither for the radius nor the curvature functions; instead, we detect a difference in terms of WSS. Being the support of the B-splines localized with respect to the arc-length (x-axis), we can impute the rejection to the segment of the carotid associated to the arc-length interval ($-2.783\,\text{cm}, -1.632\,\text{cm}$), depicted in the gray region in bottom panels of Figure 5. We found lower WSS for very severe subjects (upper group) and higher WSS for less severe subjects (lower group).

Hemodynamics might explain this finding: the latter region corresponds to the second bend of the ICA, which is the segment where a second peak of curvature is present and where the ICA starts getting narrower. Bends of the ICA are “guardians” of the arteries of upper part of the brain, being among the weakest in the entire body, as they are not sur-



(a) Estimated point-wise probability of rejection as a function of h

(b) Estimated point-wise probability of rejection as a function of v

Figure 3. Estimated point-wise probability of rejection at a 5% level for the considered multiple testing procedures on each scenario.

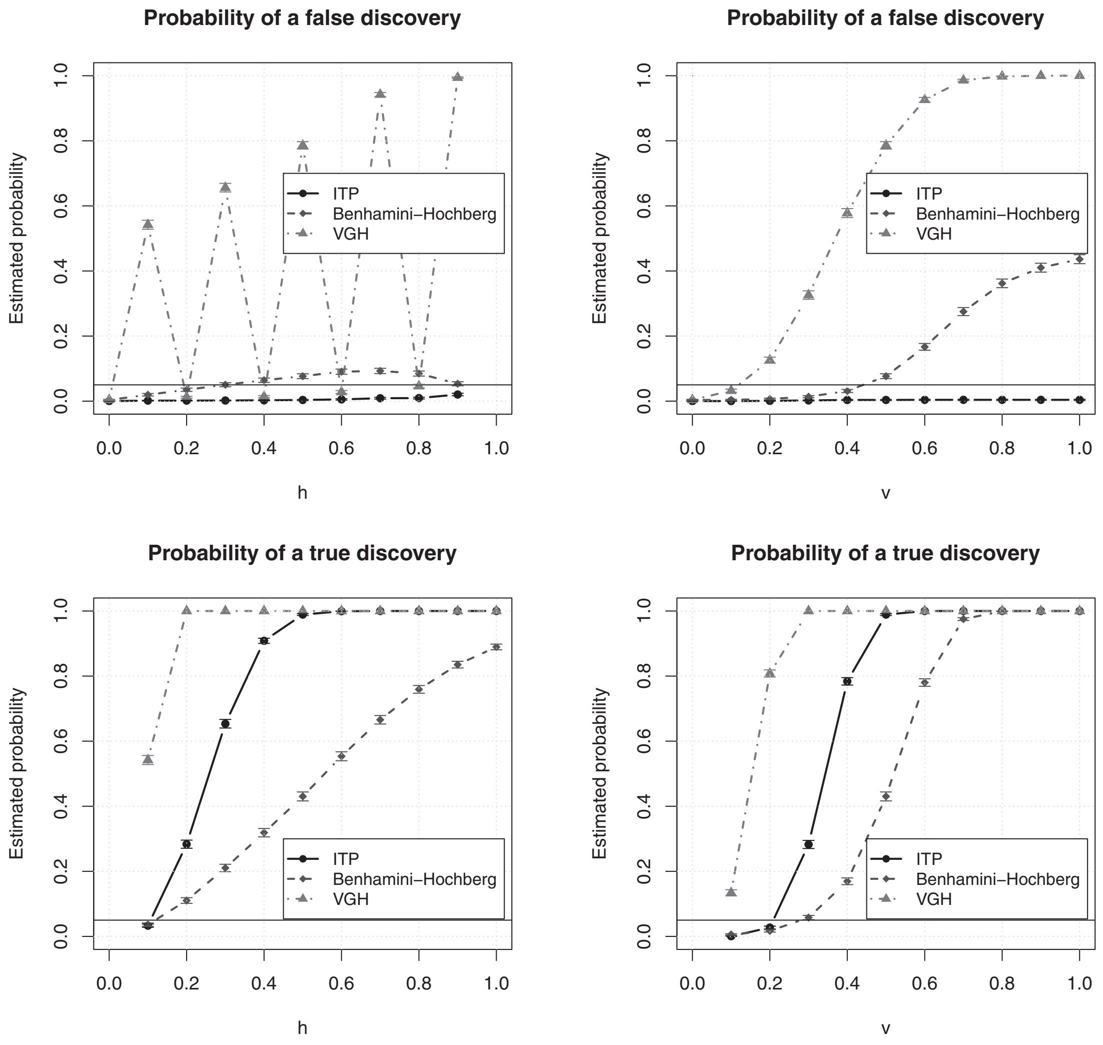


Figure 4. Estimated probability of having at least one false discovery (top) and estimated probability of having at least one true discovery (bottom) as functions of h (left) and v (right), for $h, v \in \{0, 0.1, 0.2, ..., 1\}$. The error bands indicate the 95% confidence intervals for the real probability.

rounded by muscular tissues. Thanks to the passage through the bends the unsteady blood flow from the heart becomes steadier before entering the brain. This "stabilization" effect is related to the loss of energy of the blood flow related in turn to the magnitude of WSS within the bends. Hence, ITP provides the statistician (S) with a tool for answering questions normally pointed out by a practitioner (P) dealing with this application.

P: "Are the curves belonging to upper and lower groups statistically different?"

S: "Curves are not different with respect to radius and curvature, but they are with respect to WSS."

P: "What significant differences are there in WSS curves?"

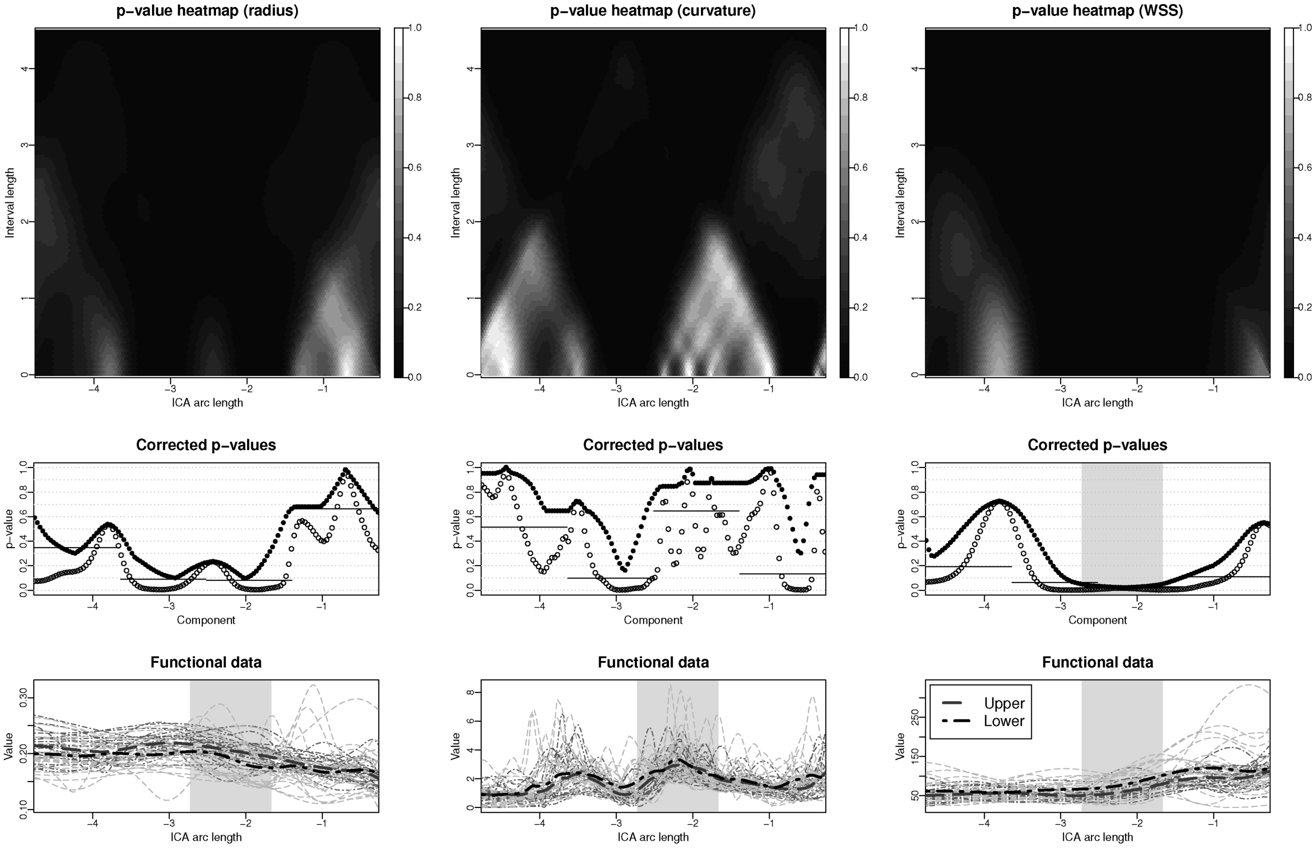
S: "Differences between the two groups are significant in the arc-length interval (−2.783 cm, −1.632 cm)."

P: "Can we state that those differences did not just pop up by chance?"

S: "The probability that this result popped up by chance is lower than 5%. Indeed, if distributions between the two population had no difference in segment (−2.783 cm, −1.632 cm), the probability of pointing out that segment would be less than 5%."

For theoretical interest, we show on the top panels of Figure 5 the heat-map of all p-values of interval-wise tests performed in the second step of the procedure (used to compute the adjusted p-values), and in the middle panels both the adjusted and the unadjusted component p-values (full dots and empty dots, respectively).

CTP in this application is not computationally feasible, as more than 10^{38} multivariate tests would be required; GTP instead, while rejecting the null hypothesis of no difference in the WSS between the groups, is unable to detect the segment of the carotid where this difference is. Bonferroni–Holm correction is not capable of detecting any difference between the two groups. Benjamini–Hochberg correction detects an



(a) Radius (b) Curvature (c) WSS

Figure 5. Aneurisk case study analysis of radius (left), curvature (center), and WSS (right). Top: p-value heat-maps; center: unadjusted (empty dots) and adjusted (full dots) p-values, and adjusted p-values according to the VGH test (lines); bottom: curves of the upper and lower groups (light and dark gray, respectively) and sample means associated to the two groups (bold curves). The shaded part indicates the interval where significant differences area found in terms of WSS.

interval (-3.239 cm, -1.210 cm), which is larger than the one detected by ITP. This is consistent with the weaker control of the FWER provided by Benjamini–Hochberg correction, and with the property that, on average, up to 5% of this interval is expected to be composed of false discoveries.

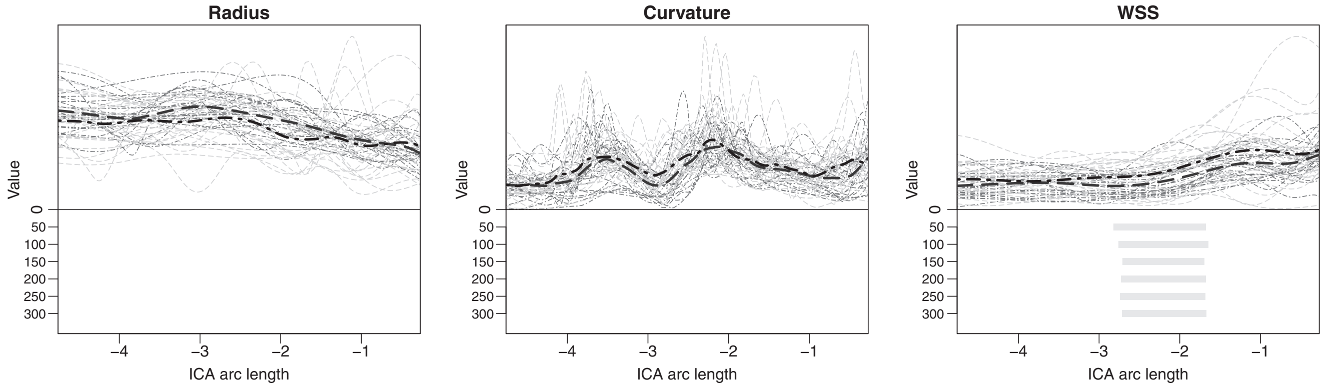
The adjusted p-values of VGH test based on four equally spaced intervals are reported in Figure 5 (black lines). Coherently with ITP, VGH test detects a significant difference between the two groups in terms of WSS on the third interval (-2.511 cm; -1.395 cm). Length and position of such interval are approximately coincident with the one detected by ITP. The main difference is that, in the case of VGH test, we know that there is a significant difference in at least a part of the interval (-2.511 cm; -1.395 cm), but we cannot locate it precisely. In the case of ITP, on the contrary, we know that data are significantly different on the whole interval (-2.783 cm, -1.632 cm).

Finally, in order to appreciate the robustness of our conclusions with respect to parameter p, we report in Figure 6 the results of the three tests on radius, curvature, and WSS based on different values of p. In the lower part of each panel, we report the gray intervals presenting significant differences, detected with six different values of p ranging from 50 to 300. Findings appear to be robust with respect to this parameter.

6. Discussion

We presented in this work, a novel inferential procedure suited for functional data analysis (FDA). Our procedure, named Interval Testing Procedure (ITP), involves three steps: (i) representing functional data on a functional basis; (ii) performing a family of multivariate tests on intervals of components; (iii) computing an adjusted p-value for each basis component. ITP can be easily modified to deal with several inferential problems occurring in FDA as, for example, the comparison of two or more functional populations. The inference carried out by ITP is semiparametric in the sense that we make use of a parametric basis expansion to represent data, but we do not introduce any strong distributional assumption on the coefficients of the expansion (for instance, we do not assume Normality).

We introduced the definition of interval-wise control of the Family Wise Error Rate (FWER) which is particularly meaningful in the framework of FDA and which ITP is provided with. In particular, interval-wise control of the FWER (lying in between weak and strong controls of the FWER) refers



(a) Radius (b) Curvature (c) WSS

Figure 6. Functional data of radius (left), curvature (center), and WSS (right). Results of the three ITPs on radius (left), curvature (center), and WSS (right). The gray intervals in the lower part of each graphic indicate intervals presenting significant differences detected with six different values of p, from 50 to 300.

to the property of controlling the FWER over all intervals of components of the basis expansion. For example this control—when associated to a B-spline expansion—implies that given any interval of the domain where there is no difference between two functional populations, the probability that at least a part of such interval is wrongly detected as significant is controlled at the desired level.

In addition to the proof of the interval-wise control property of ITP, we also showed that the component-wise and global statistical power of ITP is always higher than the Closed Testing Procedure one, this latter being a procedure that on one hand provides a strong control of FWER, but on the other hand is not computationally affordable in the functional framework. Similarly, we proved that the component-wise and global power of the ITP is always lower than the Global Testing Procedure one, this latter providing only a weak control of FWER, and no guide for the interpretation of the test result in terms of components.

Even though all theoretical properties shown in this work hold for any dimension p of the basis, the exploration of the performances of the procedure as p increases is of major interest for future research. In general, p should be set high enough in order to have a good representation of the functional data; in the case of noisy evaluations of the functional data, this may be done by applying a smoothing spline basis expansion. The higher is p, the higher is the precision in the selection of intervals presenting significant differences; this also implies that, for obtaining good results, functional data need to be evaluated on a sufficiently dense grid, in order to have a more accurate information on intervals presenting significant differences. This fact poses some limits to the adoption of ITP in very sparse settings. Finally, preliminary explorations suggest also the attainment of a gain of power as p increases, supporting further the choice of large values for p. However, as one of the reviewers pointed out, in the presence of noise in the evaluations of the functional data, increasing p can act as nuisance parameter. In this case, preliminary explorations show that if regression splines are used to estimate the basis coefficients, the parameter p plays a key role in the bias/variance tradeoff. Instead, if smoothing splines (with penalization) are used, the effect of p on the power stabilizes after reaching a sufficiently high value. Such results suggest to further explore the effect of smoothing on the power of ITP, and to perform robustness analyses to evaluate its effect on the test results when analyzing real data sets.

A comparison with Bonferroni–Holm procedures, with Benjamini–Hochberg procedure, and with a recently proposed state-of-the-art procedure (WGH, described in Vsevolozhskaya et al., 2014) has been carried out through a simulation study. In the scenarios explored in this simulation, ITP appears to be more powerful than the Bonferroni–Holm procedure. Furthermore, ITP appears to be more powerful than Benjamini–Hochberg procedure for the detection of the presence of a significant interval, while being more conservative with respect to its amplitude. On the contrary, Benjamini–Hochberg procedure is less powerful in detecting the presence of a significant interval; however, once detected, it is more capable of stating its actual amplitude. Lastly, WGH test is more powerful than ITP when a reduced number of subintervals are chosen, but it is provided only with a weak control within each preselected subinterval and it lacks of any control on intervals different from the preselected ones. Hence, it can only select intervals where there are significant differences, but it is unable to locate them precisely.

We reported the application of ITP to a case study to show its potential in practice. We performed a B-spline-based inference for the difference between radius, curvature, and WSS curves along the Internal Carotid Artery (ICA) of two pathologically different groups of subjects. The application of ITP in this case allows for the selection of a portion of ICA presenting a difference in WSS between the two groups.

The R-package `fdatest` (Pini and Vantini, 2015) with the implementation of ITP is available with this article at the *Biometrics* website on Wiley Online Library and on CRAN. The current version of the package requires functional data evaluated on a uniform grid, and it provides the following features: (i) the projection of each function on a chosen functional basis using Ordinary Least Squares; (ii) the execution of the entire family of interval-wise multivariate tests; and, (iii), the computation of the vector of adjusted p-values, to be used for the selection of the statistically significant basis components at level α. Also, the package provides a plotting function to create graphical outputs, like the ones presented in Figure 5.

7. Supplementary Materials

The web-based Supplementary Material containing the proofs of theorems referenced in Section 2, and the current version (2.1) of `fdatest` package is available with this article at the *Biometrics* website on Wiley Online Library. Please refer to CRAN for the latest version of the `fdatest` package (Pini and Vantini, 2015).

References

Abramovich, F. and Angelini, C. (2006). Testing in mixed-effects FANOVA models. *Journal of Statistical Planning and Inference* **136**, 4326–4348.

Antoniadis, A. and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis* **51**, 4793–4813.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **57**, 289–300.

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

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An ANOVA test for functional data. *Computational Statistics & Data Analysis* **47**, 111–122.

Fan, J. and Lin, S. (1998). Test of significance when data are curves. *Journal of the American Statistical Association* **93**, 1007–1021.

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

Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89**, 359–374.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, volume 200. New York: Springer.

Passerini, T., Sangalli, L. M., Vantini, S., Piccinelli, M., Bacigaluppi, S., Antiga, L., et al. (2012). An integrated statistical investigation of internal carotid arteries of patients affected by cerebral aneurysms. *Cardiovascular Engineering and Technology* **3**, 1–15.

Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. Chichester: John Wiley & Sons Inc.

Pini, A. and Vantini, S. (2015). *fdatest: Interval Testing Procedure for Functional Data*. R package version 2.1.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*, volume 77. New York: Springer.

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

Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2009). A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association* **104**, 37–48.

Schott, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *Journal of Multivariate Analysis* **98**, 1825–1839.

Shen, Q. and Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica* **14**, 1239–1258.

Spitzner, D. J., Marron, J. S., and Essick, G. K. (2003). Mixed-model functional ANOVA for studying human tactile perception. *Journal of the American Statistical Association* **98**, 263–272.

Staicu, A., Li, Y., Crainiceanu, C. M., and Ruppert, D. (2014). Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics* **41**, 932–949.

Vsevolozhskaya, O., Greenwood, M., and Holodov, D. (2014). Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *Annals of Applied Statistics* **8**, 905–925.

Zhang, J. and Liang, X. (2014). One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics* **41**, 51–71.

Received March 2015. *Revised December* 2015.
Accepted December 2015.