Multi-aspect local inference for functional data: analysis of ultrasound tongue profiles

A. Pini[a], L. Spreafico[b], S. Vantini[a] and A.Vietti[b]

[a] MOX– Department of Mathematics, Politecnico di Milano, Milan Italy
alessia.pini@polimi.it
simone.vantini@polimi.it

[b] ALPS - Alpine Laboratory of Phonetic Sciences Free University of Bozen-Bolzano, Bolzano, Italy
lorenzo.spreafico@unibz.it
alessandro.vietti@unibz.it

Abstract

Motivated by the functional data analysis of a data set of ultrasound tongue profiles, we present the multi-aspect interval-wise testing (multi-aspect IWT), i.e., a local non-parametric inferential technique for functional data embedded in Sobolev spaces. multi-aspect IWT is a non-parametric procedure that tests differences between groups of functional data jointly taking into account the curves and their derivatives. The multi-aspect IWT provides adjusted multi-aspect p-value functions that can be used to select intervals of the domain imputable for the rejection of a null hypothesis. As a result, it can impute the rejection of a functional null hypothesis to specific intervals of the domain and to specific orders of differentiation. We show that the multi-aspect p-value functions are provided with a control of the family-wise error rate, and are consistent. We apply the multi-aspect IWT to the functional data analysis of a data set of tongue profiles recorded for a study on Tyrolean, a German dialect spoken in South Tyrol. We test differences between five different manners of articulation of the uvular /r/: vocalized /r/, approximant, fricative, tap, and trill. multi-aspect IWT-based comparisons result in an informative and detailed representation of the regions of the tongue where a significant difference is located. Authors' names are alphabetically ordered.
Keywords: Functional Data Analysis, Inference, Interval-Wise Error Rate, Derivatives, Articulatory Phonetics.

1 Introduction

Speech sounds are produced by a mechanism that we may summarize in three main phases. First, the respiratory system pushes air out of the lungs, thus

providing the energy. This airflow travels the trachea up to the larynx where it passes between two muscular folds, called the vocal folds. Second, in the larynx, the passage of air may set the vocal folds vibrating or not: in the former case we obtain voiced sounds, in the latter voiceless sounds are produced. Third, the signal is modulated by changing the shape of the vocal tract. Tongue, lips, jaw and soft palate, namely the articulators, are actively involved in this "shaping" process, or articulatory process. The dynamic coordination of the articulators result in the production of linguistic sounds, such as consonants and vowels (Ladefoged and Johnson 2011). Articulatory phonetics is the branch of speech sciences, within the main field of linguistics, that studies the production of speech sounds by observing the movements of the articulators (Gick et al. 2013).

In the articulatory process, the tongue plays a central role because of its anatomy and physiology. This organ is composed by a group of highly organized muscle that work together to achieve the articulatory target: four extrinsic muscles act to change the position of the tongue in the mouth and four intrinsic muscles act to change its shape. This "interesting engineering feat" (Seikel et al. 2000, p. 336) allows fast (up to 160 cm/s, Löfqvist 2011) and at the same time fine (Qyarnström et al. 1994) motor movements of the tongue, notwithstanding its massive structure.

This work aims at developing a statistical comprehensive approach to infer if and how the tongue change while different sounds are pronounced by the same speaker. The analysis focuses on different manners of uvular - i.e., produced by the tongue back interacting with the uvula, the small piece of tissue hanging at the back of the palate - articulation of the /r/ sound in the Tyrolean dialect, a German dialect spoken in South Tyrol (Italy). Phonologically the sounds included in the data set are different variants of one abstract underlying category of /r/ sound (called allophones). As a consequence, they are minimally different objects from the articulatory point of view, since they share the same point of articulation in the vocal tract and differ only in the manner of articulation. To this purpose, we analyze a data set of tongue sagittal profiles of five variants of uvular /r/[1]:

r TRILL: it is produced by holding the back of the tongue very close to the uvula so that the airflow between the articulators sets them in motion, alternately sticking them together and moving them apart. The trill variant stands for a consonant.

t TAP: it is produced with a rapid movement of the back of the tongue upward to contact the palato-uvular region, then returning to the floor of the mouth. It is a consonant.

f FRICATIVE: it is produced by constricting airflow through a narrow channel at the place of articulation, causing a turbulent aperiodic sound. It stands

[1]All /r/-variants were classified via a combination of auditory feedback and waveform and spectrogram inspection using the software Praat (Boersma and Weenink 2010)

for a consonant, i.e., there is narrow constriction but not full contact between the tongue and the palate.

a APPROXIMANT: as in the case of *f*, the tongue is close to the uvula but without the tract being narrowed to such an extent that a turbulent sound is produced.

voc VOCALIZATION: the airstream passes through the vocal tract without major obstacles or constriction being posed by the tongue. In this case, vocalization variant stands for a back vowel, i.e., something in between an [a] and an [o].

Data were collected by using ultrasound imaging technique at the Alpine Laboratory of Phonetic Sciences of the Free University of Bozen - Bolzano, Italy (Vietti et al. 2015).

The comparison between the groups of curves can be naturally embedded within the framework of null hypothesis significance testing of functional data. Looking at tongue profiles as functions (e.g., Ferraty and Vieu 2006; Ramsay and Silverman 2002, 2005) has two main advantages: (*i*) the whole structure of data is considered, instead on focusing only at some specific features; and (*ii*) derivatives of the data are straightforwardly defined, and they can be analyzed to provide different insights on the same data.

The literature dealing with null hypothesis significance testing of functional data has pursued different approaches which can be categorized in parametric and non parametric methods on the one hand, and global and local methods on the other one. Parametric methods rely on parametric distributional models (e.g., Gaussianity) to compute the distribution of the test statistic (or statistics) under the null hypothesis while non parametric methods rely on computationally intensive re-sampling techniques (e.g., bootstrapping or permuting) able to bypass the parametric distributional model assessment. Global methods provides the tester with a "simple" rejection or non-rejection of the null hypothesis. Local methods instead restate the testing problem at the functional domain level providing the tester with portions of the domain where the null hypothesis is rejected or not rejected.

The majority of works dealing with inference for functional data rely on global parametric methods (e.g., Spitzner et al. 2003; Cuevas et al. 2004; Abramovich and Angelini 2006; Horváth and Kokoszka 2012; Staicu et al. 2014), but there is a consistent literature pertaining also global non-parametric methods (e.g., Hall and Tajvidi 2002; Cardot et al. 2004; Corain et al. 2014). Recently, some works have been proposed in the framework of local parametric techniques (Abramovich and Heller 2005) and local non-parametric techniques (Cox and Lee 2008; Vsevolozhskaya et al. 2014; Pini and Vantini 2016, 2017).

In this work we consider local non-parametric inferential methods as a starting point, and we extend the literature of this field by introducing the possibility

of jointly testing multiple aspects of the data related to differential information. When dealing with functional data, it is natural to compute derivatives, which can carry a lot of information on the data themselves. Nevertheless, all mentioned inferential techniques just focus on the vertical position of the data, without considering the information carried out by derivatives. The impact of considering differential quantities together with functional data in the analysis is well known the literature. It has been deeply investigated in different areas of it such as smoothing (e.g., Ramsay and Silverman 2005), registration (e.g., Srivastava et al. 2011), or dimensional reduction (e.g., Dalla Rosa et al. 2014; Poyton et al. 2006; Ramsay and Silverman 2005). Nevertheless, derivative-based approaches have been completely (and surprisingly) overlooked by the literature focusing on inference on functional data.

In the literature of non-parametric tests of univariate data, some methods have been proposed to jointly test several aspects of the data (e.g., differences between two populations in terms of mean and variance). This problem is often referred-to as "multi-aspect" testing (Brombin and Salmaso 2009; Pesarin and Salmaso 2010; Salmaso and Solari 2005). In this work, we extend local non-parametric inferential methods and multi-aspect testing in order to jointly exploit all information carried out by the functions and their derivatives. Our proposed technique - namely, multi-aspect interval-wise testing (multi-aspect IWT) - is an inferential tool for functional data able to select the portions of the domain imputable for the rejection of a null hypothesis, and to assess whether the rejection is imputable to specific derivatives (e.g., vertical positions, slopes, or concavities) of the functions.

The paper is structured as follows. In Section 2 we describe the multi-aspect interval-wise testing. The method is first described in the case of testing differences between two functional populations, and then extended to more complex null hypothesis testing problems. Section 3 reports the analysis of the tongue profiles. Section A describes a simulation study assessing the performances of our proposed method, and finally Section 4 draws some conclusions and discuss future developments. All computations and images have been created using R (R Core Team 2016). The procedure and analyses presented in this work have been implemented in the R package `tongue.analysis`.

2 Methods: Multi-Aspect Interval-Wise Testing

We start describing our approach to functional inference by tackling one of the simplest and most frequently encountered null hypothesis testing problem, i.e.: the two-population test. Later in the manuscript we will discuss how to extend the approach to other more complex null hypothesis testing problems (e.g., functional ANOVA).

2.1 Functional Two-population Test

The aim of the multi-aspect IWT in the two-population framework is to test differences between two groups of curves and to select the orders of differentiation and the parts of the domain which are imputable for the rejection of the functional null hypothesis. We embed the testing problem in the space $H^d(T)$ of all real-valued squared-integrable functions on the domain T with squared-integrable derivatives up to order d (where T is an interval of $\mathbb{R}$ of the form (a, b). We first describe how to split the problem by performing a test on each derivative order separately (Subsection 2.1.2), and then discuss how to combine such partial tests to take into account multiplicity (Subsection 2.1.2). Finally, we discuss the theoretical properties of the obtained procedure (Subsection 2.1.3).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on the space $H^d(T)$. Assume that $\{\xi_{1i}\}_{i=1,\dots,n_1} \overset{\text{iid}}{\sim} \boldsymbol{\xi}_1$ and $\{\xi_{2i}\}_{i=1,\dots,n_2} \overset{\text{iid}}{\sim} \boldsymbol{\xi}_2$ are two independent random samples drawn from the random functions $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, respectively, mapping from Ω to $H^d(T)$. We aim at testing - in a nonparametric framework - the null hypothesis of distributional equality of the random functions $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ against the alternative hypothesis of difference in distribution:

$$H_0 : \boldsymbol{\xi}_1 \overset{\text{d}}{=} \boldsymbol{\xi}_2 \quad \text{against} \quad H_1 : \boldsymbol{\xi}_1 \overset{\text{d}}{\neq} \boldsymbol{\xi}_2. \tag{1}$$

With respect to the standard approach, in case of rejection of the null hypothesis, we aim at imputing the rejection to: (i) specific intervals of the domain of functional data, and (ii) specific aspects of functional data that can be naturally conveyed by derivatives.

Our proposal starts from the work by Pini and Vantini (2017) which approaches the problem of testing hypotheses (1) in the $L^2(T)$ setting by defining an adjusted p-value function $\widetilde{p} : T \to [0, 1]$ that can be thresholded to select intervals of the domain imputable of the rejection of H_0. To define the adjusted p-value function, the authors propose to perform a functional test comparing the means of the two populations on every interval $\mathcal{I} \subseteq T$. Then, for each point $t \in T$, the adjusted p-value $\widetilde{p}(t)$ is defined as the supremum of the p-values of all tests pertaining to intervals including t. Finally, the authors prove that this approach provides a control of the so-called interval-wise error rate (IWER), i.e., for every interval of the domain where the null hypothesis is true, the selection procedure allows to control the probability that the interval is wrongly selected.

In this work, to address the problem of testing differences between the two data distributions taking derivatives into account, first we define - for each derivative order - an adjusted *partial* p-value function, which controls the IWER, singularly for each order of differentiation. Then, we define - for each order of differentiation - an adjusted *multi-aspect* p-value function, to jointly control the IWER on derivatives of order $1, \dots, d$.

2.1.1 Adjusted partial p-value functions.

Let $\mathcal{I} \subseteq T$ be a generic interval of the form $[t_1, t_2]$, with $a \leq t_1 < t_2 \leq b$. Consider the restriction of test (1) on interval $\mathcal{I}$:

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \stackrel{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \quad \text{against} \quad H_1^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \stackrel{\mathrm{d}}{\neq} \boldsymbol{\xi}_2^{\mathcal{I}}, \tag{2}$$

being $\boldsymbol{\xi}_j^{\mathcal{I}}$, $j = 1, 2$ the restriction of $\boldsymbol{\xi}_j$ on interval $\mathcal{I}$. A standard functional permutation test can be performed to test (2) by selecting a global test statistic stochastically greater under $H_1^{\mathcal{I}}$ than under $H_0^{\mathcal{I}}$.For instance, Hall and Tajvidi (2002) propose a test statistic based on the L^2 distance between the two sample means $T^{\mathcal{I}} = \int_{\mathcal{I}} \left(\overline{\xi}_1(t) - \overline{\xi}_2(t)\right)^2 \mathrm{d}t$. The test statistic is evaluated under all possible rearrangements (permutations) of the data $\xi_{11}, \ldots, \xi_{1n_1}, \xi_{21}, \ldots, \xi_{2n_2}$ over the sample units, and the p-value is computed as the proportion of permutations leading to a test statistic larger or equal to the one computed on the non-permuted data. The resulting test is exact regardless of the test statistic chosen to compare the two samples and of data distribution (Pesarin and Salmaso 2010). This yields the possibility of testing several aspects of the data distribution by choosing different test statistics, each focusing on a particular deviation from the null hypothesis. Let T_{D^k} denote a test statistic focusing on the order of differentiation $k \in \{1, \ldots, d\}$ (some examples of test statistics T_{D^k} will be discussed in Subsection 2.2). Let $p_{D^k}^{\mathcal{I}}$ be the p-value of the permutation test (2) based on test statistics T_{D^k}. For each order of derivative, an adjusted p-value function can be computed following the line depicted in Pini and Vantini (2017):

$$\widetilde{p}_{D^k}(t) = \sup_{\mathcal{I} \ni t} p_{D^k}^{\mathcal{I}}. \tag{3}$$

The function $\widetilde{p}_{D^k}(t)$ provides information about the kth order of differentiation, and is provided with a control of the IWER. Specifically, $\forall \alpha \in (0, 1)$:

$$\forall \mathcal{I} \subseteq T : H_0^{\mathcal{I}} \text{ true} \Rightarrow \mathbb{P}\left[\forall t \in \mathcal{I}, \widetilde{p}_{D^k}(t) \leq \alpha\right] \leq \alpha. \tag{4}$$

For $k = 0, \ldots, d$ we refer to $\widetilde{p}_{D^K}(t)$ as the adjusted **partial** p-value function of order k.

2.1.2 Adjusted multi-aspect p-value functions.

For any $k = 0, \ldots, d$, the corresponding adjusted partial p-value function $\widetilde{p}_{D^k}(t)$ can be legitimately used to perform the statistical test (1) (Pini and Vantini 2017). We are here not interested in a-priori selecting a specific value of k such to explore a specific order of differentiation but rather in exploring all the $d+1$ orders of differentiation. The driving idea is to simultaneously exploit different test statistics to solve the same null hypothesis testing problem such to have a deeper insight on the rejection of the null hypothesis. If - to achieve this task - a naive thresholding of the $d+1$ adjusted partial p-value functions $\widetilde{p}_{D^k}(t)$ were

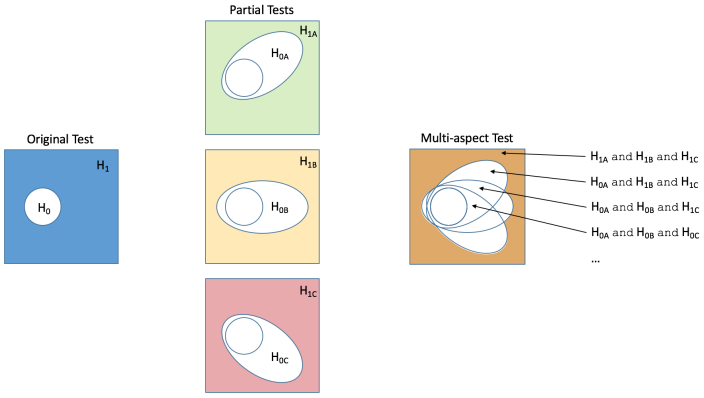

Figure 1: Left: original test of H_0 against H_1; Center: partial tests focusing on three different aspects associated to alternative hypotheses H_{1k} for $k = A, B, C$ s.t. $H_{1k} \subset H_1$; Right: regions identified by means of multi-aspect testing.

performed, the control of IWER would be lost and a biased domain selection criterion would be obtained. Specifically - for each interval where the null hypothesis is true - the probability that at least one among the $d + 1$ adjusted partial p-value functions $\widetilde{p}_{D^k}(t)$ is s.t. $\widetilde{p}_{D^k}(t) \leq \alpha$ would not be controlled. To recover this property and regain the control of the IWER while simultaneously exploring different levels of differentiation we propose to embed the IWT approach in a multi-aspect testing framework based on the Close Testing Procedure (CTP, e.g., Marcus et al. 1976) of partial tests. The integration of the IWT and multi-aspect testing - which will be detailed in the following - will lead to the definition of $d+1$ adjusted **multi-aspect** p-value functions that are used to perform domain selection while controlling the FWER simultaneously on the $d + 1$ levels of differentiation.

To better understand the idea of multi-aspect testing, look at Figure 1. Assume that we were aiming at testing a null hypothesis H_0 against the complementary alternative hypothesis $H_1 = H_0^C$. In addition, assume for instance that we are focusing on differences in three specific aspects of the data distribution, related to the three alternative hypotheses H_{1A}, H_{1B}, and H_{1C}. The key points of multi-aspect tests are: (i) when the original null hypothesis H_0 is true, all partial null hypotheses (i.e., $H_{0A} = H_{1A}^C$, $H_{0B} = H_{1B}^C$, and $H_{0C} = H_{1C}^C$) are true (i.e., for $k = A, B, C$: $H_0 \subset H_{0k}$) and thus any partial test (i.e., H_{0A} vs H_{1A}, H_{0B} vs H_{1B}, and H_{0C} vs H_{1C}) can be used also to test the original null hypothesis H_0 against H_1; (ii) when the partial tests are performed simultaneously the probability of type-I error has to be controlled for the original null hypothesis H_0 (i.e., a control of the Family-wise error rate is required on the entire family of partial tests). The major advantage of multi-aspect tests with respect to a standard test is to provide a better and more detailed insight on the rejection of the original null hypothesis H_0 (i.e., indeed when d partial tests are performed, 2^d possible conclusions are possible). The drawback is basically their computa-

tional burden and a possible lost of power due to both the conservativeness of the procedure used to strongly control the Family-wise error rate and to the fact that the truth of all partial null hypotheses does not necessarily imply the truth of the original null hypothesis (i.e., typically $H_0 \neq \bigcap_{k=A,B,C} H_{0k}$).

Going back to our main goal, which is testing the equality in distribution of two H^d-valued random functions, the multiple aspects that we aim at investigating are differences in the $d+1$ orders of differentiation. For each interval $\mathcal{I}$ we replace the original test (2):

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \quad \text{against} \quad H_1^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{\neq} \boldsymbol{\xi}_2^{\mathcal{I}} \ ,$$

with $d+1$ partial tests, each focusing on a derivative order:

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \quad \text{against} \quad H_{1k}^{\mathcal{I}} : \mathrm{A}_k[D^k \boldsymbol{\xi}_1^{\mathcal{I}}] \neq \mathrm{A}_k[D^k \boldsymbol{\xi}_2^{\mathcal{I}}], \tag{5}$$

with $k = 0, 1, \ldots, d$. In partial tests (5), we introduced the operators A_k which specifically denote the specific distributional aspect of the distribution of $D^k \boldsymbol{\xi}_j^{\mathcal{I}}$ that we aim at testing. A_k maps from the space of $L^2(T)$-valued random functions to the space of $\mathbb{R}$-valued functions on the domain T. Depending on the desired focus, A_k can extract a moment-related function, such as the mean function (i.e., $H_{1k}^{\mathcal{I}} : \mathbb{E}[D^k \boldsymbol{\xi}_1^{\mathcal{I}}] \neq \mathbb{E}[D^k \boldsymbol{\xi}_2^{\mathcal{I}}]$), the variance function (i.e., $H_{1k}^{\mathcal{I}} : \mathrm{Var}[D^k \boldsymbol{\xi}_1^{\mathcal{I}}] \neq \mathrm{Var}[D^k \boldsymbol{\xi}_2^{\mathcal{I}}]$), a higher moment function (e.g., skewness, kurtosis), or alternatively a quantile-related function, such as the median function or the inter-quartile range function. For simplicity, in the rest of the paper we assume that the same distributional aspect is tested for all orders of differentiation (i.e., $\mathrm{A}_0 = \ldots = \mathrm{A}_d$). We will indicate it as A. Note that, without any computational or theoretical difficulty, our proposal can be used also to test different distributional aspects for different orders of differentiation and/or different distributional aspects for the same order of differentiation.

To perform the $d+1$ partial tests (5) we rely on the possibility (provided by permutation tests) of using different test statistics to test the same null hypothesis. We indeed implement $d+1$ permutation tests, specifically, relying on test statistics which are sensitive to specific violations of the null hypothesis $H_0^{\mathcal{I}}$ (i.e., $H_{1k}^{\mathcal{I}}$). Finally, to achieve the strong control of the FWER onto the $d+1$ partial tests we rely on CTP. In detail, one has to simultaneously test H_0 against each possible combination of the partial alternative hypotheses of the family, and reject the null hypothesis against each partial alternative hypothesis only if all joint tests involving that partial alternative hypothesis lead to the rejection of the null hypothesis H_0. Specifically, we perform $2^{d+1} - 1$ IWTs based on "multi-derivative" tests:

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \quad \text{against} \quad H_{1\mathbf{k}}^{\mathcal{I}} : \mathrm{A}[D^{\mathbf{k}} \boldsymbol{\xi}_1^{\mathcal{I}}] \neq \mathrm{A}[D^{\mathbf{k}} \boldsymbol{\xi}_2^{\mathcal{I}}], \tag{6}$$

with $\mathbf{k}$ being any non empty subset of $\{0, 1, \ldots, d\}$ and $D^{\mathbf{k}}$ the linear differential operator mapping each function of $H^d(T)$ in the column vector of the

corresponding derivatives. Let $p^{\mathcal{I}}_{D^{\mathbf{k}}}$ denote the p-value of test (6). The adjusted p-value function $\widetilde{p}_{D^{\mathbf{k}}}(t)$ for test (6) is $\widetilde{p}_{D^{\mathbf{k}}}(t) = \sup_{\mathcal{I}\ni t} p^{\mathcal{I}}_{D^{\mathbf{k}}}$. Finally, the $d+1$ adjusted **multi-aspect** p-value functions $\widetilde{\widetilde{p}}_{D^k}(t)$ are calculated by taking for each order of differentiation the point-wise maximum of all adjusted p-value functions $\widetilde{p}_{D^{\mathbf{k}}}(t)$ involving that order of differentiation, i.e., for $k = 0, 1, \ldots, d$:

$$\widetilde{\widetilde{p}}_{D^k}(t) = \sup_{\mathbf{k}\ni k} \widetilde{p}_{D^{\mathbf{k}}}(t) = \sup_{\mathbf{k}\ni k} \sup_{\mathcal{I}\ni t} p^{\mathcal{I}}_{D^{\mathbf{k}}} \ . \tag{7}$$

2.1.3 Theoretical properties

The following theorems characterize the inferential properties of the adjusted multi-aspect p-value functions $\widetilde{\widetilde{p}}_{D^k}(t)$. All proofs are reported in the Appendix.

Theorem 1. Weak control of the IWER. *Assume that all multi-derivative tests* (6) *of* $H^{\mathcal{I}}_0$ *against* $H^{\mathcal{I}}_{1\mathbf{k}}$ *are exact. Then, if* $H^{\mathcal{I}}_0$ *is true, the* $d+1$ *adjusted multi-aspect p-value functions* $\widetilde{\widetilde{p}}_{D^k}(t)$ *defined in* (7) *are provided with a control of the interval-wise error rate over all* $d+1$ *orders of differentiation simultaneously. In detail,* $\forall \alpha \in (0,1)$*:*

$$\forall \mathcal{I} \subseteq T : H^{\mathcal{I}}_0 \text{ true} \Rightarrow \mathbb{P}\left(\exists t \in \mathcal{I}, \exists k \in \{0,1,\ldots,d\} \ : \ \widetilde{\widetilde{p}}_{D^k}(t) \leq \alpha\right) \leq \alpha.$$

Theorem 2. Strong control of the IWER. *Assume that all multi-derivative tests of* $H^{\mathcal{I}}_{0\mathbf{k}}$ *against* $H^{\mathcal{I}}_{1\mathbf{k}}$ *are exact. Then, the* $d+1$ *adjusted multi-aspect p-value functions* $\widetilde{\widetilde{p}}_{D^k}(t)$ *defined in* (7) *are provided with a strong control of the interval-wise error rate over all* $d+1$ *orders of differentiation simultaneously. Specifically, for all* $\mathcal{I} \subseteq T$*, let* $\mathbf{k}^{\mathcal{I}}_{null} = \{k \in \{0,1,\ldots,d\} \text{ s.t. } H^{\mathcal{I}}_{0k} \text{ is true}\}$ *and* $H^{\mathcal{I}}_{0\mathbf{k}_{null}} = \bigcap_{k\in\mathbf{k}_{null}} H^{\mathcal{I}}_{0k}$*. Then,* $\forall \alpha \in (0,1)$*:*

$$\forall \mathcal{I} \subseteq T : H^{\mathcal{I}}_{0\mathbf{k}_{null}} \text{ true} \Rightarrow \mathbb{P}\left(\exists t \in \mathcal{I}, \exists k \in \mathbf{k}^{\mathcal{I}}_{null} \ : \ \widetilde{\widetilde{p}}_{D^k}(t) \leq \alpha\right) \leq \alpha.$$

Theorem 3. Interval-wise consistency. *Assume that all multi-derivative tests* (6) *of* $H^{\mathcal{I}}_0$ *against* $H^{\mathcal{I}}_{1\mathbf{k}}$ *are consistent. Then, if* $H^{\mathcal{I}}_{1k}$ *is true on the whole interval* $\mathcal{I}$*, the* $d+1$ *adjusted multi-aspect p-value functions* $\widetilde{\widetilde{p}}_{D^k}(t)$ *defined in* (7) *are marginally consistent. In detail,* $\forall \alpha \in (0,1)$*,* $\forall k \in \{1,\ldots,d\}$*:*

$$\forall \mathcal{I} \subseteq T : \forall \mathcal{J} \subseteq \mathcal{I}, H^{\mathcal{J}}_{1k} \text{ true} \Rightarrow \mathbb{P}\left(\forall t \in \mathcal{I}, \widetilde{\widetilde{p}}_{D^k}(t) \leq \alpha\right) \xrightarrow[n_1+n_2\to\infty]{} 1.$$

Theorem 1 states that – even for finite sample sizes – if a thresholding of the adjusted multi-aspect p-value functions $\widetilde{\widetilde{p}}_{D^k}(t)$ is performed at level α, for each interval where the two functional populations $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are identically distributed, the probability of wrongly detecting a difference on at least one order of differentiation in at least one point is guaranteed to be lower than α. Note that, by Theorem 1, the control of the IWER only holds when the null

hypothesis of equality in distribution is not violated. If on the same interval the two populations are not identically distributed, and such difference is imputable to just few derivatives orders, Theorem 1 does not guarantee a control over the subset of derivatives related to true partial null hypotheses. This latter stronger control of the IWER - which is not guaranteed when tests $H_0^{\mathcal{I}}$ against $H_{1\mathbf{k}}^{\mathcal{I}}$ are exact - is instead guaranteed if also tests $H_{0\mathbf{k}}^{\mathcal{I}} := H_{1\mathbf{k}}^{\mathcal{I}^C}$ against $H_{1\mathbf{k}}^{\mathcal{I}}$ are exact (Theorem 2). As we will show in Sections 2.2 and 2.3, in the practice the type of control (i.e., strong or weak) of the IWER provided by the procedure depends on the specific choice for the test statistics and on the particular distributional aspect that is tested. Finally, Theorem 3 guarantees that if on an interval the two populations are different in terms of the k-th derivative, the probability of correctly detecting the complete interval goes to one as the sample size increases.

Remark. In the practice, functional data are the result of a smoothing process of point-wise noisy evaluations of the functional datum. On the one hand, all used smoothing techniques (e.g., penalized and regression splines or local regression) allow to obtain estimates of functional data as regular as one desires (i.e., choosing d as large as one may desire). On the other hand, accuracy in estimating the function and its derivatives dramatically decreases as the order of the estimated derivative increases (e.g., Ramsay and Silverman 2005; Ferraty and Vieu 2006). Hence, the number d of derivatives to be included in the analysis should be determined from a compromise between the wish of including higher order derivatives able to provide new perspectives on the data and the likewise commendable desire of preserving the statistical power of the testing procedure. Indeed, a poor estimation of high order derivatives may lead to extremely large variances in the two samples of derivatives (i.e., under-smoothing) or to annihilate the differences between the two samples of derivatives (i.e., over-smoothing). Unfortunately, a satisfactory compromise between these two extremes is not always likely to exist as the order of differentiation increases. With respect to our proposal, the inclusion in the procedure of high order derivatives that were poorly estimated may possibly lead to an increase of the value of the adjusted multi-aspect p-value functions, thus still preserving the control of the IWER but resulting in a consequent loss of power. From a computational perspective, including high order (and possibly uninformative) derivatives in the analysis would anyhow increase the computational burden of the procedure. The number of multi-derivative tests to be included in the multi-aspect IWT increase indeed exponentially in the number of considered derivatives.

2.2 Comparing means or variances of two functional populations

In the multi-aspect IWT procedure, the original test (1) is replaced by a set of partial tests (5) on a particular distributional aspect of the distribution of the k-th derivatives. The choice of the test statistic for partial tests (5) and multi-derivative tests (6) directly depends on the investigated aspect of the dis-

tribution. In this Subsection we give some possible choices of test statistics depending on the choice of A, and discuss the exactness and consistency properties of the corresponding multi-aspect IWT in the light of Theorems 1, 2, and 3. In detail, we particularly focus on the two following cases:

A = $\mathbb{E}$**.** Applying the multi-aspect IWT for testing differences in the means of the first d orders of differentiation (Subsection 2.2.1);

A = Var**.** Applying the multi-aspect IWT for testing differences in the variances of the first d orders of differentiation (Subsection 2.2.2).

2.2.1 Comparing the means of two functional populations

Assume that we want to perform a comparison between the two populations in terms of means of the first d orders of differentiation:

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \stackrel{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \quad \text{against} \quad H_{1k}^{\mathcal{I}} : \mathbb{E}[D^k \boldsymbol{\xi}_1^{\mathcal{I}}] \neq \mathbb{E}[D^k \boldsymbol{\xi}_2^{\mathcal{I}}], \tag{8}$$

where $\mathbb{E}$ denotes the expectation. For each order of differentiation $k = 0, \ldots, d$, we propose to employ global test statistics which simply integrates over $\mathcal{I}$ the square of the classical asymptotic z-test statistic used in the scalar setting, i.e.:

$$T_{D^k}^{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \int_{\mathcal{I}} \left(\frac{s^2_{D^k \xi_1}(t)}{n_1} + \frac{s^2_{D^k \xi_2}(t)}{n_2} \right)^{-1} \left(\overline{D^k \xi}_1(t) - \overline{D^k \xi}_2(t) \right)^2 \mathrm{d}t. \tag{9}$$

Where the integral is defined in a Lebesgue sense. In equation (9), and for $j = 1, 2$, $\overline{D^k \xi}_j$ and $s^2_{D^k \xi_j}$ are the sample mean function and the sample variance variance function of the kth derivatives $D^k \xi_{ji}$, respectively. The corresponding test statistic for multi-derivative tests is the integrated version of the following multivariate Hotelling's T^2-like statistics:

$$T_{D^{\mathbf{k}}}^{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \int_{\mathcal{I}} \left(\overline{D^{\mathbf{k}} \xi}_1(t) - \overline{D^{\mathbf{k}} \xi}_2(t) \right)' \left(\frac{S_{D^{\mathbf{k}} \xi_1}(t)}{n_1} + \frac{S_{D^{\mathbf{k}} \xi_2}(t)}{n_2} \right)^{-1} \left(\overline{D^{\mathbf{k}} \xi}_1(t) - \overline{D^{\mathbf{k}} \xi}_2(t) \right) \mathrm{d}t, \tag{10}$$

where $\overline{D^{\mathbf{k}} \xi}_j(t)$ and $S_{D^{\mathbf{k}} \xi_j}(t)$ respectively indicate the sample mean function and sample variance-covariance matrix function of the random vector $D^{\mathbf{k}} \boldsymbol{\xi}_j(t)$ for $j = 1, 2$. Note that - for every differentiation order k and every combination of differentiation orders $\mathbf{k}$ respectively - the test statistics $T_{D^k}^{\mathcal{I}}$ and $T_{D^{\mathbf{k}}}^{\mathcal{I}}$ are dimensionless quantities. Hence, they does not depend on the units of measure of the functional data and of the domain.

With such choices partial tests (5) and multi-derivative tests (6) are exact and consistent. Hence, both Theorems 1 and 3 hold, and the resulting multi-aspect IWT is provided with a weak control of the IWER and it is consistent regardless of the functional distributions and sample sizes. For having instead the strong control, the tests of $H_{0k}^{\mathcal{I}}$ against $H_{1k}^{\mathcal{I}}$ and the tests of $H_{0\mathbf{k}}^{\mathcal{I}}$ against

$H^{\mathcal{I}}_{1\mathbf{k}}$ are needed to be exact. Such property is not generally true for permutation tests above regardless of the data distribution. Under $H^{\mathcal{I}}_{0k}$, the two populations does not share necessarily the same distribution, and consequently data are not exchangeable with respect to units. So, the distribution of $T^{\mathcal{I}}_{D^{\mathbf{k}}}$ might not be a uniform discrete. Nevertheless, further knowledge about the specific application may turn out into further distributional assumptions which can be sufficient for obtaining a strong control. A renowned case is the one of **shifted populations**. If we know that the two functional populations differ at most for an additive term Δ (i.e., $\boldsymbol{\xi}_1 \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2 + \Delta$ with $\Delta \in H^d(T)$), we trivially have $H^{\mathcal{I}}_{0k} : D^k\Delta^{\mathcal{I}} = \mathbf{0}$ against $H^{\mathcal{I}}_{1k} : D^k\Delta^{\mathcal{I}} \neq \mathbf{0}$. In this case, it is straightforward to prove that under H_{0k} the $n_1 + n_2$ kth derivatives are exchangeable. Being the test statistics $T^{\mathcal{I}}_{D^{\mathbf{k}}}$ based on the kth derivatives with $k \in \mathbf{k}$ exclusively, all multi-derivative tests of $H^{\mathcal{I}}_{0\mathbf{k}}$ against $H^{\mathcal{I}}_{1\mathbf{k}}$ are exact. The standard scenario of two homoscedastic Gaussian populations, which is often encountered in the literature, fits within this special case.

2.2.2 Comparing the variances of two functional populations

Assume that we want to perform a comparison between the two populations in terms of variances of the first d orders of differentiation:

$$H^{\mathcal{I}}_0 : \boldsymbol{\xi}^{\mathcal{I}}_1 \overset{\mathrm{d}}{=} \boldsymbol{\xi}^{\mathcal{I}}_2 \quad \text{against} \quad H^{\mathcal{I}}_{1k} : \mathrm{Var}[D^k\boldsymbol{\xi}^{\mathcal{I}}_1] \neq \mathrm{Var}[D^k\boldsymbol{\xi}^{\mathcal{I}}_2]. \tag{11}$$

When comparing two groups in terms of variances, the adjusted partial p-value functions for derivative orders k can be computed from permutation tests based on the integrated point-wise statistic $V^{\mathcal{I}}_{D^k}$:

$$V^{\mathcal{I}}_{D^k} = \int_{\mathcal{I}} \left(\log \frac{s_{D^k\xi_1}(t)}{s_{D^k\xi_2}(t)} \right)^2 \mathrm{d}t, \tag{12}$$

where, for $j = 1, 2$: $s_{D^k\xi_j}(t)$ is the sample standard deviation of the kth derivatives at time t for the jth sample. The multi-derivative tests indexed by $\mathbf{k}$ can be instead built on the following extension of statistic $V^{\mathcal{I}}_{D^k}$ to the multivariate case:

$$U^{\mathcal{I}}_{D^{\mathbf{k}}} = \int_{\mathcal{I}} \sum_{k \in \mathbf{k}} \left(\log \frac{s_{D^k\xi_1}(t)}{s_{D^k\xi_2}(t)} \right)^2 \mathrm{d}t. \tag{13}$$

Also in this case the test statistics $V^{\mathcal{I}}_{D^k}$ and $U^{\mathcal{I}}_{D^{\mathbf{k}}}$ are dimensionless quantities, hence, they does not depend on the units of measure of the functional data and of the domain. With such choices the partial tests (5) and the multi-derivative tests (6) are all exact and consistent. Hence, Theorems 1 and 3 hold, and the resulting multi-aspect IWT is provided with a weak control of the IWER and it is consistent regardless of the family of the two functional distributions and of the sample sizes. Also in thsi case further knowledge about the specific application may turn out into further distributional assumptions which can be

sufficient for obtaining a strong control of the IWER. A renowned case is the one of **shrunk/dilated populations around the common mean**. If we know that the two functional populations share the common mean μ and could differ at most for a multiplicative term δ (i.e., $(\boldsymbol{\xi}_1 - \mu) \stackrel{\mathrm{d}}{=} \delta \cdot (\boldsymbol{\xi}_2 - \mu)$ with $\delta \in \mathbb{R}^+$ and $\mu \in H^d(T)$), we trivially have $H_{0k}^{\mathcal{I}} : \delta = 0$ and $H_{1k}^{\mathcal{I}} : \delta \neq 0$. In this case we have that $H_{0\mathbf{k}}^{\mathcal{I}} \equiv H_0$ for all $\mathbf{k}$ and so that all multi-derivative tests of $H_{0\mathbf{k}}^{\mathcal{I}}$ against $H_{1\mathbf{k}}^{\mathcal{I}}$ are exact. The scenario of two Gaussian populations with the same mean fits within this case. Moreover, for large sample sizes the strong control of the IWER can be achieved also in the case of different means by simply re-centering the two samples to a common value.

Remark. In the framework that we proposed - and in most applicative cases - the target of the analysis is comparing the two populations either in terms of means, or in terms of variances. Nevertheless, means and variances can be considered themselves as different aspect of the same distribution. Hence, it is possible - if needed - to compare the two populations in terms of mean and variance jointly, in the framework of multi-aspect testing. This would require to extend the procedure by performing multiple tests on each possible combination of the considered aspects, i.e., mean and variances of each set of derivative orders. All tests can be constructed, for instance, by applying the non-parametric combination of single-aspect tests (Pesarin and Salmaso 2010). Note that, as on one hand, this extension would increase the amount of information provided by the procedure, on the other hand it would increase the computational complexity, and possibly decrease the power of the resulting overall procedure, due to the high amount of tests included in the CTP. Hence, also in this case the amount of aspects jointly included in the analysis should be determined from a compromise between the wish of adding relevant information to the study and the desire of preserving the statistical power of the procedure.

2.3 Extensions to the Functional ANOVA

The procedure detailed in Subsection 2.1 for testing differences between two populations can be extended to test different functional hypotheses. What is required to perform the multi-aspect IWT in a general framework is to define suitable test statistics to be applied to test the null hypothesis on each interval of the domain, both for partial tests, and for multi-derivative tests. We here detail how to extend multi-aspect IWT to test differences between several functional populations.

2.3.1 Functional ANOVA for comparing means

Assume that we observe $J > 2$ groups of functional data. Assume that, for all group $j = 1, \ldots, J$, data $\{\xi_{ji}\}_{i=1,\ldots,n_j}$ are *i.i.d.* observations drawn from the random function $\boldsymbol{\xi}_j$, mapping from Ω to $H^d(T)$. We aim at testing the

null hypothesis of distributional equality between all $\boldsymbol{\xi}_j$ against the alternative hypothesis of difference in distribution between at least two random functions:

$$H_0 : \boldsymbol{\xi}_1 \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2 \overset{\mathrm{d}}{=} \ldots \overset{\mathrm{d}}{=} \boldsymbol{\xi}_J \quad \text{against} \quad H_1 : \exists j, j' \text{ s.t. } \boldsymbol{\xi}_j \overset{\mathrm{d}}{\neq} \boldsymbol{\xi}_{j'}. \tag{14}$$

In the framework of multi-aspect testing, and keeping in mind that we are interested in comparing the means, test (14) can be split in the following partial tests on interval $\mathcal{I}$:

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \overset{\mathrm{d}}{=} \ldots \overset{\mathrm{d}}{=} \boldsymbol{\xi}_J^{\mathcal{I}} \quad \text{against} \quad H_{1k}^{\mathcal{I}} : \exists j, j' \text{ s.t. } \mathbb{E}[D^k \boldsymbol{\xi}_j] \neq \mathbb{E}[D^k \boldsymbol{\xi}_{j'}]. \tag{15}$$

The adjusted partial p-value functions for FANOVA test (14) can be computed from permutation tests based on the integrated F-test statistic. Let $k \in \{0, \ldots, d\}$ denote the explored derivative order. The associated test statistic $F_{D^k}^{\mathcal{I}}$ is:

$$F_{D^k}^{\mathcal{I}} = \int_{\mathcal{I}} \frac{\sum_{j=1}^{J} n_j \left(\overline{D^k \xi}_j(t) - \overline{D^k \xi}(t) \right) / (J-1)}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(D^k \xi_{ji}(t) - \overline{D^k \xi}_j(t) \right) / (n-J)} \mathrm{d}t. \tag{16}$$

The multi-derivative tests can be instead built on the following statistic, that is the opposite integrated Wilk's lambda statistic $L_{D^{\mathbf{k}}}^{\mathcal{I}}$:

$$L_{D^{\mathbf{k}}}^{\mathcal{I}} = -\int_{\mathcal{I}} \log \frac{|W_{D^{\mathbf{k}}\xi}(t)|}{|B_{D^{\mathbf{k}}\xi}(t) + W_{D^{\mathbf{k}}\xi}(t)|} \mathrm{d}t, \tag{17}$$

where $B_{D^{\mathbf{k}}\xi}(t)$ and $W_{D^{\mathbf{k}}\xi}(t)$ are respectively the between and within sample variance-covariance matrices of data $D^{\mathbf{k}}\xi(t)$:

$$\begin{aligned} B_{D^{\mathbf{k}}\xi}(t) &= \sum_{j=1}^{J} n_j \left(\overline{D^{\mathbf{k}}\xi}_j(t) - \overline{D^{\mathbf{k}}\xi}(t) \right) \left(\overline{D^{\mathbf{k}}\xi}_j(t) - \overline{D^{\mathbf{k}}\xi}_(t) \right)' \\ W_{D^{\mathbf{k}}\xi}(t) &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(D^{\mathbf{k}}\xi_{ij}(t) - \overline{D^{\mathbf{k}}\xi}_j(t) \right) \left(D^{\mathbf{k}}\xi_{ij}(t) - \overline{D^{\mathbf{k}}\xi}_j(t) \right)'. \end{aligned}$$

Note that the theoretical properties of the multi-aspect IWT in this case are the same as for the case of comparing two means. In detail, all multi-derivative tests $H_0^{\mathcal{I}}$ against $H_{1\mathbf{k}}^{\mathcal{I}}$ are exact for any sample size and consistent, so we have weak control of the IWER and consistency. Similarly to the two population case we have the strong control of the IWER in the case of "at-most-shifted" populations with the special case of homoscedastic Gaussian population included.

2.3.2 Functional ANOVA for comparing variances

Assume that we observe $J > 2$ groups of functional data. Assume that, for all group $j = 1, \ldots, J$, data $\{\xi_{ji}\}_{i=1,\ldots,n_j}$ are *i.i.d.* observations drawn from

the random function $\boldsymbol{\xi}_j$, mapping from Ω to $H^d(T)$. We aim at testing the hypotheses (14), and we are interested in comparing the variances of the gropus. Test (14) can be then split in the following partial tests on interval $\mathcal{I}$:

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \overset{\mathrm{d}}{=} \ldots \overset{\mathrm{d}}{=} \boldsymbol{\xi}_J^{\mathcal{I}} \quad \text{against} \quad H_{1k}^{\mathcal{I}} : \exists j, j' \text{ s.t. } \mathrm{Var}[D^k \boldsymbol{\xi}_j] \neq \mathrm{Var}[D^k \boldsymbol{\xi}_{j'}]. \tag{18}$$

When comparing $J > 2$ groups in terms of variances, the adjusted partial p-value functions for the derivative orders $k \in \{0, \ldots, d\}$, can be computed from permutation tests based on the integrated Bartlett's test statistic $B_{D^k}^{\mathcal{I}}$:

$$B_{D^k}^{\mathcal{I}} = \int_{\mathcal{I}} \frac{(n-J)\ln|S_{pooled_{D^k\xi}}(t)| - \sum_{j=1}^{J}(n_j-1)\ln|S_{D^k\xi_j}(t)|}{1+\frac{1}{3(J-1)}\left[\sum_{j=1}^{J}\left(\frac{1}{n_j-1}\right)-\frac{1}{n-J}\right]}\mathrm{d}t,$$

where $S_{pooled_{D^k\xi}}(t) = \frac{1}{n-J}\sum_{j=1}^{J}(n_j-1)S_{D^k\xi_j}(t)$.

The multi-derivative tests indexed by $\mathbf{k}$ can be instead built on the integrated Box's M test statistic $M_{D^{\mathbf{k}}}^{\mathcal{I}}$:

$$M_{D^{\mathbf{k}}}^{\mathcal{I}} = \int_{\mathcal{I}} \left[(n-J)\ln|S_{pooled_{D^{\mathbf{k}}\xi}}(t)| - \sum_{j=1}^{J}(n_j-1)\ln|S_{D^{\mathbf{k}}\xi_j(t)|}\right]\mathrm{d}t.$$

As before we always have the weak control of the IWER and consistency since all all multi-derivative tests $H_0^{\mathcal{I}}$ against $H_{1\mathbf{k}}^{\mathcal{I}}$ are exact for any sample size and consistent. For example, similarly to the two population case, the strong control of the IWER holds in the case of "at-most-shrunk/dilated" populations with the same mean with the special case of Gaussian populations with the same mean included. Moreover, for large sample sizes the strong control of the IWER can be achieved also in the case of different means by simply re-centering the J samples to a common value.

3 Functional data analysis of articulatory data

3.1 Data acquisition and processing

Eighty Tyrolean words containing /r/ were selected in order to elicit the phoneme in all possible syllable contexts and positions: our dataset is the result of a two repetition of the list of the selected words by a 33 y.o. female native Tyrolean speaker with no reported speech disorders. The tongue positions during the pronunciation of each word was obtained via the Ultrasound Tongue Imaging technique (UTI). UTI makes it possible to record midsagittal or coronal plane tongue movements by placing an ultrasound (US) transceiver under the speakers' chin. The transceiver produces an ultrasound beam that travels upward through the tongue body. When the beam reaches the upper surface of the tongue it is scattered, refracted or reflected back because of the mismatch between the high

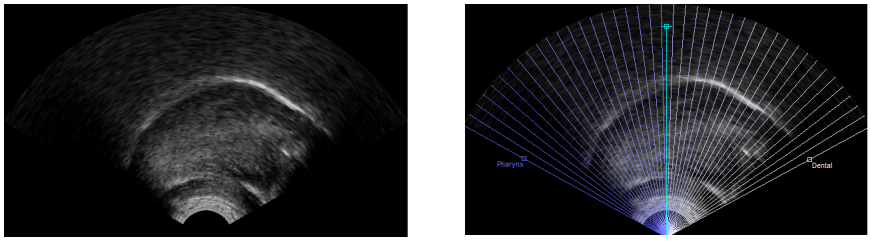

Figure 2: Left: snapshot of a dynamic ultrasound image. Right: same snapshot processed by the Articulate Assistant Advance software.

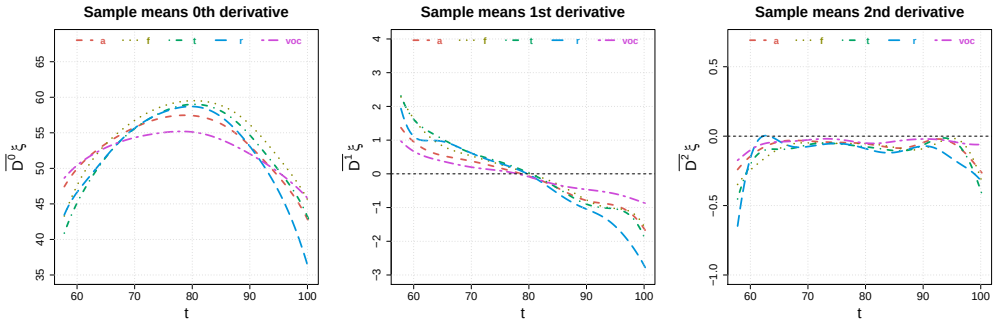

Figure 3: Sample mean curves of the five groups identified by the /r/ variants. Each image displays the tongue root on the left side and the tongue tip on the right side.

density of the muscles and the low density of the air (acoustic impedance). The reflection of the beam off the tongue surfaces is then detected by the transceiver and a white line is produced on the ultrasound image. The higher is the density mismatch, the brighter is the imaged tongue profile. Left panel of Figure 2 is an example of result of the US imaging procedure.

In order to record and analyze ultrasound tongue images the Ultrasonix Tablet Research system coupled to a C9-5/10 transceiver and the Articulate Assistant Advance (AAA) software were used. The AAA software (v. 2.16) enables to draw curves on top of ultrasound video frames and extract tongue curves from US recordings. Figure 2 on the right reports an illustration of the AAA environment. A fan shaped grid composed of 42 radial axes is constructed matching the path taken by the US beams that radiate out from the probe so to specify the area of the image that contains valid US data. An upper limit corresponding to the palate and a lower limit roughtly corresponding to the genioglossus muscle are also set. Within the defined area the AAA software searches for the brighter point on each of the 42 axes by means of looking for the gradient of the image. The result of this process is a set of 42 coordinates of the identified points for each tongue profile.

A penalized cubic B-spline smoothing procedure was used to obtain the analyzed functional data from the 42 discrete observations. As suggested in the literature (e.g., Ramsay and Silverman 2005), since our analysis aims at test-

ing differences up to the second derivative of data, we used B-splines of order six, introducing a penalty term on squared L^2 norm of the fourth derivative of the curve. Even though spatial coordinates are different for each curve, for sake of replicability we choose to use a common grid of knots for all curves. The knots position on the t-axis is found as the projection of the intersection points between the half circle that rounds up the mean curve of the n functions, and the radiant split of the AAA software (see the right panel of Figure 2). The smoothing parameter is computed through the generalized cross-validation method. Figure 3 shows the obtained smoothed vertical positions, slopes and concavities (first row), and the sample mean/variance curves of the five groups identified by the /r/ variants (second/third row). Different colors are linked to different /r/ variants emerged from the spectroacoustic analysis.

3.2 Results

First, a multi-aspect IWT-based FANOVA (14) is performed to test if there is a significant difference between at least two among the five variants, focusing on detecting differences between the means (test statistics (9), (10)). The result is summarized in Figure 4. The left panel shows the three level-2 adjusted p-value functions $\widetilde{\widetilde{p}}_{D^0}(t)$ (red), $\widetilde{\widetilde{p}}_{D^1}(t)$ (green), and $\widetilde{\widetilde{p}}_{D^2}(t)$ (blue). On the right panel we report the intervals presenting statistically significant differences between the groups. The titles of the two images report the overall p-value (i.e., the p-value of the functional test on the whole domain jointly on all derivatives), as well as the three global p-values for the test based on curves, slopes, and concavities.

First of all, for each domain point there is a significant difference between at least two variants in the vertical positions of the tongue. The slopes of the five variants are significantly different on each domain point except for the central part, i.e. in the range $[75, 90]$, due to the fact that the place of articulation is common to all variants: the speaker can not change the tongue slope in that portion without changing the place of articulation. Significant differences in concavity may be noticed in an anterior region of the tongue, if compared to the major place of articulation. This result might represent a novel acquisition of the analysis, indicating the presence a secondary articulation feature in some of the r-variants.

To understand more in detail the specific differences between the five variants, we performed a multi-aspect IWT-based analysis on each pairwise group comparison. The results are summarized in Figure 5. The main diagonal of the figure reports the smoothed tongue profiles of the five /r/ variants. Lower extra-diagonals report the multi-aspect adjusted p-value functions of the three derivative orders. Upper extra-diagonals report the intervals where statistically significant differences between each couple of groups occur for each derivative order. First of all, note that *voc* curves are statistically different from all other variants except for *a*. This sounds reasonable if you remind that the vocalized /r/ is a vowel-like sound while all the others are consonants, and that from

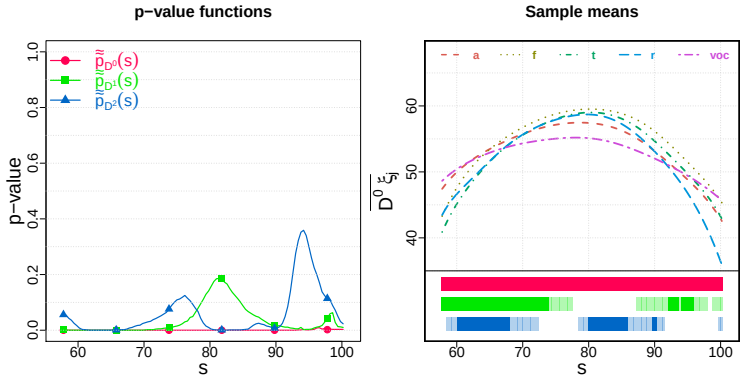

Figure 4: Results of the FANOVA test on the mean equality of groups. Left: multi-aspect adjusted p-value function for curves (red), first derivatives (green), and second derivatives (blue). Right: domain intervals selected as statistically relevant for each derivative order. Selected intervals are marked by using different color gradations corresponding to 5% and 1% levels. Each box displays the tongue root/tip on the left/right side.

the articulatory perspective the approximant variant is much more similar to a vowel than to a consonant. In addition, only *voc* distinguishes itself because of concavity. This can be explained observing that vocalized variants do not have to touch the palate. Indeed concavity of *voc* curves should be smaller smaller than the one of consonants.

Finally, let us focus on the images reporting *a* vs. *f* and *voc* vs. *f*. They are similar if we consider vertical positions. If we look at comparisons involving the first and second derivative orders, the picture changes: *a* vs. *f* box does not show relevant differences while *voc* vs. *f* box does. This supports the choice of distinguishing between vocalized and approximant variants. The approximant /r/ is like a vowel if we look at the degree of constriction, but it shows slope and concavity own of consonants. This confirms that the proposed method makes it possible to capture different articulatory properties of sounds.

4 Conclusions

We presented the multi-aspect IWT procedure, an inferential method for functional data able to take into account the information about derivatives. Multi-aspect IWT is first presented in the framework of testing differences between two functional populations. In this framework, in case of rejecting the null hypothesis of equality in distribution, multi-aspect IWT is able to select the derivative orders presenting significant differences between the two populations, and the intervals of the domain imputable for such a difference for each derivative order. This is done by providing a multi-aspect adjusted p-value function for each explored derivative order. We proved that the technique is provided with a multi-aspect control of the interval-wise error rate. If significant inter-

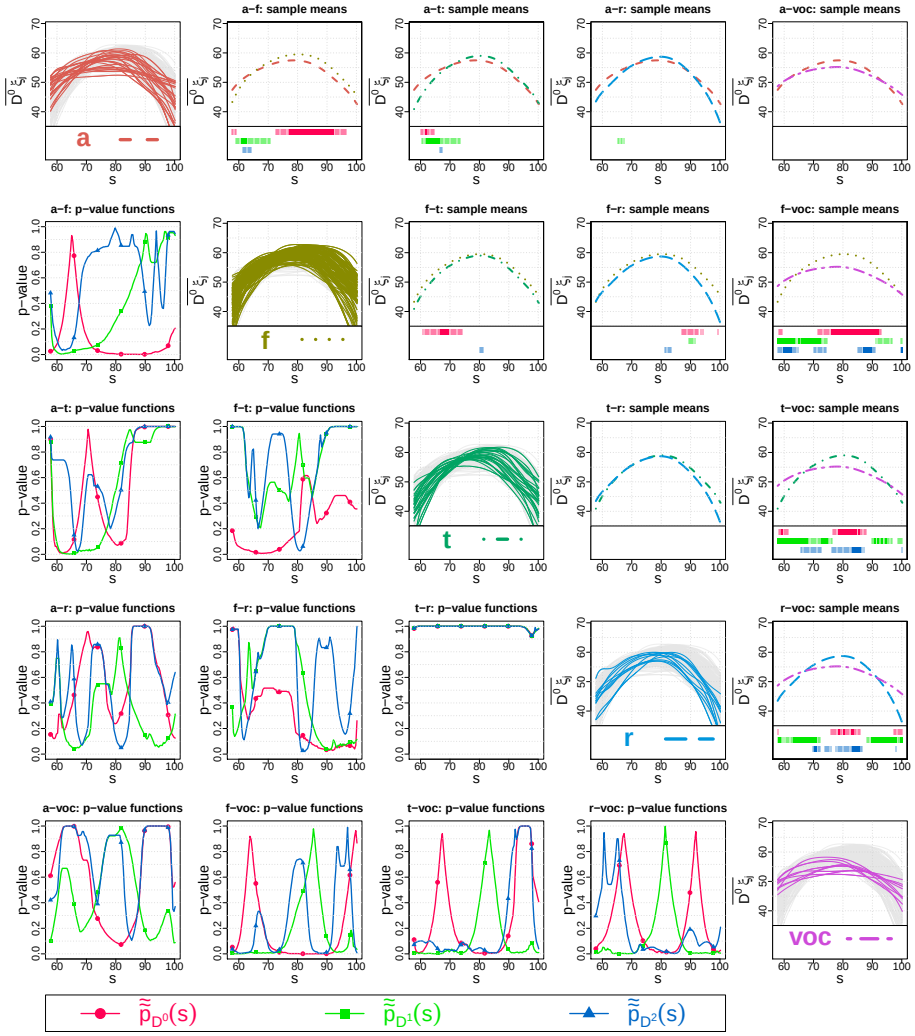

Figure 5: Results of the pairwise tests on the mean equality of groups. Main diagonal: smoothed tongue profiles. Lower extradiagonals: level-2 adjusted p-value functions. Upper extra-diagonals: mean functions of the compared groups and domain intervals selected as statistically relevant at 5% (light) and 1% levels (dark). Each box displays the tongue root/tip on the left/right side.

vals are selected by thresholding the multi-aspect adjusted p-value functions at level α: for any interval where the null hypothesis is true, the probability of detecting a difference is guaranteed to be lower than α. We discussed how to extend multi-aspect IWT to the framework of testing differences between more than two functional populations (Functional ANOVA), and to test differences between the variances of two or more populations.

The procedure can be applied considering derivatives up to any order. Nevertheless considering higher derivative orders poses two main disadvantages: (i) high derivative orders are often difficult to estimate when dealing with real data; and (ii) adding up derivative orders would decrease the power of the procedure. In general, we suggest studying the orders of derivatives that are informative for the interpretation of the results.

The implementation of the multi-aspect IWT procedure and part of the data analyzed in the current paper are provided in the `tongue.analysys` R package, included within this paper as supplementary material. The package provides functions to smooth the data, to perform the multi-aspect IWT for comparing means and/or variances, and plotting functions to create graphical outputs like the ones presented in Figure 5.

We applied the multi-aspect IWT to the analysis of tongue profiles for a study on allophonic variations of /r/ in Tyrolean, a German dialect spoken in South Tyrol. We tested differences between the profiles of the tongue of a native speaker of Tyrolean pronouncing five variants of uvular /r/. The five groups of curves correspond to five different manners of articulation: vocalized /r/, approximant, fricative, tap, and trill. We showed how the multi-aspect IWT is able in this case of giving a deep understanding of the differences between the five groups, providing practitioners with a lot of useful information.

In this work we studied the tongue profiles of a single person during a session of recording. Future works might go deeper into the matter by addressing multi-subject and multi-session cases. That would require aligning functional data (Ramsay and Silverman 2005). Note that in such a case, the multi-aspect IWT will be still an appropriate statistical tool.

5 Acknowledgements

The authors wish to acknowledge Elena Giarratano for her help in writing the computer code of the `tongue.analysis` package.

References

F. Abramovich and C. Angelini. Testing in mixed-effects FANOVA models. *J. Statist. Plann. Inference*, 136(12):4326–4348, 2006.

F. Abramovich and R. Heller. Local functional hypothesis testing. *Mathematical Methods of Statistics*, 14(3):253, 2005.

P. Boersma and D. Weenink. Praat: doing phonetics by computer, 2010.

C. Brombin and L. Salmaso. Multi-aspect permutation tests in shape analysis with small sample size. *Computational Statistics & Data Analysis*, 53(12): 3921–3931, 2009.

H. Cardot, A. Goia, and P. Sarda. Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics - Simulation and Computation*, 33(1):179–199, 2004.

L. Corain, V. B. Melas, A. Pepelyshev, and L. Salmaso. New insights on permutation approach for hypothesis testing on functional data. *Advances in Data Analysis and Classification*, 8(3):339–356, 2014.

D. D. Cox and J. S. Lee. Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika*, 95(3):621–634, 2008.

A. Cuevas, M. Febrero, and R. Fraiman. An ANOVA test for functional data. *Comput. Statist. Data Anal.*, 47(1):111–122, 2004.

M. Dalla Rosa, L. M. Sangalli, and S. Vantini. Principal differential analysis of the aneurisk65 data set. *Advances in Data Analysis and Classification*, 8(3): 287–302, 2014.

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

B. Gick, I. Wilson, and D. Derrick. *Articulatory Phonetics*. Wiley-Blackwell, 2013.

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer, 2012.

P. Ladefoged and K. Johnson. *A Course in Phonetics*. Wadsworth, sixth edition, 2011.

A. Löfqvist. Vowel-related tongue movements in speech: Straight or curved paths?(l). *The Journal of the Acoustical Society of America*, 129(3):1149–1152, 2011.

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

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

A. Pini and S. Vantini. The interval testing procedure: a general framework for inference in functional data analysis. *Biometrics*, 72:835–845, 2016.

A. Pini and S. Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, in press, 2017.

A. A. Poyton, M. S. Varziri, K. B. McAuley, P. J. McLellan, and J. O. Ramsay. Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computers & chemical engineering*, 30(4):698–708, 2006.

M. J. Qyarnström, S. M. Jaroma, and MT Laine. Changes in the peripheral speech mechanism of children from the age of 7 to 10 years. *Folia phoniatrica et logopaedica*, 46(4):193–202, 1994.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `http://www.R-project.org/`.

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

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, New York, 2005.

L. Salmaso and A. Solari. Multiple aspect testing for case-control designs. *Metrika*, 62(2):331–340, 2005.

J. A. Seikel, D. G. Drumright, and D. W. King. *Anatomy & physiology for speech, language, and hearing*. Singular, 2000.

D. J. Spitzner, J. S. Marron, and G. K. Essick. Mixed-model functional ANOVA for studying human tactile perception. *J. Amer. Statist. Assoc.*, 98(462):263–272, 2003.

A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428, 2011.

A. Staicu, Y. Li, C. M. Crainiceanu, and D. Ruppert. Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scand. J. Stat.*, 2014.

A. Vietti, L. Spreafico, and V. Galatà. An ultrasound study of the phonetic allophony of Tyrolean /r/. *ICPhS 2015 Proceedings*, 2015.

O. Vsevolozhskaya, M. Greenwood, and D. Holodov. Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *Ann. Appl. Stat.*, 8(2):905–925, 2014.

A Simulation study

In this Section we report the results of a simulation study aiming at assessing the properties of the multi-aspect IWT. The simulation study is devoted to explore an example in which we aim at testing differences between groups in terms of variance. In such a case, on one hand Theorems 1 and 3 guarantee respectively the weak control of the IWER and the marginal interval-wise consistency. Theorem 2 - on the other hand - can only be applied in the case of shrunk/dilated populations around the common mean.

Functional data are simulated on the interval $[0, 1]$ from a simple model consisting of one constant term plus one harmonic. We simulate two independent samples of functional data $\xi_{ji} : [0, 1] \to \mathbb{R}$ of sizes $n_1 = n_2 = 10$ according to the following models:

$$\begin{aligned} \xi_{1i}(t) &= A_{1i} + B_{1i}\sin(2\pi t) + C_{1i}\cos(2\pi t) && i = 1, \dots, n_1; \\ \xi_{2i}(t) &= A_{2i} + B_{2i}\sin(2\pi f_{Infl} t) + C_{2i}\cos(2\pi f_{Infl} t) && i = 1, \dots, n_2, \end{aligned}$$

where the term f_{Infl} is a fixed inflation factor for the frequency of the second population, while terms A_{ji}, B_{ji}, and C_{ji} (with $j = 1, 2$) are independent random variables sampled independently from the following uniform distributions:

$$A_{1i}, B_{1i}, C_{1i} \overset{\text{iid}}{\sim} U\left[-\frac{\sqrt{12}}{2}, +\frac{\sqrt{12}}{2}\right]$$

$$A_{2i} \overset{\text{iid}}{\sim} U\left[-\frac{\sqrt{12}}{2}\sigma_{Infl}, +\frac{\sqrt{12}}{2}\sigma_{Infl}\right]; \quad B_{2i}, C_{2i} \overset{\text{iid}}{\sim} U\left[-\frac{\sqrt{12}}{2}, +\frac{\sqrt{12}}{2}\right].$$

Consequently, the mean and variance of the coefficients are:

$$\begin{gathered} \mathbb{E}[A_{1i}] = \mathbb{E}[B_{1i}] = \mathbb{E}[C_{1i}] = 0 \\ \mathbb{E}[A_{2i}] = \mathbb{E}[B_{2i}] = \mathbb{E}[C_{2i}] = 0 \\ \mathrm{Var}[A_{1i}] = \mathrm{Var}[B_{1i}] = \mathrm{Var}[C_{1i}] = 1 \\ \mathrm{Var}[A_{2i}] = \sigma_{Infl}; \quad \mathrm{Var}[B_{2i}] = \mathrm{Var}[C_{2i}] = 1. \end{gathered}$$

The term σ_{Infl} is a fixed inflation factor for the variance of the second population. Hence, the two populations are identically distributed if and only if $\sigma_{Infl} = f_{Infl} = 1$. Note that if $f_{Infl} = 1$, the two populations are shrunk/dilated around the common mean and conditions of Theorem 2 are met.

We test differences between these two functional samples by means of the multi-aspect IWT by embedding the data in $H^2([0,1])$ and jointly testing the data and their first derivatives (i.e., $d = 1$). Note that we have $\mathbb{E}[\xi_{1i}(t)] = \mathbb{E}[D\xi_{1i}(t)] = \mathbb{E}[\xi_{2i}(t)] = \mathbb{E}[D\xi_{2i}(t)] = 0$, i.e., the means of the functions and on the first derivatives of the two populations coincide. The variances of the data of the two samples and of the corresponding first derivatives are instead:

$$\begin{aligned}
\mathrm{Var}[\xi_{1i}(t)] &= 1 + \sin^2(2\pi t) + \cos^2(2\pi t) = 2;\\
\mathrm{Var}[\xi_{2i}(t)] &= \sigma_{Infl}^2 + \sin^2(2\pi t) + \cos^2(2\pi t) = \sigma_{Infl}^2 + 1;\\
\mathrm{Var}[D\xi_{1i}(t)] &= (2\pi)^2 \sin^2(2\pi t) + (2\pi)^2 \cos^2(2\pi t) = (2\pi)^2;\\
\mathrm{Var}[D\xi_{2i}(t)] &= (2\pi f_{Infl})^2 \sin^2(2\pi t) + (2\pi f_{Infl})^2 \cos^2(2\pi t) = (2\pi f_{Infl})^2.
\end{aligned}$$

So, the two populations differ for their variances in a constant way through the whole domain. The terms σ_{Infl} and f_{Infl} influence separately the variance of the functional data and the one of the first derivatives. A natural choice in this case is then to apply the multi-aspect IWT for the comparison of variances 11 based on statistics (12) and (13).

We first describe the results of the multi-aspect IWT when applied to one instance of the simulated data. Then, we present the results - in terms of point-wise probability of rejecting the null hypothesis and IWER - obtained by simulating the data-sets from the described model 5000 times.

A.1 Analysis of one instance of the simulated data.

Figure 6 on the top displays an instance of the simulated data for the values $(\sigma_{Infl}, f_{Infl}) \in \{1,7\} \times \{1,5\}$. Figure 6 on the bottom displays the adjusted multiple-aspect p-value functions for testing the functional data displayed on the top panels of the same Figure. The four panels are associated to the four simulated data sets shown in the top panels of the same Figure. The adjusted p-value functions corresponding to functional data and first derivatives are displayed with red and green lines, respectively.

The results can be summarized as follows:

$\sigma_{Infl} = 1$, $f_{Infl} = 1$. We have $\forall t \in [0,1]$: $\tilde{\tilde{p}}_{D^0}(t) > 0.1$, $\tilde{\tilde{p}}_{D^1}(t) > 0.1$. Hence, we have no evidence for rejecting the null hypothesis of equality in distribution.

$\sigma_{Infl} = 7$, $f_{Infl} = 1$. We have $\forall t \in [0,1]$: $\tilde{\tilde{p}}_{D^0}(t) < 0.05$, $\tilde{\tilde{p}}_{D^1}(t) > 0.1$. Hence, we have a strong evidence for stating that the two populations differ in terms of vertical positions, but not in terms of slopes.

$\sigma_{Infl} = 1$, $f_{Infl} = 5$. We have $\forall t \in [0,1]$: $\tilde{\tilde{p}}_{D^0}(t) > 0.8$, $\tilde{\tilde{p}}_{D^1}(t) < 0.05$. Hence, we have a strong evidence for stating that the two populations differ in terms of slopes, but not in terms of vertical positions.

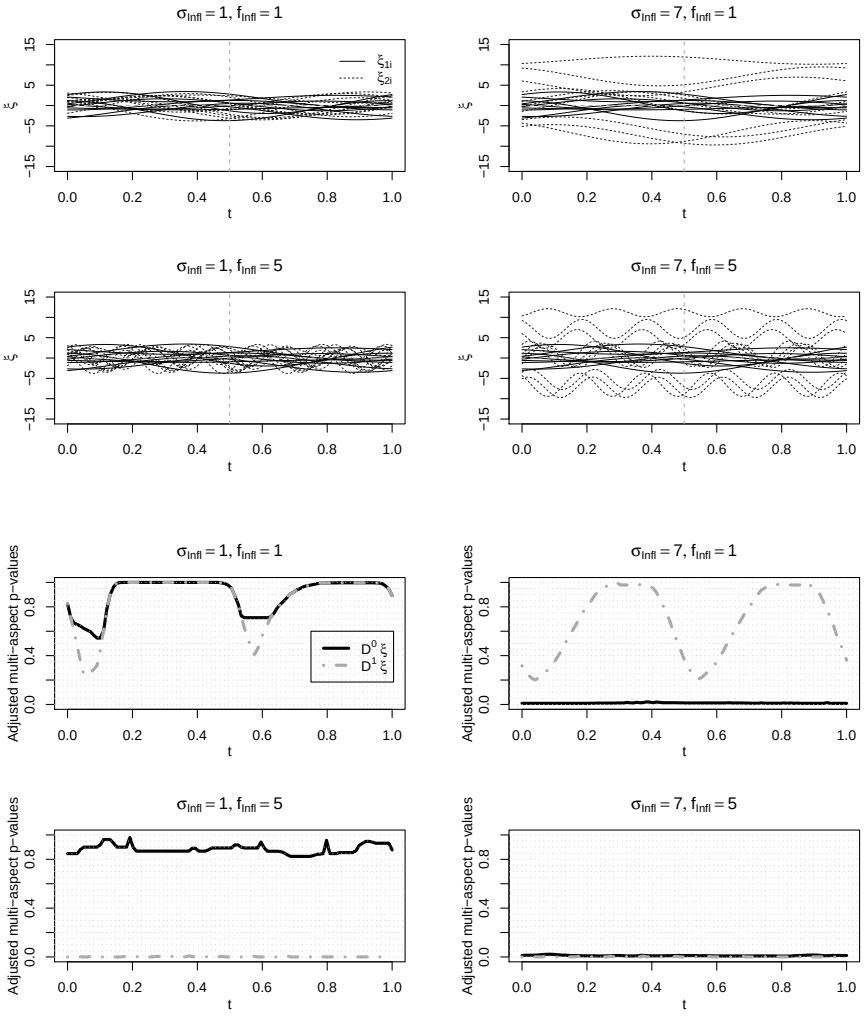

Figure 6: Top panels: instance of the simulated data of the two samples for the values $(\sigma_{Infl}, f_{Infl}) \in \{1, 7\} \times \{1, 5\}$; Bottom panels: adjusted multiple-aspect p-value functions $\widetilde{\widetilde{p}}_{D^0}(t)$ and $\widetilde{\widetilde{p}}_{D^1}(t)$ for the same data.

		σ_{Infl}				
		1.0	2.5	4.0	5.5	7.0
f_{Infl}	1	0.1	9.3	57.9	85.7	95.4
	2	0.0	10.5	60.3	86.8	95.5
	3	0.1	10.9	62.1	87.0	95.6
	4	0.0	11.6	61.8	87.5	95.8
	5	0.1	11.6	62.9	87.6	95.9

Table 1: Probability of rejection (in percentage) of H_0 against H_{10} at level $\alpha = 5\%$.

		σ_{Infl}				
		1.0	2.5	4.0	5.5	7.0
f_{Infl}	1	0.1	0.1	0.1	0.1	0.1
	2	30.4	32.6	34.3	34.5	34.6
	3	82.2	83.6	84.8	85.0	85.0
	4	95.9	96.5	96.8	96.8	96.8
	5	99.0	99.0	99.1	99.1	99.1

Table 2: Probability of rejection (in percentage) of H_0 against H_{11} at level $\alpha = 5\%$.

		σ_{Infl}				
		1.0	2.5	4.0	5.5	7.0
f_{Infl}	1	2.4	1.5	1.6	1.6	1.6
	2	0.9				
	3	0.8				
	4	1.0				
	5	0.9				

Table 3: IWER at level $\alpha = 5\%$.

$\sigma_{Infl} = 7$, $f_{Infl} = 5$**.** We have $\forall t \in [0,1]$: $\tilde{\tilde{p}}_{D^0}(t) < 0.05$, $\tilde{\tilde{p}}_{D^1}(t) < 0.05$. Hence, we have a strong evidence for stating that the two populations differ both in terms of vertical positions and in terms of slopes.

A.2 Point-wise probability of rejection and IWER.

We now discuss the point-wise probability of rejecting the null hypothesis H_0 in favor of H_{10} (i.e., difference in the vertical positions) and H_{11} (i.e., difference in terms of slopes), and the IWER. Probabilities are estimated by generating 5000 data-sets from the described models and computing the probabilities of rejecting the null hypothesis at the nominal level $\alpha = 0.05$. First note that - since the generative model is stationary - the point-wise probability of rejection is constant through all the domain. Hence, without loss of generality we report in Tables 1 and 2 the probability of rejecting the null hypothesis on the point $t = 0.5$, respectively in favor of H_{10} and H_{11}.

Table 3 reports instead the IWER, that is in this case, the probability of rejecting H_0 on at least one point of the domain in at least one derivative order in the scenario $\sigma_{Infl} = f_{Infl} = 1$, the probability of rejecting H_0 on at least one point of the domain in the test on vertical positions in the scenarios $\sigma_{Infl} = 1$, $f_{Infl} \neq 1$, and the probability of rejecting H_0 on at least one point of the domain in the test on slopes in the scenarios $\sigma_{Infl} \neq 1$, $f_{Infl} = 1$. The weak control of the IWER is achieved if $IWER \leq \alpha$ in the scenario $\sigma_{Infl} = f_{Infl} = 1$, while the strong control is achieved if $IWER \leq \alpha$ also in all other scenarios where at least one null hypothesis is true.

Note that as expected when the null hypothesis is true ($\sigma_{Infl} = f_{Infl} = 1$)

the point-wise probability of rejection in both partial tests and the IWER are controlled. The probability of correctly detecting a difference in the variance of vertical positions increases in the term σ_{Infl} for all values of f_{Infl}. Conversely, the probability of correctly detecting a difference in the variance of slopes increases in the term f_{Infl} for all values of σ_{Infl}. Interestingly, the IWER is also controlled here in a strong sense in all explored cases (i.e., also when $f_{Infl} \neq 1$). In this explored case the strong control of the IWER seems to be quite robust with respect to violations of the assumptions of Theorem 2.

B Proofs

Theorem 1. Let $\mathcal{I}$ denote an interval where $H_0^{\mathcal{I}}$ is true, i.e., $\boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}}$. Consider the test of hypotheses:

$$H_0^{\mathcal{I}} : \boldsymbol{\xi}_1^{\mathcal{I}} \overset{\mathrm{d}}{=} \boldsymbol{\xi}_2^{\mathcal{I}} \text{ against } H_{1\{0,\dots,d\}}^{\mathcal{I}} : \exists k \in \{0,\dots,d\} \text{ s.t. } \mathrm{A}[D^k \boldsymbol{\xi}_1^{\mathcal{I}}] \neq A[D^k \boldsymbol{\xi}_2^{\mathcal{I}}],$$

that is the multi-derivative test (6) including all explored derivative orders. This test is also exact. This implies that we are on the conditions of Theorem 3 of Pini and Vantini (2017). Specifically, $\forall \alpha \in (0,1)$:

$$\mathbb{P}\left(\exists t \in \mathcal{I}, \exists k \in \{0,\dots,d\} \text{ s.t. } \tilde{p}_{D^{\{0\dots,d\}}}(t) \leq \alpha\right) \leq \alpha.$$

Finally, note that, $\forall k \in \{0,\dots,d\}$, the k-th order multi-aspect adjusted p-value function $\tilde{\tilde{p}}_{D^k}(t)$ is the supremum of all adjusted p-value functions of tests including k. So, we have, $\forall k \in \{0,\dots,d\}$, $\forall t \in \mathcal{I}$:

$$\tilde{\tilde{p}}_{D^k}(t) \geq \tilde{p}_{D^{\{0\dots,d\}}}(t).$$

Hence, we can conclude that $\forall \alpha \in (0,1)$:

$$\mathbb{P}\left(\exists t \in \mathcal{I}, \exists k \in \{0,\dots,d\} \text{ s.t. } \tilde{\tilde{p}}_{D^k}(t) \leq \alpha\right) \leq \alpha.$$

□

Theorem 2. Let $\mathcal{I} \subseteq T$ be an interval of the domain, and let us define

$$\mathbf{k}_{null}^{\mathcal{I}} = \left\{k \in \{0,\dots,d\} s.t. H_{0k}^{\mathcal{I}} \text{ is true}\right\}.$$

Consider the test of hypotheses

$$H_{0\mathbf{k}_{null}^{\mathcal{I}}}^{\mathcal{I}} : \forall k \in \mathbf{k}_{null}^{\mathcal{I}} A[D^k \boldsymbol{\xi}_1^{\mathcal{I}}] = A[D^k \boldsymbol{\xi}_2^{\mathcal{I}}]$$

against

$$H_{1\mathbf{k}_{null}^{\mathcal{I}}}^{\mathcal{I}} : \exists k \in \mathbf{k}_{null}^{\mathcal{I}} \text{ s.t. } \mathrm{A}[D^k \boldsymbol{\xi}_1^{\mathcal{I}}] \neq A[D^k \boldsymbol{\xi}_2^{\mathcal{I}}].$$

This test is exact, so we are on the conditions of Theorem 3 of Pini and Vantini (2017). Specifically, $\forall \alpha \in (0,1)$:

$$\mathbb{P}\left(\exists t \in \mathcal{I}, \exists k \in \mathbf{k}^{\mathcal{I}}_{null} \text{ s.t. } \tilde{p}_{D^{\mathbf{k}^{\mathcal{I}}_{null}}}(t) \leq \alpha\right) \leq \alpha.$$

Finally, note that, $\forall k \in \mathbf{k}^{\mathcal{I}}_{null}$, the k-th order multi-aspect adjusted p-value function $\tilde{\tilde{p}}_{D^k}(t)$ is the supremum of all adjusted p-value functions of tests including k. So, we have, $\forall k \in \mathbf{k}^{\mathcal{I}}_{null}$, $\forall t \in \mathcal{I}$:

$$\tilde{\tilde{p}}_{D^k}(t) \geq \tilde{p}_{D^{\mathbf{k}^{\mathcal{I}}_{null}}}(t).$$

Hence, we can conclude that $\forall \alpha \in (0,1)$:

$$\mathbb{P}\left(\exists t \in \mathcal{I}, \exists k \in \mathbf{k}^{\mathcal{I}}_{null} \text{ s.t. } \tilde{\tilde{p}}_{D^k}(t) \leq \alpha\right) \leq \alpha.$$

□

Theorem 3. Let $k \in \{0, \ldots, d\}$, and let $\mathcal{I}$ denote an interval such that $\forall \mathcal{J} \subseteq \mathcal{I}$, $H^{\mathcal{J}}_{1k}$ is false. This also implies that $\forall \mathbf{k} \ni k$, $\forall \mathcal{J} \subseteq \mathcal{I}$: $H^{\mathcal{J}}_{1\mathbf{k}}$ is false. All partial tests of $H^{\mathcal{J}}_0$ against $H^{\mathcal{J}}_{1k}$ and all multi-derivarive tests of $H^{\mathcal{J}}_0$ against $H^{\mathcal{J}}_{1\mathbf{k}}$ are consistent. Hence, the conditions of Theorem 4 of Pini and Vantini (2017) hold, and all partial and multi-derivative interval-wise tests are consistent. Specifically, $\forall \mathbf{k} \ni k$, $\forall \alpha \in (0,1)$ and for $n \to \infty$:

$$\mathbb{P}\left(\forall t \in \mathcal{I},\ \tilde{p}_{D^{\mathbf{k}}}(t) \leq \alpha\right) \to 1.$$

i.e., $\forall \mathbf{k} \ni k$, $\forall t \in \mathcal{I}$, $\tilde{p}_{D^{\mathbf{k}}}(t) \xrightarrow[n\to\infty]{a.s.} 0$. At this point, note that $\tilde{\tilde{p}}_{D^k}(t) = \sup_{k \ni k} \tilde{p}_{D^k}(t)$. Hence we also have $\forall t \in \mathcal{I}$, $\tilde{\tilde{p}}_{D^{\mathbf{k}}}(t) \xrightarrow[n\to\infty]{a.s.} 0$, i.e., $\forall \alpha \in (0,1)$ and for $n \to \infty$:

$$\mathbb{P}\left(\forall t \in \mathcal{I},\ \tilde{\tilde{p}}_{D^{\mathbf{k}}}(t) \leq \alpha\right) \to 1.$$

□

C Results of variance inspection

Figures 7 and 8 report the results of the FANOVA test and pairwise tests comparing the variants in terms of variances (see Subsection 2.2.2).

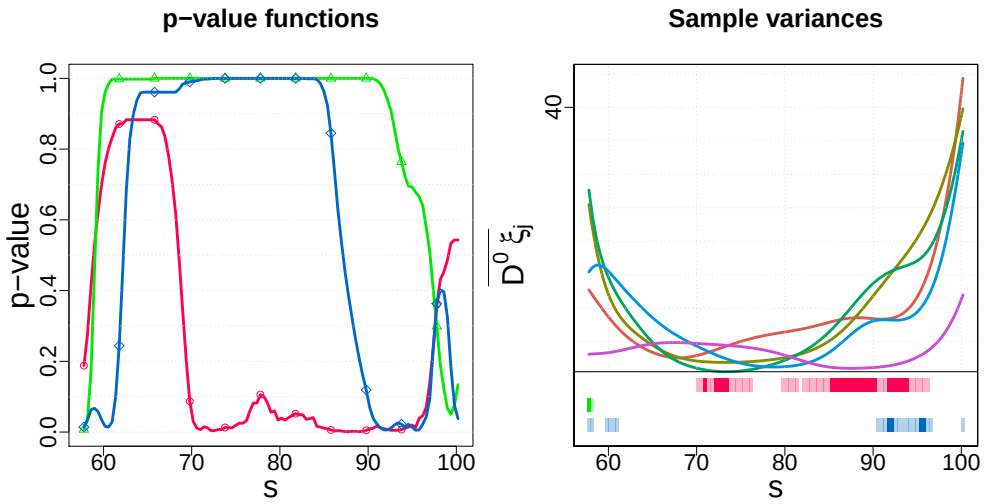

Figure 7: Results of the FANOVA test on the variance equality of groups. On the left side, the level-2 adjusted p-value function for curves (red), first derivatives (green), and second derivatives (blue). On the right side, the domain intervals selected as statistically relevant for each derivative order. Selected intervals are marked by using different color gradations corresponding to 5% and 1% levels, visualized by the dashed grey lines in the image on the left side. Each box displays the tongue root/tip on the right/left side.

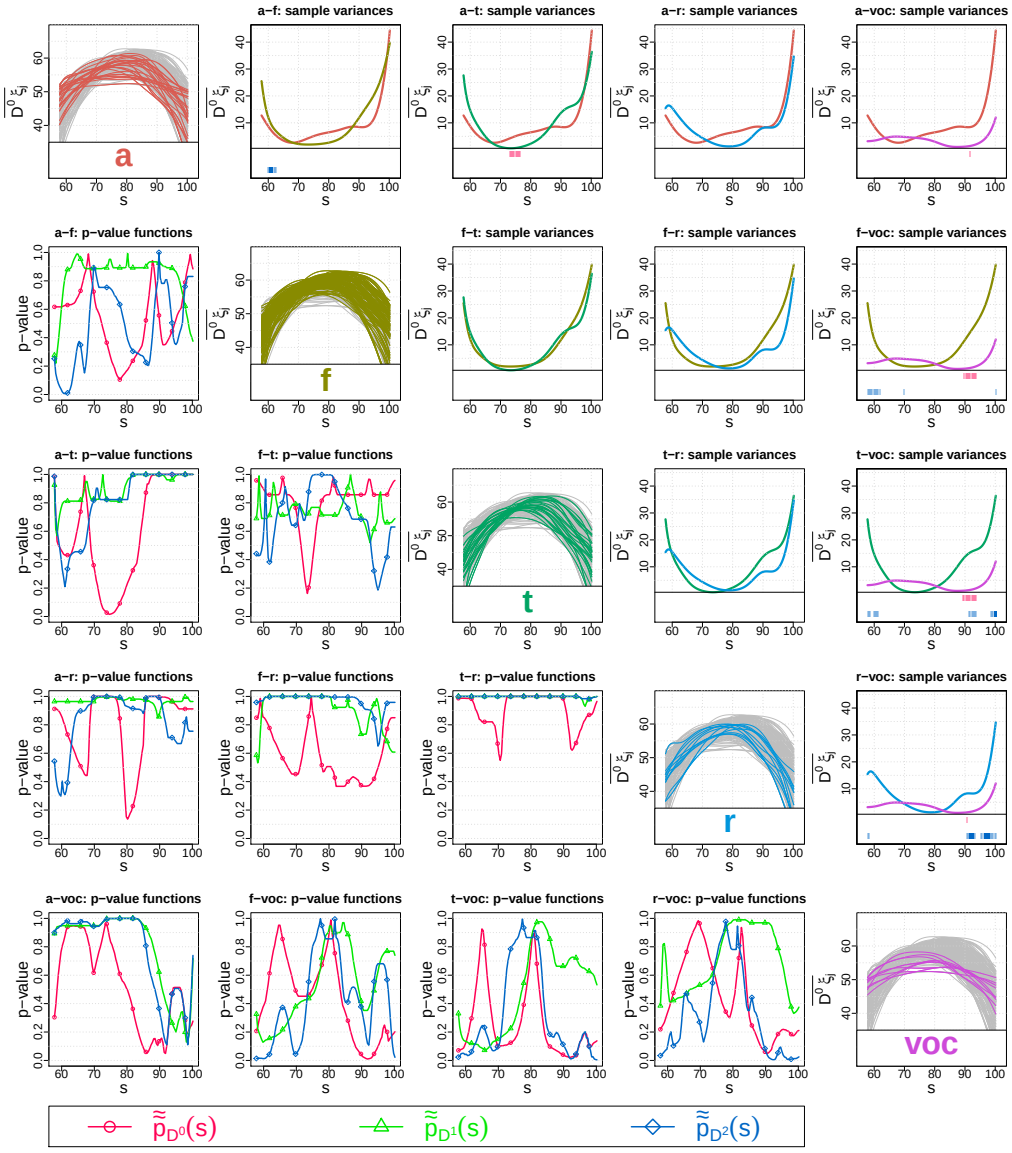

Figure 8: Results of the pairwise tests on the variance equality of groups. Main diagonal: smoothed tongue profiles. Lower extradiagonals: level-2 adjusted p-value function for curves (red), first derivatives (green), and second derivatives (blue). Upper extra-diagonals: domain intervals selected as statistically relevant for each derivative order. Selected intervals are marked by using different color gradations corresponding to 5% and 1% levels, visualized by the dashed grey lines in the image on the left side. Each box displays the tongue root/tip on the right/left side.