MOX–Report No. 13/2013

The Interval Testing Procedure: Inference for Functional Data Controlling the Family Wise Error Rate on Intervals.

Pini, A.; Vantini, S.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

The Interval Testing Procedure: Inference for Functional Data Controlling the Family Wise Error Rate on Intervals.

A. Pini [a], S. Vantini [a]

[a] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano

alessia.pini@mail.polimi.it
simone.vantini@polimi.it

Keywords: Functional Data, Inference, Family Wise Error Rate, Permutation Test.

AMS Subject Classification: 62H15, 62G10, 62H99.

Abstract

We propose a novel inferential technique based on permutation tests that enables the statistical comparison between two functional populations. The procedure (i.e., Interval Testing Procedure) involves three steps: (*i*) representing functional data on a suitable high-dimensional ordered functional basis; (*ii*) jointly performing univariate permutation tests on the coefficients of the expansion; (*iii*) combining the results obtaining a suitable family of multivariate tests and a *p*-value heat-map to be used to correct the univariate *p*-values. The procedure is provided with an interval-wise control of the Family Wise Error Rate. For instance this control, which lies in between the weak and the strong control of the Family Wise Error Rate, can imply that, given any interval of the domain in which there is no difference between the two functional populations, the probability that at least a part of the domain is wrongly detected as significant is always controlled. Moreover, we prove that the statistical power of the Interval Testing Procedure is always higher than the one provided by the Closed Testing Procedure (which provides a strong control of the Family Wise Error Rate but it is computationally unfeasible in the functional framework). On the contrary, we prove that the power of the Interval Testing Procedure is always lower than the Global Testing Procedure one (which however provides only a weak control of the Family Wise Error Rate and does not

This research has been funded by the Large *p* Small *n* Project - Politecnico di Milano.

provide any guide to the interpretation of the test result). The Interval Testing Procedure is also extended to the comparison of several functional populations and to the estimation of the central function of a symmetric functional population. Finally, we apply the Interval Testing Procedure to two case studies: Fourier-based inference for the mean function of yearly recorded daily temperature profiles in Milan, Italy; and B-spline-based inference for the difference between curvature, radius and wall shear stress profiles along the Internal Carotid Artery of two pathologically-different groups of subjects. In the supplementary materials we report the results of a simulation study aiming at comparing the novel procedure with other possible approaches. An R-package implementing the Interval Testing Procedure is available as supplementary material.

1 Introduction

During the last years, due to the fast development of more and more precise acquisition devices in many research areas, scientists have started dealing with the analysis of high dimensional data sets, that is, data sets characterized by a number p of features observed for each sample unit much larger than the number n of sample units. In such situations, we generally talk about "large p small n problems".

An extreme example of a large p small n data is constituted by functional data. Indeed, in functional data analysis (FDA) data are even no longer p-dimensional vectors, but functions observed in a continuous domain and lying in an infinite dimensional separable Hilbert space (Ramsay and Silverman 2005, 2002; Ferraty and Vieu 2006). In the practice, statisticians deal with this kind of data by projecting them on a finite dimensional space spanned by a suitable truncated functional basis (Ramsay and Silverman 2005), which may be fixed (e.g., Fourier basis, B-splines, wavelet basis, polynomial basis) or data driven (e.g., functional principal components).

The major issue for the analysis of large p small n data is constituted by the fact that many classical multivariate inferential tools (e.g., Hotelling's theorem) become pretty useless in this framework, since they require the number of sample units to be greater than the dimension of the space in which data are observed. Consequently, the growing interest for the analysis of this type of data is urging the development of inferential techniques suited for any value of n and p.

Many methods dealing with the large p small n problems are object of statistical investigation. In particular, these techniques may be classified in two different categories: the ones just focusing on "global inference" and the one focusing on "component-wise inference" as well. Techniques for global inference are made by a unique global test that provides a unique result. These procedures controls the global level of the test (i.e., weak control of the Family Wise Error Rate). In the parametric case, examples of such techniques are derived from

suitable generalizations of Hotelling's theorem (Srivastava 2007; Secchi et al. 2011). In the non parametric case, an example of these technique is constituted by the NPC permutation test (Pesarin and Salmaso 2010). On the one hand, these procedures are feasible even when the number of components is very large but, on the other one, they can just state if there is enough evidence to reject the null hypothesis without imputing the rejection to specific components. On the contrary, in large p small n problems, a method allowing the selection of the components of data set which are significantly different in distribution is often desirable. As an example, suppose that a functional data set is represented through the expansion on a suitable basis. In order to perform a dimensional reduction of the data set, it might be useful to select the components of the expansion that are significant for the specific test in exam. A global test would not be able to provide this kind of information.

Techniques that focus on components are instead based on the joint use of com-ponent-specific univariate statistics. Indeed, the central idea is the decomposition of the initial high-dimensional test into lower dimensional subproblems, which are usually characterized by marginal univariate hypotheses. Then, each subproblem is tested with a classical technique and finally the test results are corrected in order to assure the control of the level of the test for each possible set of true null hypotheses (i.e., strong control of the Family Wise Error Rate). According to these techniques, a global result for the test is given as in the previous case, and furthermore, a selection of rejected sub-hypotheses is provided. Examples of such approach are the Bonferroni and the Bonferroni-Holm correction (Holm 1979), and the Closed Testing Procedure (CTP) (Marcus et al. 1976). On the one hand, these procedures provide a strong control of the Family Wise Error Rate (FWER) and enables the selection of a smaller subset of significant components. On the other one, they are generally not suited to deal with large p small n data. Indeed as p increases their computational cost might explode and/or their power can become very low. In particular, if $p \gg n$, the Bonferroni correction, which implies the division of the level of the test by p, leads to a highly conservative and low-power procedure while the CTP requires to test the closure family of the set of marginal hypotheses (i.e., $2^p - 1$ tests) making the computations quickly unfeasible.

The Interval Testing Procedure (ITP), which we hereby propose, is meant for dealing with functional data and it lies in between these two different approaches. Similarly to the component-wise inferential techniques, it is able, in case of rejection, to highlight which components are significantly different. Differently from them, even when the number of components is very large, its power remains comparable with the one provided by the global inference techniques and its computational cost grows just quadratically in p. Since "there is no such thing as a free lunch" the ITP lacks the strong control of the FWER. Indeed it just provides an "interval-wise" control of the FWER (which is stronger than the weak control provided by global test but weaker than the strong control provided by component-wise procedures). In the FDA framework this is a minor drawback

since this kind of control is often the only one you might want. Indeed, all functional bases commonly used in FDA present a natural ordered structure: Fourier components are ordered according to frequency, B-spline components according to the abscissa, wavelet components according to the abscissa and the frequency, Taylor components (or polynomial-inspired components in general) according to roughness, functional principal components according to variance. For example, when testing for the difference between two functional populations, the "interval-wise" control of the FWER in the Fourier expansion implies that, for any band of frequencies (including the entire set of explored frequencies or the set of all frequencies greater or lower than a certain threshold), if there is no difference between the two functional populations in that band, it means that the probability that at least one component of the band is wrongly detected as differently distributed in the two populations is controlled. As a second example, you might think a the B-spline representation. Indeed, because of the compact support of the B-spline basis elements you have the control of the FWER on intervals of the domain, i.e., if there is no difference between the two populations in an interval of the domain, the probability that the two population are detected as significantly different on part of this interval is controlled. Finally, note that the ITP is based on permutation tests which just require the exchangeability of the sample units under the null hypothesis and since the sample functional principal component scores result exchangeable, we have that the ITP is also suited to deal with the functional principal component basis expansion.

The paper is outlined as follows: in Section 2 the ITP is described for the two population framework (i.e., testing for the difference between two functional populations in both the coupled and the uncoupled scenario). In Section 3, the ITP is declined in the test for the mean of one functional population and in the test for the difference among $g > 2$ populations. The theoretical properties of the ITP, both in terms of control of the FWER and of its power, are proven in Section 4. In Sections 5 and 6 the ITP is applied to two case studies, respectively. In particular, the first one pertains the estimation of the mean function of yearly-recorded daily temperature profiles in Milan, Italy (NASA 2008). The second case study is instead devoted to the analysis the Aneurisk data set (Sangalli et al. 2009a) and it concerns the comparison between geometric and hemodynamic features of the internal carotid artery in two groups of patients associated to different levels of severity of the cerebral aneurysm pathology. Finally, in the supplementary materials we provide the results of a simulation study comparing the performances of the ITP with other multiple testing procedures, as well as an R-package implementing the ITP for one or two populations of functional data evaluated on a uniform grid. All computations and images have been created using R (R Core Team 2012).

2 The ITP in the Two-Population Framework

Let $\mathbf{y} = \{\mathbf{y}_{11}, \mathbf{y}_{21}, ..., \mathbf{y}_{n_1 1}, \mathbf{y}_{12}, ..., \mathbf{y}_{n_2 2}\}$ be a collection of $n = n_1 + n_2$ functions, and assume that the set of functions $\mathbf{y_1} = \{\mathbf{y}_{i1}\}$, $i = 1, ..., n_1$ represents a random sample from a first functional population $\mathbf{Y}_1$ and the remaining $\mathbf{y_2} = \{\mathbf{y}_{i2}\}$, $i = 1, ..., n_2$ a sample from a second functional population $\mathbf{Y}_2$. Consequently, assume that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are two random functions taking values in a separable functional space $\mathcal{Y}$. We aim at testing the null hypothesis $\mathbf{Y}_1 \overset{\text{d}}{=} \mathbf{Y}_2$ against the alternative $\mathbf{Y}_1 \overset{\text{d}}{\neq} \mathbf{Y}_2$, in both the uncoupled and the coupled scenario:

- **Uncoupled scenario**: we assume independence between the first n_1 units and the remaining n_2 units. We make the assumption that $\mathbf{y}_{11}, ..., \mathbf{y}_{n_1 1} \overset{\text{iid}}{\sim} \mathbf{Y}_1$, $\mathbf{y}_{12}, ..., \mathbf{y}_{n_2 2} \overset{\text{iid}}{\sim} \mathbf{Y}_2$, where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are two independent random functions.
- **Coupled scenario**: we assume a coupled dependence between the units of the first group and the ones of the second group. In particular, $n_1 = n_2$ and units are coupled across groups, i.e., $(\mathbf{y}_{11}, \mathbf{y}_{12}), ..., (\mathbf{y}_{n_1 1}, \mathbf{y}_{n_1 2}) \overset{\text{iid}}{\sim} (\mathbf{Y}_1, \mathbf{Y}_2)$.

The testing procedure we hereby propose is composed by the following steps:

1. **Basis Expansion**: functional data are represented through the coefficients of a suitable basis expansion;
2. **Joint Univariate Tests**: univariate permutation tests for the basis coefficients are jointly performed;
3. **Interval-wise Combination and Correction of the Univariate Tests**: the univariate tests are suitably combined and then p-values are corrected to obtain an interval-wise control of the FWER.

2.1 First Step: Basis Expansion

Theoretically, each function can be univocally represented through a countable sequence of coefficients associated to a suitable basis of the functional space $\mathcal{Y}$ (i.e., Fourier harmonics, B-splines, wavelets, ...). In the practice, very rarely functional data come with an analytic expression. More often, just some pointwise evaluations of a function (possibly with some noise) are available, and thus just a reduced number of components can be estimated. It is thus necessary to represent data by means of an expansion on a reduced basis $\{\phi^{(k)}\}_{k=1,...,p}$:

$$y_{ij}(t) = \sum_{k=1}^{p} c_{ij}^{(k)} \phi^{(k)}(t) \tag{1}$$

This projection constitutes the first step in most FDA procedures, and it is presented in detail in Ramsay and Silverman (2005). The integer p represents

the finite dimension of the functional space in which data are represented and, to have a good description of data, it is generally greater than n, at least when a strong data smoothing is not planned. In particular, when data are constituted by J observations of each function, we have typically $n \ll p \leq J$. In detail, in this particular context we are interested in representing data without loss of information, rather than reducing *a priori* data dimension. Thus, we will always set p as big as possible.

In the end, we can represent each of the n units by means of the p coefficients $c_{ij}^{(k)} \in \mathcal{C} \subseteq \mathbb{R}$, $i = 1, ..., n_j$, $j = 1, 2$ associated to the expansion (1). In particular, for each component k of the expansion, we obtain n_1 coefficients associated to units of the first group $\mathbf{c}_1^{(k)} = \{c_{i1}^{(k)},\, i = 1, ..., n_1\}$ and n_2 coefficients associated to units of the second group $\mathbf{c}_2^{(k)} = \{c_{i2}^{(k)},\, i = 1, ..., n_2\}$. As the basis coefficients represent sampled data, the hypotheses made for the functional populations can be re-stated in terms of the expansion coefficients: in the uncoupled case we have for each k, $c_{11}^{(k)}, ..., c_{n_1 1}^{(k)} \overset{\text{iid}}{\sim} C_1^{(k)}$, $c_{12}^{(k)}, ..., c_{n_2 2}^{(k)} \overset{\text{iid}}{\sim} C_2^{(k)}$, and in the coupled one $(c_{11}^{(k)}, c_{12}^{(k)}), ..., (c_{n_1 1}^{(k)}, c_{n_1 2}^{(k)}) \overset{\text{iid}}{\sim} (C_1^{(k)}, C_2^{(k)})$.

2.2 Second Step: Joint Univariate Tests

The second step of the ITP consists in jointly performing p univariate permutation tests for the coefficients of the basis expansion (1). In particular, we aim at testing the differences between the two populations for each $k = 1, ..., p$ by means of an univariate test on the kth coefficient, defined by:

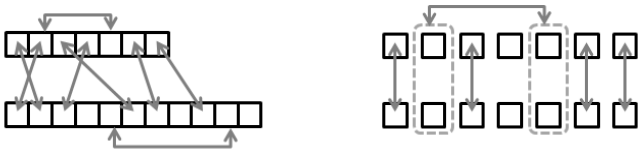
$$H_0^{(k)} : C_1^{(k)} \overset{\text{d}}{=} C_2^{(k)} \quad \text{vs} \quad H_1^{(k)} : C_1^{(k)} \overset{\text{d}}{\neq} C_2^{(k)} \tag{2}$$

In order to perform a marginal test for each k, we introduce a suitable permutation test, based on a family of data transformations which preserve the likelihood under the null hypothesis $H_0^{(k)}$ and a suitable test statistic, stochastically larger under $H_1^{(k)}$ than under $H_0^{(k)}$. The family of transformations depends on the assumptions of independence between the two samples (i.e., coupled or uncoupled test). The test statistic depends instead on the structure of the basis.

In particular, fix the basis component k, and let $\mathbf{c}^{(k)} = (\mathbf{c}_1^{(k)}, \mathbf{c}_2^{(k)})$ the $n_1 + n_2$ dimensional vector of the coefficients associated to units of the two groups, and $\mathbf{c}^{(k)^*} = (\mathbf{c}_1^{(k)^*}, \mathbf{c}_2^{(k)^*})$ the vector of the permuted coefficients. The family of likelihood-invariant transformations depends on the type of test:

- in the **uncoupled** case, we have the total exchangeability under $H_0^{(k)}$, thus the family of transformations is composed by any permutation over the sample units of the observed values.

- in the **coupled** case, under $H_0^{(k)}$ exchangeability is just between and within couples (e.g., if we want to preserve likelihood, couples cannot be splitted



(a) Uncoupled permutations (b) Coupled permutations

Figure 1: Examples of some possible likelihood-invariant transformations of the original data set in the coupled and uncoupled scenarios.

up). The family of transformations is thus composed by between and within-couple permutations of the observed values.

Examples of possible likelihood-invariant transformations of the original data set in the uncoupled and coupled scenario are presented in Figure 1.

It is important to note that, being the different components $C_1, C_2, \ldots, C_p$ possibly dependent, the permutations of the coefficients need to be jointly performed, i.e., each permutation is applied simultaneously to the entire set of coefficients. This is the key to build the multivariate tests needed in the next step.

The test statistic $T(\mathbf{c}^{(k)^*})$ used for the univariate permutation tests of the expansion coefficients depend on the type of test to be performed (i.e., coupled or uncoupled), and on the functional basis used to describe data. Indeed, in the permutation framework the test statistic has to be properly chosen in order to reflect the characteristics of data which are expected to change the most under the alternative hypothesis.

Once chosen the test statistic, for each k, the p-value of the corresponding test (2) is estimated through a conditional MC algorithm (Pesarin and Salmaso 2010), as the proportion of $T(\mathbf{c}^{(k)^*})$ exceeding the value $T(\mathbf{c}^{(k)})$ calculated on the original data set.

To better understand how a test statistic may be selected and how the test statistic may depend on the type of test and on the basis used for the analysis, we report some examples that will be used in the two applications.

Example 1: B-spline Basis. Suppose that a difference between the two functional populations is suspected to occur exclusively on an unknown region of the domain. Then, a quite natural choice to target this problem is the use of the B-spline basis. In particular, we fix a grid of knots along the abscissa, and express each data through the p coefficients associated to the B-spline basis functions $b_m^{(k)}(t)$ of order m: $y_{ij}(t) = \sum_{k=1}^{p} c_{ij}^{(k)} b_m^{(k)}(t)$ (Bosq 2000).

If we consider the uncoupled case, then, a possible test statistic for each test (2) can be defined as the difference between the two sample means of the

coefficients:

$$T(\mathbf{c}^{(k)^*}) = \frac{1}{n_1}\sum_{i=1}^{n_1} c_{i1}^{(k)^*} - \frac{1}{n_2}\sum_{i=1}^{n_2} c_{i2}^{(k)^*}.$$

If, on the contrary, we consider the coupled scenario, the same test statistic can be properly rewritten as the sample mean of the differences between the coupled coefficients:

$$T(\mathbf{c}^{(k)^*}) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(c_{i1}^{(k)^*} - c_{i2}^{(k)^*}\right).$$

Example 2: Fourier Basis. Suppose now that data are T-periodic curves, and that we expect a difference between the two populations in a frequency band. Thus, it is natural to express data in the frequency domain by means of a Fourier expansion, which can be expressed in both following representations:

$$y_{ij}(t) = m_{ij}^{(0)} + \sum_{k=1}^{p}\left(a_{ij}^{(k)}\cos\left(\frac{2\pi}{T}kt\right) + b_{ij}^{(k)}\sin\left(\frac{2\pi}{T}kt\right)\right); \tag{3}$$

$$y_{ij}(t) = m_{ij}^{(0)} + \sum_{k=1}^{p}\alpha_{ij}^{(k)}\cos\left(\frac{2\pi}{T}kt + \phi_{ij}^{(k)}\right), \tag{4}$$

The first expression (3) is exactly of the type (1), and associates each frequency k to the coefficients $a_{ij}^{(k)}$ and $b_{ij}^{(k)}$. The second expression (4) associates instead each frequency to an amplitude and to a phase coefficient (i.e., $\alpha_{ij}^{(k)}$ and $\phi_{ij}^{(k)}$) leading to a more interesting interpretation. Coherently, for the "$0th$" frequency, we can define the amplitude and phase coefficients as $\alpha_{ij}^{(0)} = |m_{ij}^{(0)}|$ and $\phi_{ij}^{(0)} = \pi[1 - \text{sign}(m_{ij}^{(0)})]/2$. Amplitude and phase coefficients have different properties: amplitude coefficients are defined on $[0, +\infty)$, while phase coefficients are angles defined on $[0, 2\pi]$ and invariant by 2π translations. Thus, it is clear that different test statistics need to be used for testing the two quantities.

In particular in the uncoupled scenario, for the amplitude coefficients we will rely on the logarithmic distance of geometric sample means:

$$T_{amp}(\boldsymbol{\alpha}^{(k)^*}) = \left|\log\left(\frac{\left(\prod_{i=1}^{n_1}\alpha_{i1}^{(k)^*}\right)^{1/n_1}}{\left(\prod_{i=1}^{n_2}\alpha_{i2}^{(k)^*}\right)^{1/n_2}}\right)\right|.$$

In the coupled case, the same test statistic can be more properly rewritten as:

$$T_{amp}(\boldsymbol{\alpha}^{(k)^*}) = \left|\log\left(\prod_{i=1}^{n_1}\frac{\alpha_{i1}^{(k)^*}}{\alpha_{i2}^{(k)^*}}\right)^{1/n_1}\right|.$$

Instead, for testing the phase coefficients in the uncoupled scenario, we will rely on the signed geodesic distance (on the circle S^1) between the geodesic sample means:

$$T_{ph}(\boldsymbol{\phi}^{(k)^*}) = \text{sign}(m_{geo}(\boldsymbol{\phi}_2^{(k)^*}) - m_{geo}(\boldsymbol{\phi}_1^{(k)^*}))d_{geo}(m_{geo}(\boldsymbol{\phi}_1^{(k)^*}), m_{geo}(\boldsymbol{\phi}_2^{(k)^*})).$$

In the coupled case, we will use the geodesic sample mean of the signed geodesic distances:

$$T_{ph}(\boldsymbol{\phi}^{(k)^*}) = m_{geo}[\{\text{sign}(\phi_{i2}^{(k)^*} - \phi_{i1}^{(k)^*})d_{geo}(\phi_{i1}^{(k)^*}, \phi_{i2}^{(k)^*})\}_{i=1,\ldots,n_1}],$$

with the signed geodesic distance and the geodesic sample mean defined according to:

$$d_{geo}(\phi_1, \phi_2) = min\{|\phi_1 - \phi_2|, |2\pi - (\phi_1 - \phi_2)|\}, \quad \phi_1, \phi_2 \in [0, 2\pi);$$

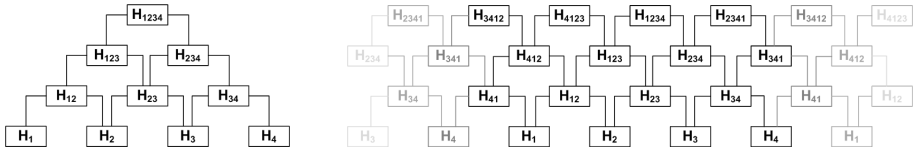
$$m_{geo}(\phi_1, \phi_2, \ldots, \phi_q) = \underset{\phi}{\text{argmin}} \sum_{l=1}^{q} [d_{geo}(\phi_l, \phi)]^2 \quad \phi_i \in [0, 2\pi);$$

$$\text{sign}(\phi_2 - \phi_1) = \begin{cases} +1 & \text{if } 0 \leq \phi_2 - \phi_1 \leq \pi \text{ or } -2\pi \leq \phi_2 - \phi_1 < -\pi \\ -1 & \text{if } \pi < \phi_2 - \phi_1 \leq -2\pi \text{ or } \pi < \phi_2 - \phi_1 \leq 2\pi \end{cases}.$$

2.3 Third Step: Interval-wise Combination and Correction

The third step of the ITP consists in the construction of suitable combinations of the univariate test statistics in order to obtain an interval-wise control of the FWER. Our proposal is to combine the p univariate test statistics by means of multivariate non parametric combinations (i.e., NPC). In the following, as an example, we will illustrate how to obtain a bivariate test from two univariate tests along the NPC philosophy. The extension to multivariate NPC's is straightforward. It is however detailed in Pesarin and Salmaso (2010).

Let us indicate with $T_0^{(1)^*}$ and $T_0^{(2)^*}$ the observed values of the two univariate statistics related to variables X_1 and X_2 and with $T_b^{(1)^*}$ and $T_b^{(2)^*}$ the values induced by the permutation b of the bivariate data set containing the realizations of (X_1, X_2). Then, select a combining function, i.e., a continuous non increasing function $\psi : [0,1]^2 \to \mathbb{R}$ which is symmetric on the two arguments and attain its supremum value when at least one argument attains zero. The Fisher combining function $\psi(x_1, x_2) = -2(\log x_1 + \log x_2)$ is an example. Finally define $T_b^{(1,2)^*} = \psi(L_b^{(1)}, L_b^{(2)})$ where $L_b^{(1)}$ and $L_b^{(2)}$ are the marginal survival functions of the the two test statistics $T^{(1)}$ and $T^{(2)}$ evaluated in $T_b^{(1)^*}$ and $T_b^{(2)^*}$, respectively. Analogously define $T_0^{(1,2)^*} = \psi(L_0^{(1)}, L_0^{(2)})$. The p-value of the joint bivariate test is now simply defined as the proportion of permutations providing $T_b^{(1,2)^*} > T_0^{(1,2)^*}$. Note that $L_0^{(1)}$ and $L_0^{(2)}$ coincide with the p-values of the two original tests. In the practice, the marginal survival functions (and the descending p-values) can be estimated by means of a conditional MC



(a) Unrecycled (b) Recycled

Figure 2: Example (with $p = 4$) of the family of multivariate tests explored by the ITP: unrecycled version on the left and recycled version on the right.

(i.e., just B randomly selected permutations are used). In this case we have: $\hat{L}_b^{(1)} = \frac{\sum_{q=1}^{B} \mathbb{I}(T_q^{(1)^*} \leq T_b^{(1)^*}) + 1/2}{B+1}$ and $\hat{L}_b^{(2)} = \frac{\sum_{q=1}^{B} \mathbb{I}(T_q^{(2)^*} \leq T_b^{(2)^*}) + 1/2}{B+1}$. The p-value of the joint test is of course estimated by $\hat{L}_0^{(1,2)} = \frac{\sum_{q=1}^{B} \mathbb{I}(T_q^{(1,2)^*} \leq T_0^{(1,2)^*}) + 1/2}{B+1}$. For further details about NPCs procedure please refer to Pesarin and Salmaso (2010).

Applying this procedure to all couples of subsequent coefficients, then to all triplets of subsequent coefficients and so on progressively exploring larger intervals of coefficients, up to the global test obtained combining all p coefficients, we obtain a family of tests with their associated p-values (e.g., Figure 2(a)). Finally, we obtain the corrected p-value for the kth coefficient by associating to the kth coefficient the maximum p-value observed over the p-values of all tests of the previous family whose null hypothesis implies $H_0^{(k)}$: the univariate test for $H_0^{(k)}$; the bivariate tests for $H_0^{(k-1)} \cap H_0^{(k)}$ and for $H_0^{(k)} \cap H_0^{(k+1)}$; the three-variate tests for $H_0^{(k-2)} \cap H_0^{(k-1)} \cap H_0^{(k)}$, for $H_0^{(k-1)} \cap H_0^{(k)} \cap H_0^{(k+1)}$, and for $H_0^{(k)} \cap H_0^{(k+1)} \cap H_0^{(k+2)}$; and so on up to the global p-variate test $\bigcap_{k'=1,\ldots,p} H_0^{(k')}$. Then, using these corrected p-values to detect the components respect to which the two functional populations are significantly different, we obtain an inferential procedure (i.e., the ITP) controlling the FWER on any interval of components. This property is proven in Theorem 4.1.

2.4 Remarks

According to the latter combination strategy the hypotheses in the "middle" are tested more times than the ones at the "edges" (in the example of Figure 2(a) with $p = 4$, $H_0^{(2)}$ and $H_0^{(3)}$ are included in 6 tests, whereas the hypotheses $H_0^{(1)}$ and $H_0^{(4)}$ are only tested 4 times). In order to avoid this asymmetry, which may favor the rejection of the hypotheses which are tested less times, one can introduce a correction that consists in recycling the marginal hypotheses at the edges of the structure. The resulting family is represented in an example with $p = 4$ in Figure 2(b). This recycled version of the ITP has two major advantages: (i) each components is tested the same number of times (i.e., $p(p+1)/2$), and (ii) the FWER is controlled not only on intervals but also on their respective complementary sets. This is the implementation of the ITP we will refer to in the two applications. The theoretical results presented in Section 4 are valid for

both the recycled and the un-recycled version of the ITP.

Non parametric combinations can be used to build other family of multivariate test. In the framework hereby depicted, we can think at building two extreme procedure: the Global Testing Procedure (GTP), which is associated to a degenerative family made by the global test only, and the Closed Testing Procedure (CTP), which is associated to the family made by all $2^p - 1$ possible multivariate tests and whose kth corrected p-values are obtained by computing the maximum p-value observed over the p-values of all tests whose null hypothesis implies $H_0^{(k)}$. The GTP provides a weak control of the FWER (the probability of wrongly rejecting at least one null hypothesis is controlled only if all null hypotheses are true) while the CTP provides a strong control of the FWER (the probability of wrongly rejecting at least one null hypothesis is controlled over any set made of true null hypotheses). Theorem 4.1 proves that the control of the FWER provided by the ITP is intermediate between the two above. Theorems 4.2 and 4.3 instead prove (both globally and component-wise respectively) that the power of the GTP is always higher then the power of the ITP which is always higher than the power of the CTP. The same theorems prove also (both globally and component-wise respectively) that the CTP is always more conservative than the ITP which is always more conservative than the GTP, which is indeed exact. In the supplementary materials a simulation study is performed to explore the tightness of the latter inequalities.

Finally, note that to implement the GTP just one test needs to be performed, p^2 tests are needed for the recycled version of the ITP, and $2^p - 1$ are needed for the CTP. The CTP becomes thus quickly unfeasible for the typical values of p used in FDA.

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

3 The ITP in Different Frameworks

The idea of combining and correcting suitable univariate permutation tests along the line described in Section 2.3 is very general, and may be applied more or less straightforwardly to many other situations, provided that a suitable family of univariate permutation tests is defined. In particular, we describe here the application of the ITP in two situations: the functional ANOVA framework, where the objective is to detect differences among $g > 2$ independent functional populations, and the one population framework, where the objective is testing for the center of symmetry of a symmetric functional population.

3.1 The ITP in the Multi-Population Framework

Suppose to observe a collection of n functions $\mathbf{y} = \{\mathbf{y}_{11}, ..., \mathbf{y}_{n_1 1}, ..., \mathbf{y}_{1g}, ..., \mathbf{y}_{n_g g}\}$ from $g > 2$ different populations and to aim at testing the equality in distribution of all functional populations against the difference in distribution of at least one population from the other ones.

Once again, we calculate for each function the p coefficients of a suitable basis expansion (1). For each component k of the basis, we have a vector of $n = n_1 + n_2 + ... + n_g$ coefficients associated to data $\mathbf{c}^{(k)} = (c_{11}^{(k)}, ..., c_{n_1 1}^{(k)}, ..., c_{1g}^{(k)}, ..., c_{n_g g}^{(k)})$, with $c_{i1} \sim C_1$, $c_{i2} \sim C_2$, ..., $c_{ig} \sim C_g$, and then perform the univariate permutation tests $H_0^{(k)} : C_1^{(k)} \overset{\mathrm{d}}{=} C_2^{(k)} \overset{\mathrm{d}}{=} ... \overset{\mathrm{d}}{=} C_g^{(k)}$ vs $H_1^{(k)} : \exists\, \tau_1, \tau_2 \; s.t. \; C_{\tau_1}^{(k)} \overset{\mathrm{d}}{\neq} C_{\tau_2}^{(k)}$. In detail, the permutations to be used depend on the test structure: in the independent scenario (i.e., functional ANOVA framework), under the null hypothesis, data are completely exchangeable. Consequently the family of transformations is constituted by all possible permutations of the observed values, as in the uncoupled case of the two-populations scenario. In the dependent scenario (i.e., the functional repeated measurements) where the same $n_1 = ... = n_g$ sample units are observed $g > 2$ times, as in the coupled case of the two-population scenario, the groups of values associated to the same sample unit i cannot be split up. Thus, the family of transformations is composed by within and between sample unit permutations of the observed values.

In the independent case, the Fisher's test statistic can be used to perform the marginal tests:

$$T((c_{11}^{(k)}, ..., c_{n_1 1}^{(k)}, ..., c_{1g}^{(k)}, ..., c_{n_g g}^{(k)})^*) = \frac{\sum_{\tau=1}^{g} n_\tau (\bar{c}_\tau^{(k)^*} - \bar{c}^{(k)^*})^2/(g-1)}{\sum_{\tau=1}^{g} \sum_{j=1}^{n_\tau} (c_{j\tau}^{(k)^*} - \bar{c}^{(k)^*})^2/(n-g)},$$

where $\bar{c}^{(k)^*}$ is the sample mean of all permuted coefficients (which is identical to the original sample mean), and $\bar{c}_\tau^{(k)^*}$ is the sample mean of the permuted coefficients associated the group τ.

In the repeated measure scenario, the Hotelling T^2 for $g-1$ independent contrasts can be used:

$$T^2((c_{11}^{(k)}, ..., c_{n_1 1}^{(k)}, ..., c_{1g}^{(k)}, ..., c_{n_g g}^{(k)})^*) = n_1 (\Delta \bar{\mathbf{c}}^{(k)^*})' (\Delta S_{\mathbf{c}}^* \Delta')^{-1} (\Delta \bar{\mathbf{c}}^{(k)^*}),$$

where $\Delta \in \mathbb{R}^{(g-1)\times g}$ is a contrast matrix, $\bar{\mathbf{c}}^{(k)^*} = (\bar{c}_1^{(k)^*}, \ldots, \bar{c}_\tau^{(k)^*})$ is the vector of the sample means of the permuted coefficients and $S_{\mathbf{c}}^*$ is the sample variance-covariance matrix of the permuted coefficients.

3.2 The ITP in the One-Population Framework

Suppose $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ to be a random sample drawn from a symmetric functional population $\mathbf{Y}$ and to aim at testing if the center of symmetry of the functional population is equal to a certain function μ_0. Note that if the mean

of a symmetric functional distribution exists this is identical to its center of symmetry and thus, in the latter case, the test become also a test for the mean function.

As in the previous cases, expand functional data and the m_0 on a suitable basis and represent them through their p coefficients $c_1^{(k)}, ..., c_n^{(k)} \overset{\text{iid}}{\sim} C^{(k)}$:

$$y_i(t) = \sum_{k=1}^{p} c_i^{(k)} \phi^{(k)}(t) \quad \text{and} \quad \mu_0(t) = \sum_{k=1}^{p} c_0^{(k)} \phi^{(k)}(t).$$

By exploiting the linearity of the coefficient representation we can introduce a univariate permutation test for the center of symmetry of each coefficient of the expansion $H_0^{(k)} : \text{center}[C^{(k)}] = c_0^{(k)}$ vs $H_1^{(k)} : \text{center}[C^{(k)}] \neq c_0^{(k)}$. In particular, we need to define a suitable family of permutations and a test statistic.

In this case, the invariant transformations are the reflections of the function y_i with respect to μ_0 (i.e., $y_i \mapsto \mu_0 - (y_i - \mu_0) = 2\mu_0 - y_i$) which, in the coefficient representation, coincide with the simultaneous reflections through $c_0^{(k)}$ of all coefficients $c_i^{(k)}$ of the same unit.

Note that it is not straightforward to test for the center of symmetry when a non linear expansion is used (e.g., the amplitude-phase representation (4) of the Fourier expansion). For this reason, in the Fourier representation, the use of the sine-cosine expansion (3) together with the treatment of the coefficients $(a_i^{(k)}, b_i^{(k)})$ associated to the same frequency k as a bivariate vector is probably the more proper approach as the two coefficients jointly describe the same sinusoid: $(a_1^{(k)}, b_1^{(k)}), ..., (a_n^{(k)}, b_n^{(k)}) \overset{\text{iid}}{\sim} (A^{(k)}, B^{(k)})$. It is thus natural in this case to take a bivariate test for each frequency as the starting tests of the ITP:

$$H_0^{(k)} : \text{center}[(A^{(k)}, B^{(k)})] = (a_0^{(k)}, b_0^{(k)}) \quad \text{vs} \quad H_1^{(k)} : \text{center}[(A^{(k)}, B^{(k)})] \neq (a_0^{(k)}, b_0^{(k)}), \tag{5}$$

where $a_0^{(k)}$ and $b_0^{(k)}$ are the coefficients of the expansion of the function μ_0.

Indeed, the ITP that we will use in the case study is based on the joint reflection of all coefficients vectors $(a_i^{(k)}, b_i^{(k)})$ through the point $(a_0^{(k)}, b_0^{(k)})$, and initialized by the bivariate Hotelling T^2 test statistics:

$$T(\mathbf{a}^{(k)^*}, \mathbf{b}^{(k)^*}) = (\bar{a}^{(k)^*} - a_0^{(k)}, \bar{b}^{(k)^*} - b_0^{(k)})' S_{k,k}^* (\bar{a}^{(k)^*} - a_0^{(k)}, \bar{b}^{(k)^*} - b_0^{(k)}), \tag{6}$$

where $\bar{a}^{(k)^*}, \bar{b}^{(k)^*}$ are the two sample means of the permuted coefficients, and $S_{k,k}^*$ is the sample variance covariance matrix of the permuted coefficient vectors $(a_i^{(k)^*}, b_i^{(k)^*})$.

For the $0-th$ frequency, as we have a unique coefficient $m_i^{(0)}$ for each data, with $m_i^{(0)} \overset{\text{iid}}{\sim} M^{(0)}$, we initialize the ITP through a univariate permutation test $H_0^{(0)} : \text{center}[M^{(0)}] = m_0^{(0)}$ vs $H_1^{(0)} : \text{center}[M^{(0)}] \neq m_0^{(0)}$, based on the squared of the Student t statistic and on the reflections of the coefficient values $m_i^{(0)}$ through the corresponding value $m_0^{(0)}$ derived from the expansion of μ_0.

4 Theoretical Properties of the ITP

In this section, we prove some theoretical results regarding the control of the FWER and the power of the ITP. Note that all results hold for any implementation of the ITP. Indeed, the corresponding proofs exclusively rely on the combination and correction procedure of the p univariate tests described in Section 2.3 and not on the nature of the latter ones. Thus, results depend neither on the specific basis used nor on the test statistics used. Furthermore, being the ITP initialized by univariate tests, results hold for any dimension p of the basis and in particular even for p greater than the sample size. The first result characterizes the control of the FWER provided by the ITP.

Theorem 4.1 *Let us consider an ITP obtained by aggregation of p univariate tests associated to the p components of an ordered basis expansion. Such inferential procedure provides a control of the FWER on all closed intervals of components (i.e., interval-wise control of the FWER).*

Proof. Let $\mathbf{k} = \{k_1, k_2, ...k_d\}$ be a set of indices defining a close interval in $\{1, 2, ..., p\}$. Let $\mathcal{R}^{(k_i)}_{\alpha,ITP}$ be the event "$H_0^{(k_i)}$ is rejected by the ITP at level α" and $\mathcal{R}^{(\mathbf{k})}_{\alpha,ITP} = \bigcup_{k_i \in \mathbf{k}} \mathcal{R}^{(k_i)}_{\alpha,ITP}$ the event "at least one of the $H_0^{(k_i)}$ is rejected by the ITP at level α". Proving the interval-wise control of the FWER of the ITP means proving that, for any $\mathbf{k}$ and for any α, $\mathbb{P}[\mathcal{R}^{(\mathbf{k})}_{\alpha,ITP}] \leq \alpha$ when $\mathbf{H}_0^{(\mathbf{k})} = \bigcap_{k_i \in \mathbf{k}} H_0^{(k_i)}$ is true (i.e., when all $H_0^{(k_i)}$ are true).

Let us indicate with $\mathcal{R}^{(\mathbf{k})}_{\alpha}$ the event "$\mathbf{H}_0^{(\mathbf{k})} = \bigcap_{k_i \in \mathbf{k}} H^{(k_i)}$ is rejected at level α by the corresponding multivariate NPC test". This is the conclusion of the test included in the family of tests explored within the ITP that is derived from the aggregation of the univariate tests for $H_0^{(k_i)}$ with $k_i \in \mathbf{k}$. Thanks to the structure of the ITP, for any $k_i \in \mathbf{k}$, the latter test is among the ones used to correct the k_ith p-value. Thus, the ITP cannot reject $H_0^{(k_i)}$ if the latter test does not reject $\mathbf{H}_0^{(\mathbf{k})}$ (i.e., $\mathcal{R}^{(k_i)}_{\alpha,ITP} \subseteq \mathcal{R}^{(\mathbf{k})}_{\alpha}$). This inclusion holds for all $k_i \in \mathbf{k}$ and thus we have that $\mathcal{R}^{(\mathbf{k})}_{\alpha,ITP} = \bigcup_{k_i \in \mathbf{k}} \mathcal{R}^{(k_i)}_{\alpha,ITP} \subseteq \mathcal{R}^{(\mathbf{k})}_{\alpha}$; and consequently that $\mathbb{P}[\mathcal{R}^{(\mathbf{k})}_{\alpha,ITP}] \leq \mathbb{P}[\mathcal{R}^{(\mathbf{k})}_{\alpha}]$. Finally, due to the exactness of all tests included in the family explored by the ITP, we have that, when $\mathbf{H}_0^{(\mathbf{k})} = \bigcap_{k_i \in \mathbf{k}} H_0^{(k_i)}$ is true, the second term of the latter inequality is equal to α and thus, under the same assumption, that $\mathbb{P}[\mathcal{R}^{(\mathbf{k})}_{\alpha,ITP}] \leq \alpha$. □

In simple 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 detected as false is always less than α. Note that among the controlled intervals, we can find two interesting extreme kinds of intervals: the entire set of components and all single components as well. The former control is known in the literature as "weak control of the FWER", while the latter one as "control of the Comparison-Wise Error Rate".

Our second result compares, relatively the global hypothesis, the ITP with the CTP and the GTP. In details, it ranks these tests in terms of weak control of the FWER and of power.

Theorem 4.2 *Let us consider a CTP, an ITP, and a GTP of level α obtained by aggregation of p univariate tests associated to the p components of an ordered basis expansion. The actual global level of the CTP, of the ITP, and of the GTP (i.e., the probability of rejecting at least one $H_0^{(k)}$ when all sub-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 sub-hypotheses is false) satisfy:

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

Proof. Let $\mathcal{R}^{(k)}_{\alpha,ITP}$ be the event "$H_0^{(k)}$ is rejected by the ITP at level α", $\mathcal{R}^{(k)}_{\alpha,CTP}$ be the event "$H_0^{(k)}$ is rejected by the CTP at level α", and $\mathcal{R}_{\alpha,GTP}$ be the event "$\mathbf{H}_0 = \bigcap_{k=1,\ldots,p} H_0^{(k)}$ is rejected by the GTP at level α". Thanks to the structure of the ITP and of the CTP, all multivariate NPC tests used to correct the kth p-value in the ITP are used to correct the CTP but not viceversa. Moreover, the global test is among the test used in both the CTP and the ITP to correct the kth p-value. Thus, every time the CTP rejects $H_0^{(k)}$ also the ITP rejects it and every time the ITP rejects $H_0^{(k)}$ also the GTP rejects it. Thus we have that $\mathcal{R}^{(k)}_{\alpha,CTP} \subseteq \mathcal{R}^{(k)}_{\alpha,ITP} \subseteq \mathcal{R}_{\alpha,GTP}$, and consequently that $\mathbb{P}[\mathcal{R}^{(k)}_{\alpha,CTP}] \leq \mathbb{P}[\mathcal{R}^{(k)}_{\alpha,ITP}] \leq \mathbb{P}[\mathcal{R}_{\alpha,GTP}]$. Let us now consider the event "at least one of the $H_0^{(k)}$ is rejected by the ITP at level α" (i.e., $\bigcup_{k=1,\ldots,p} \mathcal{R}^{(k)}_{\alpha,ITP}$) and the event "at least one of the $H_0^{(k)}$ is rejected by the CTP at level α" (i.e., $\bigcup_{k=1,\ldots,p} \mathcal{R}^{(k)}_{\alpha,CTP}$). We have that $\bigcup_{k=1,\ldots,p} \mathcal{R}^{(k)}_{\alpha,CTP} \subseteq \bigcup_{k=1,\ldots,p} \mathcal{R}^{(k)}_{\alpha,ITP} \subseteq \mathcal{R}_{\alpha,GTP}$ and thus $\mathbb{P}[\bigcup_{k=1,\ldots,p} \mathcal{R}^{(k)}_{\alpha,CTP}] \leq \mathbb{P}[\bigcup_{k=1,\ldots,p} \mathcal{R}^{(k)}_{\alpha,ITP}] \leq \mathbb{P}[\mathcal{R}_{\alpha,GTP}]$. Now, if the state of nature implies that $\mathbf{H}_0 = \bigcap_{k=1,\ldots,p} H_0^{(k)}$ is true, the left term defines the actual global level of CTP, the second term the actual global level of ITP, and the third one the actual global level of the GTP which is equal to α. Thus, the first thesis is proven.

On the contrary, if the state of nature implies that $\mathbf{H}_0 = \bigcap_{k=1,\ldots,p} H_0^{(k)}$ is false, the left term defines the power of CTP, the second term the power of ITP, and the third one the power of the GTP. Thus, also the second thesis is proven. □

Our third result compares, relatively the component-specific sub-hypotheses, the ITP with the CTP and the GTP. In details, it ranks these tests in terms of Comparison-Wise Error Rate (CWER) and marginal power.

Theorem 4.3 *Let us consider a CTP, an ITP, and a GTP of level α obtained by aggregation of p univariate tests associated to the p components of an ordered basis expansion. The Comparison-Wise Error Rate of the CTP and of the ITP on each component (i.e., the probability of rejecting $H_0^{(k)}$ when the latter is true) satisfy:*

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

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

satisfy:

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

with π_{GTP} the power of the global test.

Proof. Let $k \in \{1, 2, ..., p\}$ be an index referring to the kth component of the basis representation. Let $\mathcal{R}^{(k)}_{\alpha,ITP}$ be the event "$H^{(k)}_0$ is rejected by the ITP at level α", $\mathcal{R}^{(k)}_{\alpha,CTP}$ be the event "$H^{(k)}_0$ is rejected by the CTP at level α", and $\mathcal{R}_{\alpha,GTP}$ be the event "$\mathbf{H}_0 = \bigcap_{k=1,...,p} H^{(k)}_0$ is rejected by the GTP at level α". Thanks to the structure of the ITP and of the CTP, all multivariate tests used to correct the kth p-value in the ITP are used to correct the CTP but not viceversa. Moreover, the GTP is among the test used in both the CTP and the ITP to correct the kth p-value. Thus, every time the CTP rejects $H^{(k)}_0$ also the ITP rejects it and every time the ITP rejects $H^{(k)}_0$ also the GTP rejects it. Thus we have that $\mathcal{R}^{(k)}_{\alpha,CTP} \subseteq \mathcal{R}^{(k)}_{\alpha,ITP} \subseteq \mathcal{R}_{\alpha,GTP}$, and consequently that $\mathbb{P}[\mathcal{R}^{(k)}_{\alpha,CTP}] \leq \mathbb{P}[\mathcal{R}^{(k)}_{\alpha,ITP}] \leq \mathbb{P}[\mathcal{R}_{\alpha,GTP}]$. Now, if the state of nature implies that $H^{(k)}_0$ is true, the left term defines the CWER of the CTP and the second term the CWTR of the ITP. Moreover, being single components special kind of intervals, Theorem 4.1 proves that also the CWTR is controlled by the ITP. Thus, the first thesis $CWER^{(k)}_{CTP} \leq CWER^{(k)}_{ITP} \leq \alpha$ is proven.

On the contrary, if the state of nature implies that $H^{(k)}_0$ is false, the left term defines the marginal power of the CTP, the second term the marginal power of the ITP, and the third one the power of the GTP. Thus, also the second thesis $\pi^{(k)}_{CTP} \leq \pi^{(k)}_{ITP} \leq \pi_{GTP}$ is proven. □

Previous theorems explicit the tradeoff between the control of the FWER and the power both globally (Theorem 4.2) and component-wise (Theorem 4.3). Indeed the weaker control of the FWER of the ITP with respect to the CTP is counterbalanced by the fact that the ITP is less conservative and more powerful (globally and component-wise) than the CTP. On the contrary, the stronger control of the FWER of the ITP with respect to the GTP pays the fact that the ITP is more conservative and less powerful than the GTP. This power loss is anyway countered by a big gain in interpretability of the test results with respect to the GTP. Indeed, differently from the GTP, the ITP is able to highlight the basis elements which the rejection is due to.

5 Analysis of the NASA Temperature Data

In this section we report the analysis of daily temperatures registered by NASA satellites in the region ($45^o - 46^o$ North, $8^o - 9^o$ East) including the city of Milan (Italy) from July 1983 to June 2005 and stored in the NASA database *Earth Surface Meteorology for Solar Energy* (NASA 2008). The aim of this analysis is to test for the mean function of Milan temperature yearly profiles.

In the application, we identify the 22 years available as sample units ($n = 22$) and the 365 records available for each year as 365 point-wise evaluations of the functional data ($J = 365$) (Figure 3), and we aim at testing the mean function of

the functional population which data are assumed to be drawn. Because of the periodic nature of these data and because of their daily resolution we perform an ITP starting from the coefficients of a truncated Fourier expansion (3) of dimension 365 and period T equal to one year.

In particular we aim at selecting, among the frequencies $k = 0, ..., (J - 1)/2 = 182$, the ones whose contribution to the mean function is significantly different from zero. In detail, we applied the ITP as described in subsection 3.2: assuming the functional population to be symmetrically distributed around its mean function, for each frequency $k > 0$ we perform the bivariate test (5), based on the joint changes of the signs of vectors $(a_i^{(k)}, b_i^{(k)})$ and on the Hotelling T^2 statistic (6); for the $0 - th$ frequency, as discussed in Section 3, we perform a univariate permutation test based on the squared of the univariate Student t statistic and on the change of the signs of the coefficients $m_i^{(0)}$. Finally, we obtain the p-value heat-map (top panel of Figure 3) by combine the tests mentioned above as shown in subsection 2.3 relying on the Fisher combination function. In the top panel of Figure 3, we represent the result of each test included in the family explored by the ITP. In particular, the horizontal axis is associated to the interval central frequencies and the vertical one to the amplitudes of the tested interval. Each pixel of the image represents a single multivariate test and its color represents the corresponding p-value (blue corresponds to low p-values and yellow to high p-values). Please remember that p-value heat-map is periodic in the horizontal direction.

Following the correction procedure described in subsection 2.3, the mean contribution of the kth frequency is detected as significantly different from zero at level α if all tests in the family explored by the ITP associated to intervals including the kth component provide a significant result at level α (i.e., if, in the p-value heat-map, all p-values lying in the upsidedown cone with vertex in correspondence with the univariate test for the kth frequency are less than α; the corrected p-value for the kth frequency is indeed exactly the maximum of those p-values). For convenience, the central panel of Figure 3 reports for each frequency its corrected p-value. According to the corrected p-values, just the first two frequencies (i.e., the constant term and the sinusoids of period one-year) contribute significantly to the mean function. The ITP thus suggests an easy description of the mean function as a vertically translated sinusoid of period one year. In detail, this sinusoid is characterized by an annual average temperature of 9.023°C and an annual excursion of 21.771°C. Thanks to this reduced representation we can also estimate the 18^{th} January as the coldest day of the year with a mean temperature of -1.891°C and the 20^{th} July as the hottest day of the year with a mean temperature of 19.880°C.

To appreciate the information provided by the ITP, in the lower panel of Figure 3, together with the original data (dashed light lines), we report the sample mean (bold solid red line), that would be the estimate suggested by the GTP, and the sample mean restricted just to the zero-th and the first frequencies

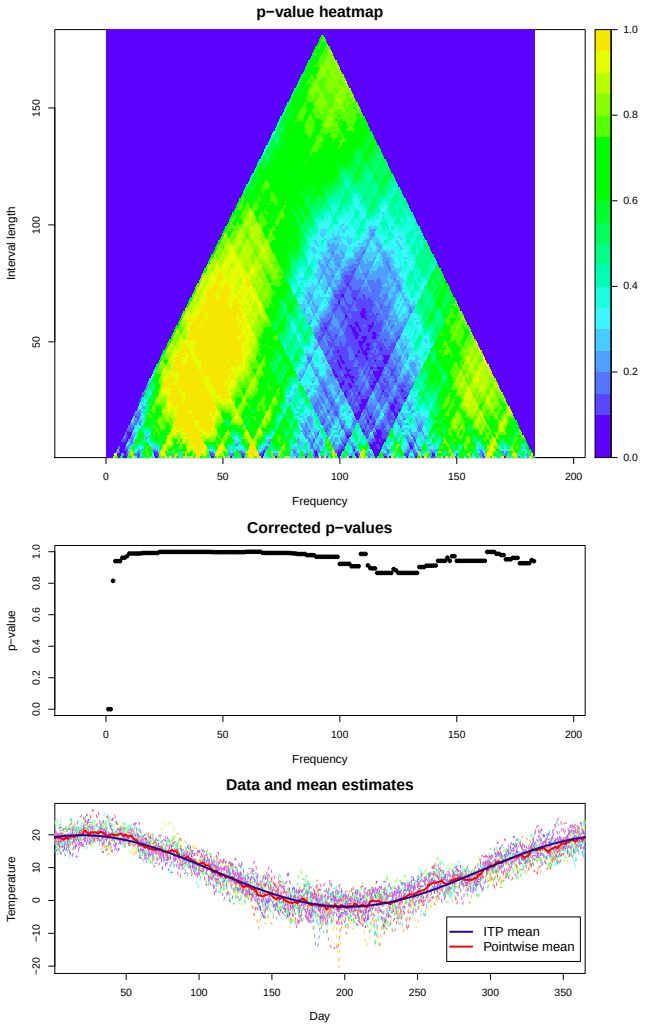


Figure 3: NASA case study. Top : p-values heat-map of the ITP. Center: corrected p-values provided by the ITP. Bottom: curves of daily temperatures data (dashed light lines), sample mean (bold solid red line), and mean as estimated according to the ITP results (bold solid blue line).

(bold solid blue line), that is the estimate of the mean suggested by the ITP. Note how the high-frequency fluctuations that characterize the sample mean (clearly related to the specific sample at hand) are instead not present in the second estimate, as considered not significant by the ITP.

As a comparison with other inferential procedures that can be applied to the coefficients of the basis expansion, let us mention the fact that: the CTP is not feasible for $p = 365$ (i.e., more than 10^{109} tests would be needed). The global test of course rejects the null hypothesis that the function population is centered on zero but it cannot detect which frequencies are not centered on zero. Finally, like the ITP, both the Bonferroni-Holm and the Benjamini-Hochberg corrections (Benjamini and Hochberg 1995) of the univariate tests detect just the zero-th and the first frequencies as not centered on zero. Note that the latter correction procedures obtain corrected p-values by comparing the p-values of univariate tests while the ITP obtains the corrected p-values by comparing the p-values of a family of multivariate tests (i.e., the ones related to intervals) thus exploiting possible dependencies among components.

As a final comment, note that the Fourier expansion of temporal signals is common practice in engineering. Nevertheless, in that field, important frequencies are detected by means of amplitude thresholding and/or frequency filters tuned according to some specific knowledge about the physics (typical amplitude and frequencies of the signal) and/or about the instruments (typical amplitude and frequencies of the noise). The selection criterion derived by the application of the ITP is instead purely statistical and exclusively relies on the observed signals, and it can thus be applied also in context not provided with any quantitative prior knowledge about the problem.

6 Analysis of the Aneurisk Data Set

In this section we present the analysis of the Aneurisk Project data set (Sangalli et al. 2009a,b), which deals with the geometrical and hemodynamical features of the internal carotid arteries (ICA) of patients affected by a cerebral aneurysm. The data set is freely available at http://ecm2.mathcs.emory.edu/aneuriskweb.

The aim of this analysis is to assess whether the geometry and/or the hemodynamics of the internal carotid artery can be 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 - as functions of the arch-length - between subjects affected by a severe form of the pathology (i.e., upper group, 25 subjects with an aneurysm in the upper part of the brain within the skull) and subjects affected by a minor form of the pathology or healthy (i.e., lower group, 25 subjects with an aneurysm in the lower part of the head outside the skull or without any aneurysm). A detailed description of data gathering and processing can be found in Passerini et al. (2012). Data of radius, curvature, and WSS are reported in the bottom left and right panels of Figure 4. Upper

group functions are reported in blue while the lower group ones in red.

In detail, we perform three separated analyses for the radius, curvature, and WSS functions, respectively, and we implement the uncoupled ITP for the differences between two independent functional populations (Section 2), using the B-spline basis representation, the difference between the two sample means of the coefficients as univariate test statistics, and the Fisher combination function. The results hereby presented refer to $p = 128$ uniformly spaced B-splines of order $m = 3$. Similar results are obtained by reducing the order m of the basis and varying (up to some extent) its dimension p.

The p-value heat-maps resulting from the radius, curvature, and wall shear stress tests are reported in the top panels of Figure 4, respectively, whereas the corresponding corrected p-values are reported in the central panels. The three ITP's, at level $\alpha = 5\%$, do not detect any statistical difference between the upper and lower groups pertaining neither the radius nor the curvature functions while a difference in terms of wall shear stress is detected. In particular, as the B-spline basis is local, by looking at the supports of the basis functions, we can identify the interval where the difference is detected as the interval $(-2.783, -1.632)$ (gray region in the lower panels of Figure 4): lower WSS for very severe subjects (i.e., upper group) while higher WSS for less severe subjects (i.e., lower group). Note that thanks to the interval-wise control of the FWER provided by the ITP, we can state that if in the interval $(-2.783, -1.632)$ no differences in distribution between the two population were present, the probability of detecting as significant at least part of the interval would be less than 5%.

Hemodynamics could explain this finding: the latter region corresponds indeed to the second bend of the ICA (i.e., the segment of the ICA where a second peak of curvature is present and where the ICA becomes getting narrower). The bends of the ICA are indeed "guardians" of the arteries of upper part of the brain, which are among the weakest in the entire body (being the latter ones not surrounded by any muscular tissue). Thanks to the passage through the bends the unsteady blood flow from the heart is made steadier before entering the brain. This "stabilizing" effect is related to the loss of energy which is in turn related to the magnitude of the wall-shear-stress within the bends.

To show in a simple way the results of the ITP, in the lower panels of Figure 4, in the region where a significant difference is detected between the two groups, the two sample means of the upper and of the lower group are plotted in bold and the overall sample mean is dotted. The opposite notation is used in the regions where no significant differences are detected. Finally, focussing on the WSS functions is possible to appreciate how the ITP takes into account the local variability. Indeed, despite the enhanced difference between two sample mean functions with respect to the gray-colored part and because of the higher variability occurring at the end of the ICA, the very last part of the ICA is not detected as significant.

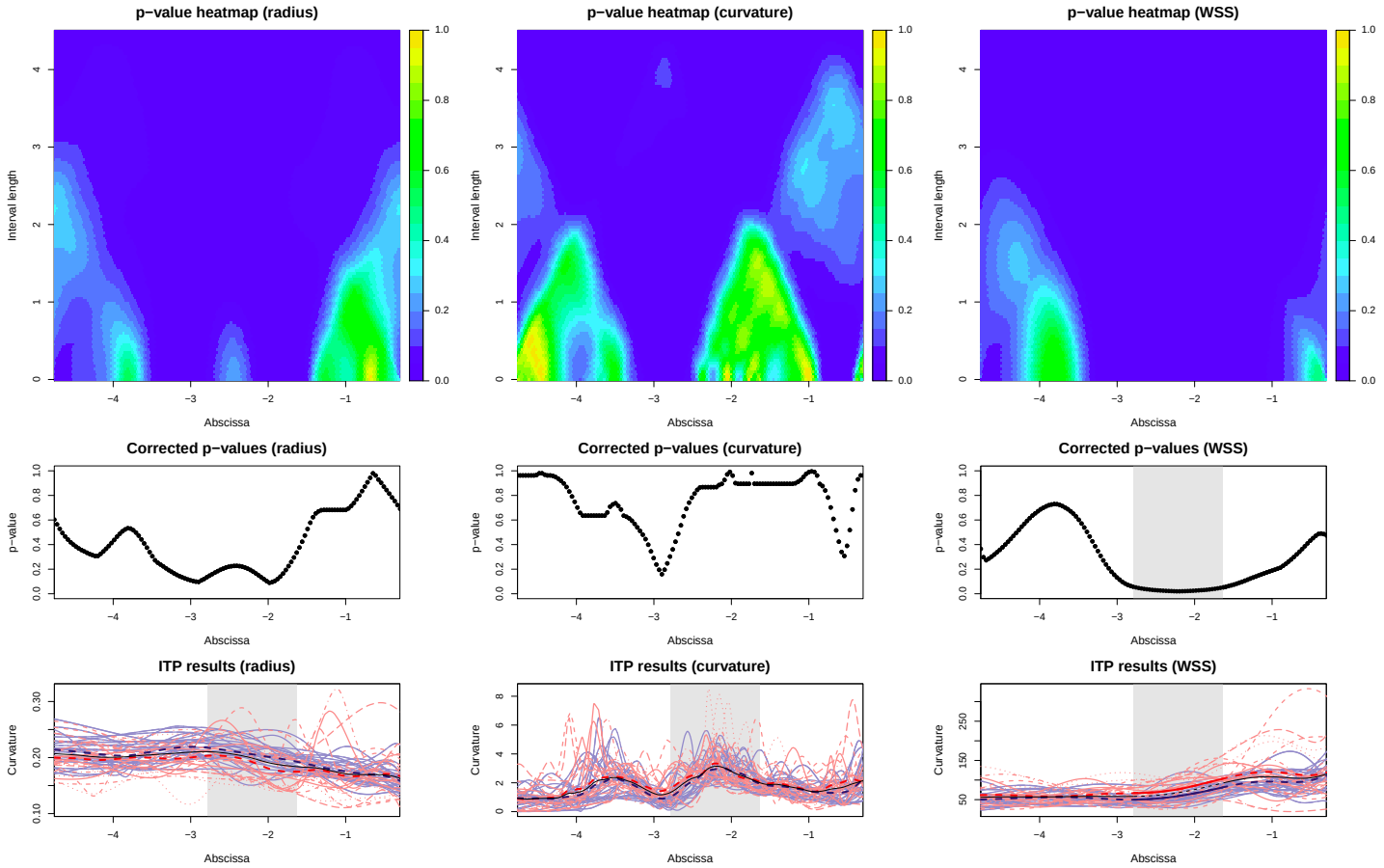


Figure 4: Aneurisk case study analysis of radius (left), curvature (center) and WSS (right). Top: p-value heat-maps; center: corrected p-values; bottom: curves of the upper and lower groups (blue and red, respectevely), sample means associated to the two groups (bold blue and red curves) and global sample means (bold black curves). The shaded part indicates the interval where significant differences area found in terms of WSS.

Similarly to the previous application, also in this case the CTP remains unfeasible (i.e., more than 10^{38} tests would be needed) and the GTP on WSS rejects the null hypothesis of no difference in the WSS between the upper and the lower group but it cannot detect in which part of the carotid this difference is shown. Differently from the previous application, the Bonferroni-Holm correction seems less powerful indeed it is not able to detect any difference between the two groups, while the Benjamini-Hochberg correction detect, in this case, a larger interval than the one detected by the ITP (i.e., the interval $(-3.239, -1.210)$). This latter finding can be explained by the fact that the Benjamini-Hochberg correction has only a weak control of the FWER.

7 Conclusions

We presented a novel inferential procedure suited for functional data analysis (FDA) and based on permutation tests. The procedure, named Interval Testing Procedure (ITP), involves three steps: (*i*) representing functional data on a suitable high-dimensional ordered functional basis; (*ii*) jointly performing univariate permutation tests on the coefficients of the expansion; (*iii*) combining the univariate tests obtaining a p-value heat-map to be used to correct the univariate p-values. The procedure is very general and it can be easily declined to deal with several inferential problems occurring in FDA: for example, the comparison of two or more functional populations, or testing for the mean function of a functional population.

In particular, in this work, we introduced the concept of interval-wise control of the Family Wise Error Rate (FWER) which is particularly meaningful in the framework of FDA and which the ITP is provided with. In detail, interval-wise control of the FWER refers to the property of controlling the FWER over all sets of subsequent coefficients of the basis expansion, meaning that, for any interval of coefficients, if there is no difference in distribution between the investigated populations the probability of incorrectly detecting as significant at least one coefficient of the interval is controlled. For instance this control, which lies in between the weak and the strong control of the FWER, if associated to a B-spline expansion implies that, given any interval of the domain in which there is no difference between the two functional populations, the probability that at least a part of the domain is wrongly detected as significant is always controlled.

In addition to having proved the interval-wise control property of the ITP, we also proved that the marginal and global statistical power of the ITP is always higher than the one provided by the Closed Testing Procedure (which provides a strong control of the FWER but it is computationally unfeasible in the functional framework). On the contrary, we proved that the marginal and global power of the ITP is always lower than the Global Testing Procedure one (which however provides only a weak control of the FWER and does not provide any guide to the interpretation of the test result).

Finally, we reported the application of the ITP to two case studies to show the potential of the ITP in the practice. In the detail, we performed a Fourier-based inference for the mean function of yearly recorded daily temperature profiles in Milan, Italy; and B-spline-based inference for the difference between wall shear stress profiles along the Internal Carotid Artery of two pathologically-different groups of subjects. In both applications we compared the findings highlighted by the ITP by the ones pointed out by the Bonferroni-Holm correction procedure (which provides a strong control of the FWER) and the Benjamini-Hochberg correction procedure (which provides a control of the False Discovery Rate and thus just a weak control of the FWER). In both applications, the ITP turned out to be equally or more powerful than the Bonferroni-Holm procedure and comparable with the Benjamini-Hochberg correction procedure.

A deeper comparison with the Bonferroni-Holm procedure and the Benjamini-Hochberg procedure has been carried out through a simulation study reported in the supplementary materials. The major finding that can be drawn from simulations is that, if one is just interested in maximizing the power under a weak control of the FWER: the use of the Benjamini-Hochberg correction procedure or of the Bonferroni-Holm correction procedure is suggested when the "false" components are expected to be sparse and isolated across components; while the use of the ITP is suggested when the latter ones are expected to be grouped in intervals or bands.

An R-package (`fdatest`) implementing the ITP is available as supplementary material. The current version of the package requires functional data evaluated on a uniform grid; it automatically projects each function on a chosen functional basis; it performs the entire family of multivariate tests; and, finally, it provides the matrix of the p-values of the previous tests and the vector of the corrected p-values. The functional basis, the coupled or uncoupled scenario, and the kind of test can be chosen by the user. The package provides also a plotting function creating a graphical output like the ones presented in Figures 3 and 4: the p-value heat-map, the plot of the corrected p-values, and the plot of the functional data.

8 Supplementary Material

Simulation Studies An extensive simulation study comparing the performances of the ITP, CTP, GTP, Bonferroni-Holm procedure and Benjamini Hochberg procedure, articulated in three parts: (i) comparison among the FWER control of the procedures; (ii) comparison among the global test and global power of the procedures; (iii) comparison among the CWER and marginal power of the procedures. (pdf file)

fdatest package An R package implementing the ITP for one or two populations of functional data evaluated on a uniform grid. The package also contains all data sets used as examples in the article. (zipped file)

References

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

D. Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer, 2000.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

R. Marcus, P. Eric, and K.R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

NASA. Surface meteorology and solar energy, a renewable energy resource web site (release 6.0). http://eosweb.larc.nasa.gov, 2008. Last visit: 20/09/2012.

T. Passerini, L.M. Sangalli, S. Vantini, M. Piccinelli, S. Bacigaluppi, L. Antiga, E. Boccardi, P. Secchi, and A. Veneziani. An integrated statistical investigation of internal carotid arteries of patients affected by cerebral aneurysms. *Cardiovascular engineering and technology*, pages 1–15, 2012.

F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons Inc, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

J.O. Ramsay and B.W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer, 2002.

J.O. Ramsay and BW Silverman. *Functional Data Analysis*. Springer, New York, 2005.

L.M. Sangalli, P. Secchi, S. Vantini, and A. Veneziani. A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, 104(485):37–48, 2009a.

L.M. Sangalli, P. Secchi, S. Vantini, and A. Veneziani. Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(3):285–306, 2009b.

P. Secchi, A. Stamm, and S. Vantini. Large p small n data: Inference for the mean. Technical Report 06, MOX, 2011.

MS Srivastava. Multivariate theory for analyzing high dimensional data. *J. Japan Statist. Soc*, 37(1):53–86, 2007.

9 Simulation Study

The aim of the simulation studies here reported is to investigate the performances of the ITP. In particular, we want to explore the tightness of the inequalities reported in theorems 4.1, 4.2 and 4.3, as the number of false hypotheses increases. The simulation study is therefore divided into three parts, each of them being related to one theorem: in the first part we compare the FWER of different testing procedures, in the second part their global level and power, and in the third part their CWER and marginal power. In each part of the simulation, we compare the ITP with the CTP and GTP (the GTP is considered only in the second part, as both the FWER and the component-wise statistics are not defined for this procedure). In addition, we compare the ITP with other multiple testing strategies, such as the Bonferroni-Holm and the Benjamini-Hochberg procedures.

In the entire simulation study, we consider a coupled test for the differences between two populations in an 8-dimensional space. Let $c_1^{(1)}, c_1^{(2)}, .., c_1^{(8)}$ be the random coefficients associated to units of the first population, and $c_2^{(1)}, c_2^{(2)}, .., c_2^{(8)}$ the random coefficients associated to units of the second population. We generate for each $k \in \{1, 2, ..., 8\}$ the differences between coupled coefficients from a normal distribution, with mean $\mu^{(k)} \in \{0, 1\}$ and standard deviation $\sigma^{(k)} = 1$. The different components are generated independently, i.e., the variance covariance matrix of the 8-dimensional vector of differences is $\Sigma = \sigma^2 I$. Other simulations have been performed with a different choice for Σ, showing that the described results do not change considering a more complicate covariance structure. Finally, we suppose to observe $n_1 = n_2 = 10$ different realizations from the two populations.

We consider the simultaneous test of $p = 8$ hypotheses on the 8 corresponding independent differences. The corresponding tests are $H_0^{(k)} : \mu^{(k)} = 0$ vs. $H_1^{(k)} : \mu_i \neq 0, \forall k \in \{1, 2, ..., 8\}$. In particular, if $\mu^{(k)} = 0$, $H_0^{(k)}$ is true and on the contrary if $\mu^{(k)} = 1$, $H_0^{(k)}$ is false. The truth values of the 8 hypotheses will change from one scenario to another. Nine scenarios are explored with the kth scenario characterized by the first k null hypotheses being false and the last $8-k$ being true, with $k = 0, 1, \ldots, 8$.

9.1 Comparison of the FWER

In Figure 5 we report, for each scenario and each testing procedure, the estimated FWER. We notice from the simulation results that, coherently with the theory, the ITP, CTP and Bonferroni-Holm procedure control the FWER on intervals. On one hand, the ITP is less conservative than the CTP in all scenarios. On the other hand, the ITP and Bonferroni-Holm procedures seem to have an opposite behavior. In addition, simulation shows that, as expected, the Benjamini-Hochberg procedure does not control the FWER.

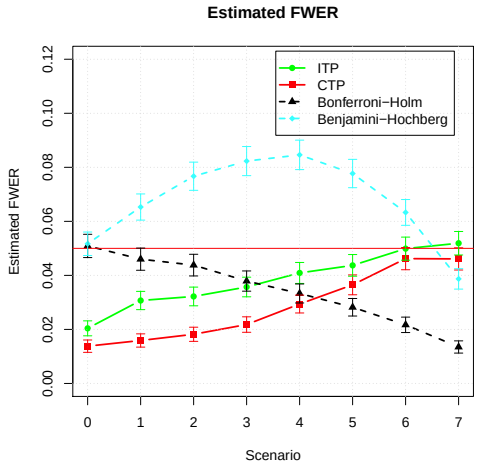


Figure 5: FWER of the considered multiple testing procedures as the number of false hypotheses increases. The error bars indicates the 95% confidence interval for the real FWER.

9.2 Comparison of the global level and power

Figure 6 reports the estimated global probability of rejection of each procedure and each scenario. In particular, in the scenario zero, with no false hypotheses, the probability of rejection is the global level, and in all other scenarios it is the global power. Here, we notice that all procedures control the global level of the test. In terms of power, the ITP seems to behave more similarly to the CTP when the number of false hypotheses is low, and more similarly to the GTP when the number of false hypotheses is high. Moreover, the Bonferroni-Holm procedure and the Benjamini-Hochberg procedure seem to outperform the ITP when the number of true hypotheses is large and the number of false hypotheses is low, while the opposite occurs in the opposite case.

9.3 Comparison of the CWER and marginal power

Figure 7 reports the component-wise probability of rejection. In particular, on each panel we report a different scenario, and on the abscissa of each panel we report the 8 different components. In addition, the shaded gray part of each panel indicates the false hypotheses on the corresponding scenario. Thus, the CWER is the probability of rejection of each true null hypothesis (that is, the values reported in the white part of each graph), and the marginal power is the probability of rejection of each false null hypothesis (that is, the values reported in the gray part of each graph).

The simulation shows that all procedures assure the control of the CWER in each scenario. As confirmed by the theory, the ITP outperforms the CTP in terms of marginal power, and the difference between the power of the two procedures depends on the component. Indeed, we notice that the CTP, Bonferroni-

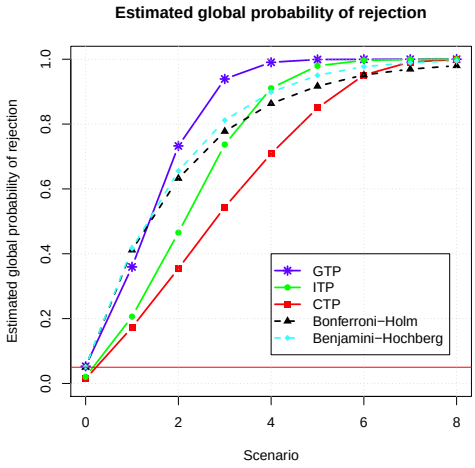


Figure 6: Estimated global probability of rejection for the considered multiple testing procedures as the number of false hypotheses increases.

Holm and Benjamini-Hochberg do not make any distinction among the different components, whereas the ITP does. In particular, the ITP maximizes the power at the center of the intervals of false hypotheses, by exploiting the ordered structure of the components. Consequently, the ITP seems to outperform the Benjamini-Hochberg procedure and (even more) the Bonferroni-Holm procedure on all false hypotheses not occurring at the boundaries between "true" and "false" regions. If one was just interested in the weak control of the FWER, this latter finding could suggest the use of the Benjamini-Hochberg correction procedure or of the Bonferroni-Holm correction procedure when the "false" components are expected to be mostly sparse and isolated while the use of the ITP when the latter ones are expected to be mostly grouped in intervals.

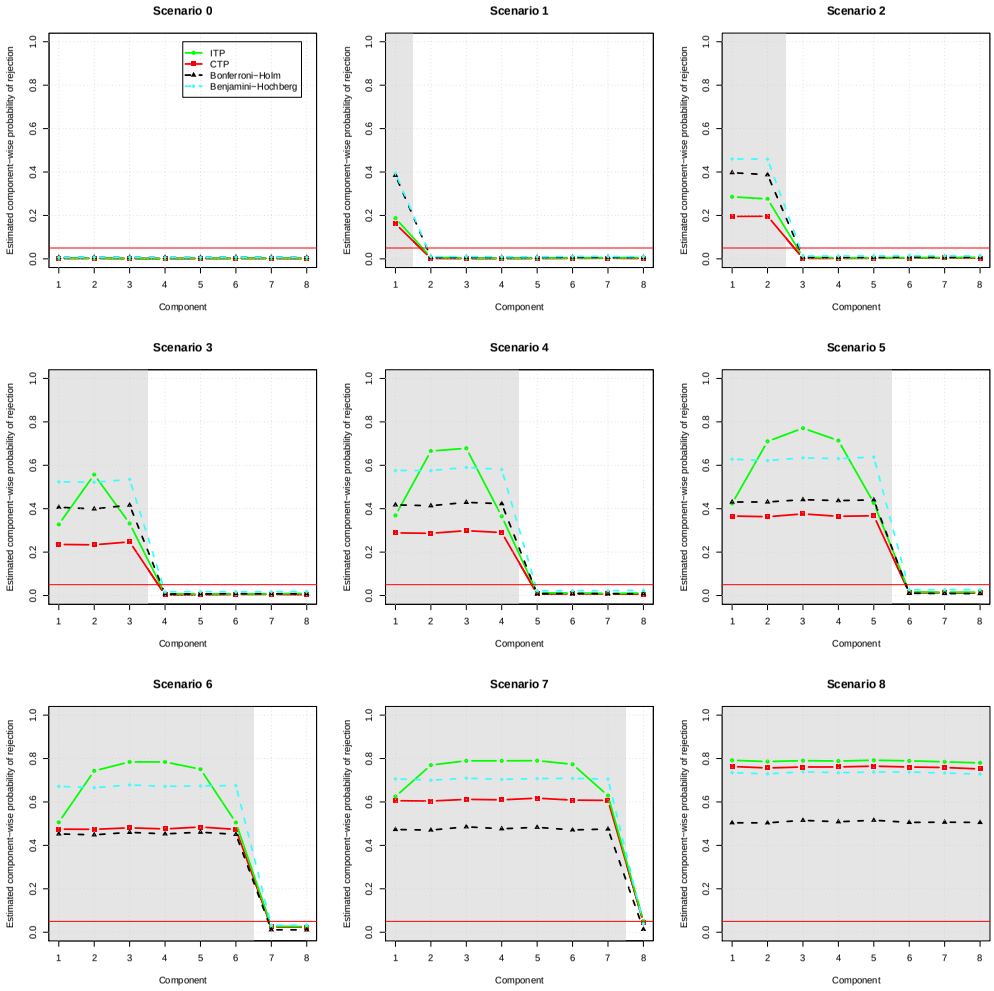


Figure 7: Estimated component-wise probability of rejection for the considered multiple testing procedures on each scenario.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

13/2013 Pini, A.; Vantini, S.
The Interval Testing Procedure: Inference for Functional Data Controlling the Family Wise Error Rate on Intervals.

12/2013 Antonietti, P.F.; Beirao da Veiga, L.; Bigoni, N.; Verani, M.
Mimetic finite differences for nonlinear and control problems

11/2013 Discacciati, M.; Gervasio, P.; Quarteroni, A.
The Interface Control Domain Decomposition (ICDD) Method for Elliptic Problems

10/2013 Antonietti, P.F.; Beirao da Veiga, L.; Mora, D.; Verani, M.
A stream virtual element formulation of the Stokes problem on polygonal meshes

09/2013 Vergara, C.; Palamara, S.; Catanzariti, D.; Pangrazzi, C.; Nobile, F.; Centonze, M.; Faggiano, E.; Maines, M.; Quarteroni, A.; Vergara, G.
Patient-specific computational generation of the Purkinje network driven by clinical measuraments

08/2013 Chen, P.; Quarteroni, A.; Rozza, G.
A Weighted Reduced Basis Method for Elliptic Partial Differential Equations with Random Input Data

07/2013 Chen, P.; Quarteroni, A.; Rozza, G.
A Weighted Empirical Interpolation Method: A-priori Convergence Analysis and Applications

06/2013 Ded, L.; Quarteroni, A.
Isogeometric Analysis for second order Partial Differential Equations on surfaces

05/2013 Caputo, M.; Chiastra, C.; Cianciolo, C.; Cutri , E.; Dubini, G.; Gunn, J.; Keller, B.; Zunino, P.;
Simulation of oxygen transfer in stented arteries and correlation with in-stent restenosis

04/2013 Morlacchi, S.; Chiastra, C.; Cutr, E.; Zunino, P.; Burzotta, F.; Formaggia, L.; Dubini, G.; Migliavacca, F.
Stent deformation, physical stress, and drug elution obtained with provisional stenting, conventional culotte and Tryton-based culotte to treat bifurcations: a virtual simulation study