Check for updates

Received: 23 December 2020 | Accepted: 23 March 2022

DOI: 10.1111/biom.13669

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Domain selection and familywise error rate for functional data: A unified framework

Konrad Abramowicz[1] | **Alessia Pini**[2] | **Lina Schelin**[3] | **Sara Sjöstedt de Luna**[1] | **Aymeric Stamm**[4] | **Simone Vantini**[5]

[1]Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden

[2]Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

[3]Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

[4]Department of Mathematics Jean Leray, UMR CNRS 6629, Nantes University, Nantes, France

[5]MOX – Modelling and Scientific Computing Laboratory, Department of Mathematics, Politecnico di Milano, Milan, Italy

Correspondence
Konrad Abramowicz, Department of Mathematics and Mathematical Statistics, UmeåUniversity, 901 87 Umeå, Sweden.
Email: konrad.abramowicz@umu.se

Funding information
Vetenskapsrådet, Grant/Award Numbers: 2016-02763, 340-2013-5203

Abstract

Functional data are smooth, often continuous, random curves, which can be seen as an extreme case of multivariate data with infinite dimensionality. Just as componentwise inference for multivariate data naturally performs feature selection, subsetwise inference for functional data performs domain selection. In this paper, we present a unified testing framework for domain selection on populations of functional data. In detail, p-values of hypothesis tests performed on pointwise evaluations of functional data are suitably adjusted for providing control of the familywise error rate (FWER) over a family of subsets of the domain. We show that several state-of-the-art domain selection methods fit within this framework and differ from each other by the choice of the family over which the control of the FWER is provided. In the existing literature, these families are always defined a priori. In this work, we also propose a novel approach, coined thresholdwise testing, in which the family of subsets is instead built in a data-driven fashion. The method seamlessly generalizes to multidimensional domains in contrast to methods based on a priori defined families. We provide theoretical results with respect to consistency and control of the FWER for the methods within the unified framework. We illustrate the performance of the methods within the unified framework on simulated and real data examples and compare their performance with other existing methods.

KEYWORDS

adjusted p-value function, functional data, local inference, permutation test

1 | INTRODUCTION

Functional data analysis (FDA) is a field of statistics that pertains to the study of datasets in which the sample unit is a smooth curve. Such data arise as the results of many experimental studies, including engineering, biology, medicine, and biomechanics. Examples of the two latter ones (diffusion magnetic resonance imaging data and kinematic data) are going to be addressed in this paper.

For functional data, besides estimation, clustering, and prediction, it is of critical importance to design appropriate statistical methodologies for inference such as testing hypotheses on populations of functional data, which is the objective of the present work. For example, suppose that random functions are observed for two populations, and we want to test if the mean functions μ_1 and μ_2 are the same, testing $H_0 : \mu_1(\cdot) = \mu_2(\cdot)$ versus $H_1 : \mu_1(\cdot) \neq \mu_2(\cdot)$. In our examples, we consider the knee kinematics of two

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Biometrics* published by Wiley Periodicals LLC on behalf of International Biometric Society.

groups of patients and look for the differences in their performance, as well as compare two diffusion models for structural connectivity in the brain. Multiple methods have been devised in the literature to form *global* tests for this setting and more general scenarios, both parametrically (e.g., Horváth and Kokoszka, 2012; Staicu *et al.*, 2014) and nonparametrically (e.g., Cardot *et al.*, 2004; Corain *et al.*, 2014). With the help of such methods, we could statistically identify the existence of significant differences between the populations but their results do not tell us in which part of the domain (time of the movement, or part of the brain) the differences appear. Therefore, if H_0 of a global test is rejected, we want to identify the parts of the domain where significant differences occur, performing the so-called *local* inference. In this paper, we focus on local inference for functional data, which we refer to as *domain selection*. Few attempts have been made in this direction in the literature. A first and natural approach pertains to discretizing the functional domain and performing pointwise inference. For instance, Fan and Zhang (2000) and Reiss *et al.* (2010) derive pointwise confidence bands for functional data. This however only provides a pointwise control of errors arising in statistical hypothesis testing. Similar to the multivariate case, devising testing approaches that use multiple (or even infinite) numbers of hypotheses affects the performance of the test by increasing the overall probability of making wrong rejections. Out of the multiple concepts that can be used for controlling this overall probability, the most well known are the familywise error rate (FWER), which is the probability of rejecting at least one true null hypothesis, and false discovery rate (FDR), which quantifies the expected proportion of false discoveries (i.e., the expected ratio between the number of wrong rejections and the total number of rejections). Both measures are extensions to the multivariate setting of the type I error, though with different experimental meanings. While the control of the FWER is related to a deterministic (although unknown) partition of the domain related to where the null hypothesis is true, the control of the FDR is instead related to a random (but observed) partition of the domain related to the rejections of the null hypothesis. Controlling the FWER is stronger than controlling the FDR: if we devise a method for which the FWER is controlled at a specific level for any ground truth of the null hypothesis, then the FDR will also be controlled at that level. Additionally, if the null hypothesis is true on the whole domain, the two measures coincide. Simulation studies in biomechanics and brain imaging research fields comparing different approaches for FWER as well as comparing it to FDR elucidate differences and similarities in detection regions depending on the approach used. For kinematic data, we refer to, for example, Naouma and Pataky (2019) and Pataky *et al.* (2021), while discussions in the context of brain imaging can be found in, for example, Logan and Rowe (2004).

In our work, we focus on the problem of testing functional data by providing adjusted *p*-values controlling the FWER. Some examples of methods that instead of control the FDR in the context of functional data can be found in Perone Pacifico *et al.* (2004) and Olsen *et al.* (2021). We base our approach on properly adjusting pointwise *p*-values in order to account for the multiplicity of tests that are jointly performed when analyzing the whole domain. This issue has, in multivariate analysis, given birth to many adjustment procedures (see, e.g., Marcus *et al.*, 1976; Holm, 1979; Holmes *et al.*, 1996; Winkler *et al.*, 2014). However, functional data differ from multivariate data in that functional data feature unique properties such as smoothness and domain continuity, which can be used to improve upon classic methods for performing domain selection. Vsevolozhskaya *et al.* (2014) propose a method for domain selection that relies on the availability of a partition of the domain that allows to perform dimensionality reduction. They perform functional tests on the elements of the partition and resort to a closed testing procedure (Marcus *et al.*, 1976) to adjust the resulting *p*-values and achieve strong control of the FWER for the family generated by unions of the elements of the partition. The resulting inference heavily depends on the partition itself. In addition, the coarseness of the partition defines the depth to which local inference is performed, and the approach is of practical relevance only for relatively small predefined partitions. We refer to this method as *partition-closed testing* (PCT). Another approach— introduced for functional *t*-tests in Pini and Vantini (2017) and extended to functional-on-scalar linear models in Abramowicz *et al.* (2018)— is *interval-wise testing* (IWT). The procedure simultaneously tests a family of hypotheses generated by all intervals of the domain. This is however of practical use only for functional data defined on one-dimensional domains as it is unclear how to define "multidimensional intervals" and would be computationally overdemanding due to the curse of dimensionality. IWT provides control of the FWER for the family of all intervals: if the null hypothesis is true on more complex subsets of the domain (e.g., a union of intervals), IWT fails to control the FWER. The adjustment made on the pointwise and setwise *p*-values is only one of the possible approaches presented in the literature. Some works recently focused on providing simultaneous confidence bands for functional data: Degras (2017) develop asymptotic confidence bands, Rathnayake and Choudhary (2016) focus on parametric confidence bands, and Crainiceanu *et al.* (2012) and Park *et al.* (2017) use bootstrap confidence bands. Confidence sets based on random field theory have also been considered in, for example, Telschow and Schwartzman (2022) and Liebl and Reimherr (2020). It

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is also worth to notice that besides FWER and FDR, additional performance measures have been introduced (e.g., false discovery exceedance, false cluster rates, and false nondiscovery proportions). We do not discuss them further in this paper, but refer the reader to, for example, Perone Pacifico *et al.* (2004) for further information.

In our paper, we focus on methods that aim at providing control of the FWER. We start by formalizing the concepts in Section 2 and introduce a general framework for performing local inference in Section 3. The basic principles of the methods are based on standard pointwise inferential procedures and their setwise counterparts for a chosen family of subsets. The framework is related to a wide class of inferential problems (e.g., comparisons of population means, hypothesis tests for coefficients in models), as we utilize general concepts of null and alternative hypotheses. Furthermore, it can be applied either to a parametric or a nonparametric analysis. Using the properties of the underlying tests, we formulate and prove finite sample and asymptotic properties for methods within this framework in Sections 4 and 5 and Web Appendices A and B, respectively. In Section 4, we show how well-established methods from the literature on inference for functional data can be described in the light of our proposed unified framework. In Section 5, we present a novel method with asymptotic control of the FWER. The control is provided for the family generated by domain discretization corresponding to the resolution of the observed functional data. The computational burden of the new method is independent of the dimension and complexity of the data domain. Further, simulation studies designed to exemplify the properties of the described methods and to compare them with alternative methods existing in the literature are presented in Section 6. Real data applications are presented in Section 7, while Section 8 contains conclusions. Additional definitions and results are presented in Web Appendices C–I.

2 | DEFINITIONS AND THE INFERENTIAL PROBLEM

Consider a space of continuous random functions defined on domain D, where D is a compact subset of $\mathbb{R}^d$, $d \geq 1$. Let us consider a general inferential problem based on a sample of n independent functional observations. Without loss of generality, assume that we aim at testing a functional null hypothesis H_0 against an alternative hypothesis H_1. For instance, it could be the functional two-sample t-test where $H_0 : \mu_1(\cdot) = \mu_2(\cdot)$ is tested against $H_1 : \mu_1(\cdot) \neq \mu_2(\cdot)$. Let $\mathcal{D}_0$ and $\mathcal{D}_1$ denote the regions of D where the null hypothesis is true and false, respectively. Our goal is to construct an inferential method that correctly identifies $\mathcal{D}_0$ and $\mathcal{D}_1$ and controls the type I error along with the domain. Formally, assume to observe a random sample of n continuous functions $y_i(t), t \in D, i = 1, \ldots, n$ possibly with attached functional or scalar covariates. For all $t \in D$, we denote by H_0^t and H_1^t the restrictions of H_0 and H_1 to point t. Assume that we can obtain a test statistic $T_n(t)$ for testing H_0^t against H_1^t at point t, where H_0^t is rejected for large values of $T_n(t)$. Let $p_n(t)$ denote the p-value of the test at point t based on $T_n(t)$ and data $\{y_i(t)\}_{i=1}^n$. Depending on the assumptions of the generative process of functional data and on the sample size, $p_n(t)$ can be computed with parametric, asymptotic, or nonparametric tests.

2.1 | Pointwise and setwise test properties

Below, we define some of the properties that are typically required for pointwise tests.

Definition 2.1. *We say that the pointwise test of H_0^t against H_1^t based on the statistic $T_n(t)$ with p-value $p_n(t)$ is*

- ***valid*** *if for all $\alpha \in (0,1)$ and any $n \in \mathbb{N}_+$ the probability of rejecting H_0^t at level α when it is true is smaller or equal to α, that is, $t \in \mathcal{D}_0 \Rightarrow \mathbb{P}[p_n(t) \leq \alpha] \leq \alpha$,*
- ***asymptotically valid*** *if for all $\alpha \in (0,1)$ the probability of rejecting H_0^t at level α when it is true is asymptotically smaller or equal to α, that is, $t \in \mathcal{D}_0 \Rightarrow \lim_{n\to\infty} \mathbb{P}[p_n(t) \leq \alpha] \leq \alpha$, and*
- ***consistent*** *if for alls $\alpha \in (0,1)$ the probability of rejecting H_0^t at level α when H_0^t is false is asymptotically one, that is, $t \in \mathcal{D}_1 \Rightarrow \lim_{n\to\infty} \mathbb{P}[p_n(t) \leq \alpha] = 1$.*

Remark 2.1. In Definition 2.1, we specify in general terms $n \to \infty$. However, depending on the test that is performed, some more specific assumptions about the sample size may be required. For example, when performing a test comparing two populations, both sample sizes are required to go to infinity and not only the total sample size n.

Note that according to Definition 2.1, we allow for valid tests— with an error smaller than α— rather than exact tests— with an error equal to α— which is related to the use of permutation tests in our paper. We now introduce the following hypotheses defined on any set $A \subset D$: H_0^A is the hypothesis that H_0^t is true for all $t \in A$ while H_1^A is the alternative that H_1^t is true for some $t \in A$. Assume that tests of H_0^A against H_1^A are performed using the following statistic:

$$T_n^A = \frac{1}{|A|} \int_A T_n(t) \mathrm{d}t, \tag{1}$$

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where the integral is defined in a Lebesgue sense and $|A|$ denotes the Lebesgue measure of A. Let p_n^A be the corresponding p-value. We now provide the definitions of validity and consistency for the test on A.

Definition 2.2. *For any $A \subseteq D$ such that $|A| > 0$, we say that the test of H_0^A against H_1^A, based on the statistic T_n^A in (1) with a p-value p_n^A is*

- ***valid** if for any $\alpha \in (0,1)$ and for any $n \in \mathbb{N}_+$, $|A \cap \mathcal{D}_1| = 0 \Rightarrow \mathbb{P}[p_n^A \leq \alpha] \leq \alpha$;*
- ***asymptotically valid** if for any $\alpha \in (0,1)$, $|A \cap \mathcal{D}_1| = 0 \Rightarrow \lim_{n\to\infty} \mathbb{P}[p_n^A \leq \alpha] \leq \alpha$;*
- ***consistent** if for any $\alpha \in (0,1)$: $|A \cap \mathcal{D}_1| > 0 \Rightarrow \lim_{n\to\infty} \mathbb{P}[p_n^A \leq \alpha] = 1$.*

In the nonparametric permutation test framework, it is straightforward to build valid and consistent tests on sets from the corresponding pointwise tests under mild assumptions. Specifically, following Pesarin and Salmaso (2010, pp. 122–124), we know that if permutation tests are used and we use the same permutations for all points of the set, the (asymptotic) validity of the pointwise tests implies (asymptotic) validity of the tests on sets. Further, if for all $t \in D$, $T_n(t)$ is nonnegative and stochastically greater under H_1^t than under H_0^t, we have that consistency of the pointwise tests implies consistency of the tests on sets.

2.2 | Domain selection

Suppose that we use the pointwise p-value $p_n(t)$ for selecting the parts of the domain imputable for the rejection of H_0 by performing thresholding at level $\alpha \in (0,1)$, that is, the parts where $p_n(\cdot) < \alpha$. The probability that this selected region—or part of it—is wrongly selected is not controlled, since $p_n(t)$ is computed pointwise and cannot guarantee any control of the probability of committing at least one type I error over the whole domain. In multivariate statistical analysis, p-values are adjusted to provide global control of the type I error rate. Selection of the variables responsible for the rejection of the null hypothesis is performed by thresholding properly adjusted p-values instead of the original unadjusted ones. A type of adjustment strategy is controlling the FWER, that is, the probability of rejecting at least one true null hypothesis. There are two classical types of control of the FWER that have been defined in the literature: weak control of the FWER holds if the FWER is controlled when all null hypotheses are true, while strong control of the FWER holds if the FWER is controlled for any configuration of true and false null hypotheses. We introduce an analogous concept in FDA. We define strong control of the FWER of a test procedure based on an adjusted p-value function $\widetilde{p}_n(t)$, $t \in D$ (cf. Equation (6)) as follows.

Definition 2.3. *We say that a test procedure has a **strong control of the FWER** if for any $n \in \mathbb{N}_+$ its adjusted p-value function $\widetilde{p}_n(t)$, $t \in D$ is such that, for all $\alpha \in (0,1)$,*

$$A \subseteq \mathrm{cl}(\mathcal{D}_0) \Rightarrow \mathbb{P}(\exists t \in A : \widetilde{p}_n(t) \leq \alpha) \leq \alpha. \tag{2}$$

Here $\mathrm{cl}(\mathcal{D}_0)$ denotes the closure of the set $\mathcal{D}_0$. In some cases, we cannot control over all possible configurations of $\mathcal{D}_0$ and $\mathcal{D}_1$, only have specific types of them. We therefore define such a type of intermediate control. Consider a family of domain subsets $\mathcal{G}$, in which elements can be expressed as finite unions of closed compact subregions of D.

Definition 2.4. *We say that a test procedure has a **control of the FWER restricted to family** $\mathcal{G}$ if for all $n \in \mathbb{N}_+$ its adjusted p-value function $\widetilde{p}_n(t)$, $t \in D$ is such that, for all $\alpha \in (0,1)$,*

$$G \in \mathcal{G} \text{ with } G \subseteq \mathrm{cl}(\mathcal{D}_0) \Rightarrow \mathbb{P}(\exists t \in G : \widetilde{p}_n(t) \leq \alpha) \leq \alpha. \tag{3}$$

When $\mathcal{G}$ is the family of all possible subsets of D, the control defined as above coincides with the strong one. Finally, analogously to the multivariate framework, if a procedure has a control of the FWER restricted to $\mathcal{G} = \{D\}$, we say that it has a weak control of the FWER. Some situations may only have an asymptotic control of the FWER , that is, control of the FWER when $n \to \infty$. In the following, we formalize it for the restricted FWER.

Definition 2.5. *We say that a test procedure has an **asymptotic control of the FWER restricted to family** $\mathcal{G}$ if its adjusted p-value function $\widetilde{p}_n(t)$, $t \in D$ is such that, for all $\alpha \in (0,1)$,*

$$G \in \mathcal{G} \text{ with } G \subseteq \mathrm{cl}(\mathcal{D}_0) \Rightarrow \limsup_{n\to\infty} \mathbb{P}(\exists t \in G : \widetilde{p}_n(t) \leq \alpha) \leq \alpha. \tag{4}$$

Finally, we define the consistency of an inferential procedure, assuring that it asymptotically detects the parts of the domain where H_1 holds, that is, $\mathcal{D}_1$.

Definition 2.6. *We say that the test procedure is **consistent** if its adjusted p-value function $\widetilde{p}_n(t)$, $t \in D$ is such that, for*

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

all $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \mathbb{P}(\forall t \in \mathrm{Int}(\mathcal{D}_1) : \widetilde{p}_n(t) \le \alpha) = 1, \tag{5}$$

where $\mathrm{Int}(\mathcal{D}_1)$ *denotes the interior of set* $\mathcal{D}_1$.

Remark 2.2. Since tests on subsets are performed using an integrated pointwise test statistic, deviations from the null hypothesis at only one point or a set of null Lebesgue measures cannot be detected. In particular, the boundary of the set $\mathcal{D}_1$ cannot be detected, since it has a null measure. Hence, strong control of the FWER is extended beyond $\mathcal{D}_0$ to the closure of the set $\mathcal{D}_0$, while consistency can be reached only for the interior of $\mathcal{D}_1$.

3 | A UNIFIED FRAMEWORK

In this section, we describe a unified framework for testing local functional hypotheses on D, given a set of n independent random functions. We present a class of methods that can be used to adjust the pointwise p-values $p_n(t)$ to provide a control of the FWER over specific families. Consider a nonempty (possibly infinite) family $\mathcal{F}$ of Lebesgue-measurable subsets of the domain of nonnull measure, such that $\cup_{S\in\mathcal{F}} S = D$. The testing procedure that we propose is based on performing tests on the restrictions of H_0 and H_1 to all subsets of the family and adjusting the p-value according to the results of such tests. First, we formally describe the testing procedure for a general $\mathcal{F}$ and provide a characterization of the inferential properties of the methods depending on the choice of $\mathcal{F}$. Then, we describe several methods that can be obtained for some particular choices of $\mathcal{F}$. The unified framework consists of the following steps (presented graphically in Web Appendix I):

1. *Computation of p-values for all subsets.* For all $S \in \mathcal{F}$, compute the p-value p_n^S of the test of H_0^S against H_1^S, based on the test statistic T_n^S in (1).
2. *Computation of the adjusted p-value function. For all* $t \in D$, *compute the adjusted p-value,*

$$\widetilde{p}_n(t;\mathcal{F}) = \sup_{S\in\mathcal{F}: t\in S} p_n^S. \tag{6}$$

3. *Domain selection. Select the subsets of* D *where* H_0 *is rejected at level* $\alpha \in (0,1)$ *as*

$$\{t \in D : \quad \widetilde{p}_n(t;\mathcal{F}) \le \alpha\}.$$

In the following sections, we consider two types of families $\mathcal{F}$: a predefined type, where all subsets belonging to $\mathcal{F}$ are defined a priori, and a data-driven type, where the subsets belonging to the family depend on the data at hand. For clarity, we denote the predefined families by $\mathcal{F}_-$ and the data-driven ones by $\mathcal{F}_n$.

4 | PREDEFINED FAMILIES

In this section, we state properties of the test procedure described in Section 3 for predefined families, with proofs given in Web Appendices A and B.

Theorem 4.1. *Let* $\mathcal{F}_-$ *be a predefined nonempty family of Lebesgue-measurable subsets of domain* D. *Let* $\widetilde{p}_n(t;\mathcal{F}_-)$, $t \in D$, *be the adjusted p-value function in (6). If the tests of* H_0^S *against* H_1^S *are valid (asymptotically valid) for all* $S \in \mathcal{F}_-$, *then, the test procedure based on* $\widetilde{p}_n(t;\mathcal{F}_-)$, $t \in D$, *has a control (asymptotic control) of the FWER restricted to the family* $\mathcal{F}_-$.

Theorem 4.2. *Let* $\mathcal{F}$ *be a nonempty family of Lebesgue-measurable subsets of the domain* D. *Assume that the cardinality of family* $\mathcal{F}$ *is finite. Further, assume that all* $S \in \mathcal{F}$ *are either compact sets or a finite union of compact sets. If the tests of* H_0^S *against* H_1^S *are consistent for all* $S \in \mathcal{F}$, *the test procedure based on* $\widetilde{p}_n(t,\mathcal{F})$ *in (6) is consistent.*

Theorem 4.1 states that if the family is fixed, the probability of wrongly detecting a set where the null hypothesis is actually true is bounded by α for every set included in the family $\mathcal{F}_-$. Theorem 4.2 states the conditions under which the test procedure is consistent. Observe that the latter result is valid for both predefined and data-driven families $\mathcal{F}$.

The remainder of this section discusses test procedures for particular choices of predefined families $\mathcal{F}_-$, and theoretical properties of corresponding adjustment procedures. We focus on the case when $D = [a,b]$, leaving the discussion about higher dimensions to Section 7.2.

Global testing

Suppose that the family consists only of the whole domain, $\mathcal{F}_{Glob} := \{D\}$. The corresponding test procedure performs one test over D and assigns its p-value to all points of D, with $\widetilde{p}_n(t;\mathcal{F}_{Glob}) \equiv p_n^D$, for all $t \in D$. From Theorem 4.1. it follows that if the test on D is valid, this method has a weak control of the FWER. The consistency of the procedure follows directly from the consistency of the test. However, a global test cannot provide strong control of the FWER. Further, since the adjusted p-value function is constant, it cannot be used to select specific parts of the domain responsible for the rejection of the null hypothesis.

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Borelwise testing

The Borelwise testing procedure (BWT) is based on the choice $\mathcal{F}_{BWT} := \mathcal{B}(D)$, where $\mathcal{B}(D)$ denotes all Borel sets of nonzero measure of D. Borel subsets of zero measure are excluded since the test statistic (1) is not definite on such sets. The resulting procedure is the continuous extension of the closed testing procedure (see, e.g., Marcus *et al.*, 1976) that has been proposed in multivariate analysis. If all tests are valid, Theorem 4.1 implies that the BWT has a strong control over the FWER. The adjusted p-value function for this method is constant, with $\widetilde{p}_n(t;\mathcal{F}_{BWT}) \geq \max_{t\in D} p_n(t)$ (Proposition 1 in Web Appendix B). Hence, the BWT is not consistent and cannot be used for domain selection.

Partition-closed testing

Assume that interest lies in performing tests on an a priori selected partition of the original domain. Let $\{S_j\}_{j=1}^J$ for some finite $J \in \mathbb{N}_+$ define the sets of the partition, satisfying $S_j \subseteq D$, $S_j \cap S_{j'} = \emptyset$ for all $j \neq j'$, and $\bigcup_{j=1}^J S_j = D$. Assume that S_j is Lebesgue-measurable for all j. Then, the partition-closed testing procedure (PCT; Vsevolozhskaya *et al.*, 2014) is the inferential procedure based on a family containing all possible unions between sets S_j, with $\mathcal{F}_{PCT,J} = \{\cup_{j\in\mathbb{I}} S_j\}_{\mathbb{I}\subseteq\{1,\ldots,J\}}$. From Theorem 4.1, it follows that the PCT procedure has a control of the FWER restricted to family $\mathcal{F}_{PCT,J}$ when the tests are valid. For every finite J, the PCT method is consistent, by Theorem 4.2 if the tests on subsets are consistent. Since the method is based on performing tests on unions of sets S_j, the adjusted p-value $\widetilde{p}_n(t;\mathcal{F}_{PCT,J})$ is a stepwise constant function attaining the same value for all points belonging to the same element of the partition. If for some j, we reject the null hypothesis on S_j, we only know that S_j presents a statistically significant deviation from the null hypothesis in at least one of its points. With this method, it is not possible to decide which set of points within this subset that are responsible for the rejection of H_0. The practical use of the method is highly dependent on the choice of $\{S_j\}_{j=1}^J$. Consider two uniform partitions of the domain D, the first of size $J_0, J_0 \in \mathbb{N}_+$ and the second of size $J_1 = kJ_0$, for an arbitrary $k \in \mathbb{N}_+, k > 1$. By definition, the adjusted p-value function for the PCT method based on the partition of size J_1 cannot be smaller than the one corresponding to size J_0. Moreover, if at any $t_0 \in D$ the unadjusted p-value function is above the significance level, the corresponding adjusted p-value function increases with k, and at some point exceeds the significance level on the whole domain, resulting in no domain selection. Note that if the measure of all elements of the partition goes to zero (as $J \to \infty$) the PCT and BWT methods coincide, and for $J = 1$ the PCT method coincides with the global testing.

Intervalwise testing

IWT (Pini and Vantini, 2017) is based on performing a test on every interval of the (one-dimensional) domain. The method fits under the unified framework with family $\mathcal{F}_{IWT} = \{[t_1, t_2] : t_2 > t_1\}_{t_1,t_2\in D}$. By Theorem 4.1, the test procedure has a control of the FWER restricted to $\mathcal{F}_{IWT}$ when valid tests are used. The attained intervalwise control of the FWER is in-between the weak and the strong control. Further, the pointwise test statistic is a continuous function, and the test statistic (1) is continuous with respect to the limits of integration. This implies that $\widetilde{p}_n(t;\mathcal{F}_{IWT})$ is continuous on D, providing us with a tool for domain selection. Similar methods can be defined by replacing intervals with more complex subsets. An apparently straightforward extension of IWT would be families that also include countable unions of intervals. However, such a generalization does not lead to a method with desired properties. Indeed, for a fixed integer K, consider the testing procedure based on the family $\mathcal{F}_K = \{\cup_{j=1}^K [t_{1j}, t_{2j}] : t_{2j} > t_{1j}\}_{t_{1j},t_{2j}\in D,\, j=1,\ldots,K}$, that is, the family of all possible unions of at most K disjoint intervals. It can be shown (see Proposition 2 in Web Appendix B) that the adjusted p-value function $\widetilde{p}_n(t;\mathcal{F}_K)$ is such that, for all $K \geq 2$, $\widetilde{p}_n(t;\mathcal{F}_K)$ is constant on D and such that $\widetilde{p}_n(t;\mathcal{F}_K) \geq \max_{t\in D} p_n(t)$, making the method unsuitable for domain selection. Furthermore, for all $K < \infty$, $\widetilde{p}_n(t;\mathcal{F}_K)$ is not provided with a finite-sample strong control of the FWER.

5 | DATA-DRIVEN FAMILIES

Section 4 shows that in the case of predefined families it is not possible to guarantee both the possibility of performing domain selection and strong control of the FWER. In the following, we show that, with data-driven families, it is possible to identify families that provide an asymptotically strong control of the FWER while allowing for domain selection.

Thresholdwise testing

The thresholdwise testing (TWT) performs tests on a family, $\mathcal{F}_{TWT_{J,n}}$, which is constructed based on the unadjusted p-value function, and thus data dependent. The family is constructed in the following way: Analogously to the PCT, consider a partition of the domain $\{S_j\}_{j=1}^J$ and the corresponding family of subsets $\mathcal{F}_{PCT,J}$. We introduce the discretized version of the unadjusted p-value function as $p_{J,n}(t) = p^{S_{j^*}}$, where j^* is such that $t \in S_{j^*}$, and thus $p_{J,n}(t)$, $t \in D$ is piecewise constant. The next step is to determine the family of subsets on which the tests are being performed. In the PCT case, the family is $\mathcal{F}_{PCT,J}$ and we would perform 2^J tests. For the TWT

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

procedure, we define a much smaller family $\mathcal{F}_{TWT_{J,n}}$ which is data dependent. It consists of the sublevel and superlevel sets of the discretized unadjusted p-value function. Formally,

$$\mathcal{F}_{TWT_{J,n}} = \big\{\{t \in D : p_{J,n}(t) \le y,\}\ \{t \in D : p_{J,n}(t) > y\}\big\}_{y\in[0,1]} \tag{7}$$

From the construction of $p_{J,n}(t), t \in D$, it is straightforward to see that $\mathcal{F}_{TWT_{J,n}} \subset \mathcal{F}_{PCT,J}$ and that the maximum number of elements in $\mathcal{F}_{TWT_{J,n}}$ is $2J$. With such a choice, the adjusted p-value function $\widetilde{p}_n(t;\mathcal{F}_{TWT_{J,n}})$ as defined in (6) is a piecewise constant, satisfying $\widetilde{p}_n(t;\mathcal{F}_{TWT_{J,n}}) = \max_{S\in\mathcal{F}_{TWT_{J,n}}: t\in S} p_n^S$. Here the supremum in definition (6) is replaced by a maximum since the discretized unadjusted p-value function is a piecewise constant on a finite partition and hence attains only a finite number of levels.

For finite n, and when the tests are valid, TWT has a weak control of the FWER, since $D \in \mathcal{F}_{TWT_{J,n}}$. Naturally, given the data the TWT procedure with valid tests also provides a finite sample control of the FWER restricted to $\mathcal{F}_{TWT_{J,n}}$. However, by definition, the partition is data dependent as the sets over which we control the error change between samples. The strength of the TWT is that control of the FWER restricted to $\mathcal{F}_{PCT,J}$ is attained asymptotically, for asymptotically valid and consistent tests (see Theorem 5.1). The proof of the theorem is given in Web Appendix A.

Theorem 5.1. *Let $\mathcal{F}_{TWT_{J,n}}$ be the TWT family, based on the partition $\{S_j\}_{j=1}^J$. Assume that for all $S \in \mathcal{F}_{PCT,J}$, the tests of H_0^S against H_1^S are asymptotically valid and consistent. Then, the test procedure based on the adjusted p-value function $\widetilde{p}_n(t, \mathcal{F}_{TWT_{J,n}})$ has an asymptotic control of the FWER restricted to $\mathcal{F}_{PCT,J}$.*

The conditions of Theorem 4.2 are met if the tests are consistent for all $S \in \mathcal{F}_{PCT,J}$, since for finite J the family is finite, and all subsets in the family are composed of a finite union of compact sets. This implies that the TWT procedure is consistent. The resolution of the domain selection process is related to the coarseness of the partition $\{S_j\}_{j=1}^J$, similarly to PCT. In both cases, the largest subset we can provide a control of is $S_{0,J}$ which is the biggest set included in $\mathcal{D}_0$ that can be constructed as a union of elements of the partition. For finite J, $S_{0,J}$ is possibly smaller than $\mathcal{D}_0$, so the control provided by TWT is weaker than the asymptotic strong control of the FWER. In practice, however, by refining the partition, the difference can be made arbitrarily small. In general, one would like to increase the value of J in order to have a good approximation of the functional data and of the set $\mathcal{D}_0$ where the FWER is controlled, even though increasing the size of the partition can in principle decrease the power of the method, since a larger number of tests would be involved in the maximization. The effect of changing the partition size J is explored in a simulation study described in Web Appendix F. It illustrates that when J is sufficiently large to well approximate $\mathcal{D}_0$ by $S_{0,J}$, the method continues to have similar power for larger J.

As discussed earlier, increasing J has a significantly negative effect on power and domain selection capability for the PCT method, due to the exponential number of tests performed. It is illustrated by the simulation study in Section 6 that compares the performance of all the methods within the unified framework described in Sections 4 and 5, as the sample size grows. The study confirms the already mentioned pros and cons of the methods and shows how the power (sensitivity) of all other methods except BWT increases with the sample size. The TWT procedure is by construction more powerful than PCT, since $F_{TWT_{J,n}} \subset F_{PCT_J}$, and the number of tests increase linearly, making it suitable for high-resolution domain selection. The computational costs of TWT are not affected by the dimensionality of the domain. In the case of multidimensional domains, one only has to ensure that the partition can approximate the sets $\mathcal{D}_0$ and $\mathcal{D}_1$. This makes TWT naturally suited to deal with functional data defined on multidimensional domains or even on smooth manifolds (cf. Section 7.2). Alternative data-driven families can be constructed using preimages of the unadjusted p-value function, corresponding to a suitable family of subsets of the codomain [0,1]. Such families can be shown to share the same asymptotic properties as the TWT method, and are discussed in Web Appendix C.

6 | SIMULATION STUDIES

This simulation study has two aims. First, the performance of the methods within the general framework is compared in a finite sample setting. Second, we compare the performance of the TWT method with some additional methods provided in the literature.

Simulation model

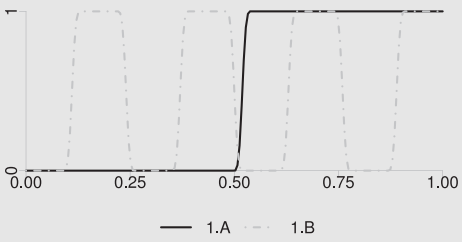
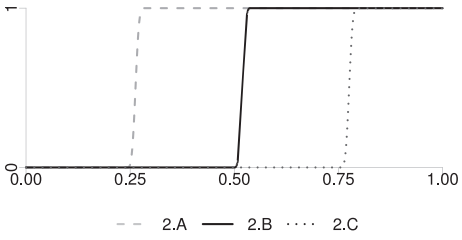
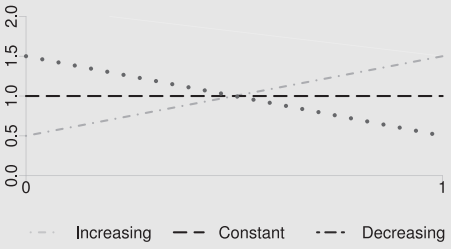
For both simulation studies, the inferential problem at hand is the comparison of means of two functional populations and we utilize the same underlying model. We consider equal size samples of two groups:

$$y_{ij}(t) = \mu_i(t) + \varepsilon_{ij}(t) \quad i = 1, \dots, n,\ j = 1, 2,\ t \in D = [0, 1].$$

The error functions $\varepsilon_{ij}(t)$ have zero mean and are independent between individuals and populations. We simulate them by simulating the coefficients of a cubic B-spline basis expansion with 81 basis functions and

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 The overview of parameter values used in the two simulation studies

Parameter	Meaning	Values
Simulation 1		
n	Samples size	5, 10, 15, 20, 30, 40
d	Effect size	1, 2, 3
$\widetilde{\mu}_2(t)$	Prototype for the mean	
$\sigma(t)$	Standard deviation function	
Simulation 2		
n	Samples size	5, 10, 15, 20, 30, 40
d	Effect size	1, 2, 3
$\widetilde{\mu}_2(t)$	Prototype for the mean	
$\sigma(t)$	Standard deviation function	

equally spaced knots, from a multivariate Gaussian distribution: $\varepsilon_{ij}(t) = \sigma(t)\sum_{k=1}^{80} c_{ijk}B_k(t)$, where $c_{ijk} \sim N(\mathbf{0}, \Sigma)$, $B_k(t), k = 1, \ldots, 80$ are B-spline basis functions and $\sigma(t)$ is a standard deviation function. We assume that the basis coefficients are correlated according to a squared exponential covariance function, that is,

$$[\Sigma]_{k_1,k_2} = \mathrm{Cov}(c_{ijk_1}, c_{ijk_2}) = \exp\left(-200\left(\frac{k_1 - k_2}{80}\right)^2\right), \quad k_1, k_2 = 1, \ldots, 81.$$

In all simulations, we use $\mu_1(t) = 0$ while we consider multiple scenarios for $\mu_2(t) = d\widetilde{\mu}_2(t)$, with varying d representing the effect size and $\widetilde{\mu}_2(t)$ representing the prototype for the mean. All prototypes are obtained using the same cubic B-spline basis, whose coefficients are sequences of zeroes and ones. In the first simulation study, we consider a division of the domain into two equisized parts, $\mathcal{D}_0$ and $\mathcal{D}_1$, using two scenarios. In the first scenario, (1.A), $\mathcal{D}_0$ is an interval, while in the second scenario, (1.B), $\mathcal{D}_0$ and $\mathcal{D}_1$ are composed of eight alternating intervals. In the second simulation study, we consider three scenarios (2.A, 2.B, and 2.C), where the domain is divided into two intervals $\mathcal{D}_0$ and $\mathcal{D}_1$, and we vary the proportion of the domain corresponding to $\mathcal{D}_0$. A summary of the parameters and their values for both studies is presented in Table 1.

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

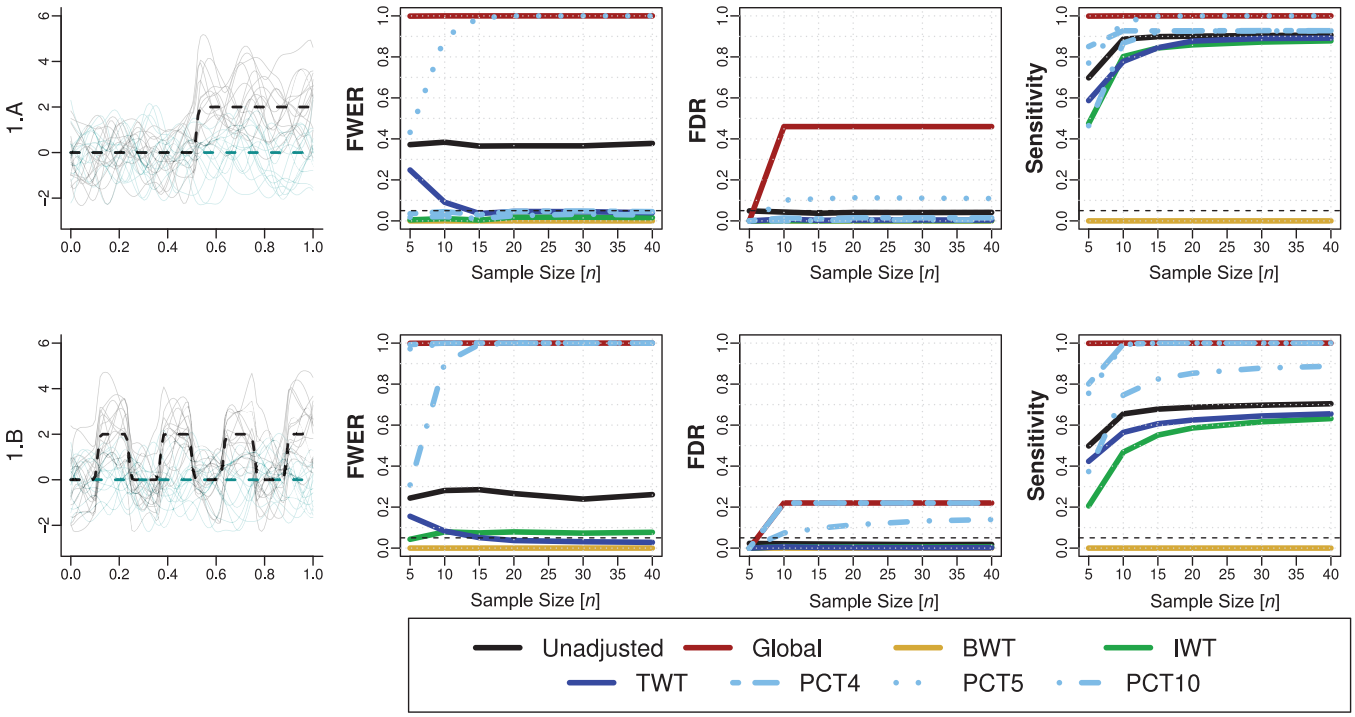


FIGURE 1 Results for simulation study 1 with effect size $d = 2$ and constant standard deviation function. Examples of $n = 15$ sample functions from both populations (distinguished by color) are presented together with their corresponding mean functions (first column). Effect of increased sample size n on the estimated FWER (second column), FDR (third column), and sensitivity (fourth column) for the introduced methods in the two scenarios. Line colors correspond to different methods, while line types correspond to different sizes of the partition for the PCT method. The dashed horizontal line corresponds to the nominal level $\alpha = 0.05$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

We test the two sample mean equality hypothesis using permutation tests with the pointwise test statistics $T_n(t) = (\overline{y}_1(t) - \overline{y}_2(t))^2$. We compare the performance of the methods by estimating FWER, FDR, and sensitivity by their empirical correspondence based on 1000 simulated experiments. For details on the definition and used estimates, as well as details of implementations, we refer to Web Appendix D.

6.1 | Simulation study 1: Comparison of the methods within the unified framework

Figure 1 presents the dynamics of the estimated measures for $d = 2$ as a function of n, with $\alpha = 0.05$. As expected, the sensitivity of all the methods, except BWT, increases as n increases. BWT is the only procedure always controlling the FWER. In practice, though, BWT does not detect any significant differences and hence is not of practical use. The IWT and PCT procedures control the FWER only if the underlying partition into $\mathcal{D}_0$ and $\mathcal{D}_1$ can be captured by the corresponding family of subsets, so the provided control is not strong. In scenario A.1, since the null hypothesis is true on an interval, IWT results in a finite sample control of the FWER. The interval can also be constructed using a partition defined by the PCT method with $J = 4$ and 10, but not with $J = 5$. In scenario A.2, none of the PCT partitions result in a separation of $\mathcal{D}_0$ and $\mathcal{D}_1$ and therefore no control is provided. TWT is the only method that possibly allows the selection of portions of the domain and provides asymptotically strong control of the FWER. This control is here reached for a reasonably small sample size (i.e., $n \approx 30$), which further supports its possible usefulness in statistical practice. Finally, as expected from theory, FDR is controlled by all procedures controlling the FWER. Since FDR is generally lower than FWER, in a few cases, procedures not controlling FWER control the FDR instead (e.g., IWT in scenario A.2), even though this is not supported by theory and could be a consequence of the parameter choice. The results presented are inherently dependent on the effect size used in the simulation studies. In Web Appendix E, we present the effect of changing the effect size on the performance of the method. As expected, increasing the effect size speeds up the convergence of TWT to the asymptotic strong control of the FWER, while lowering the value of this parameter implies higher sample sizes are required for attaining the control.

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

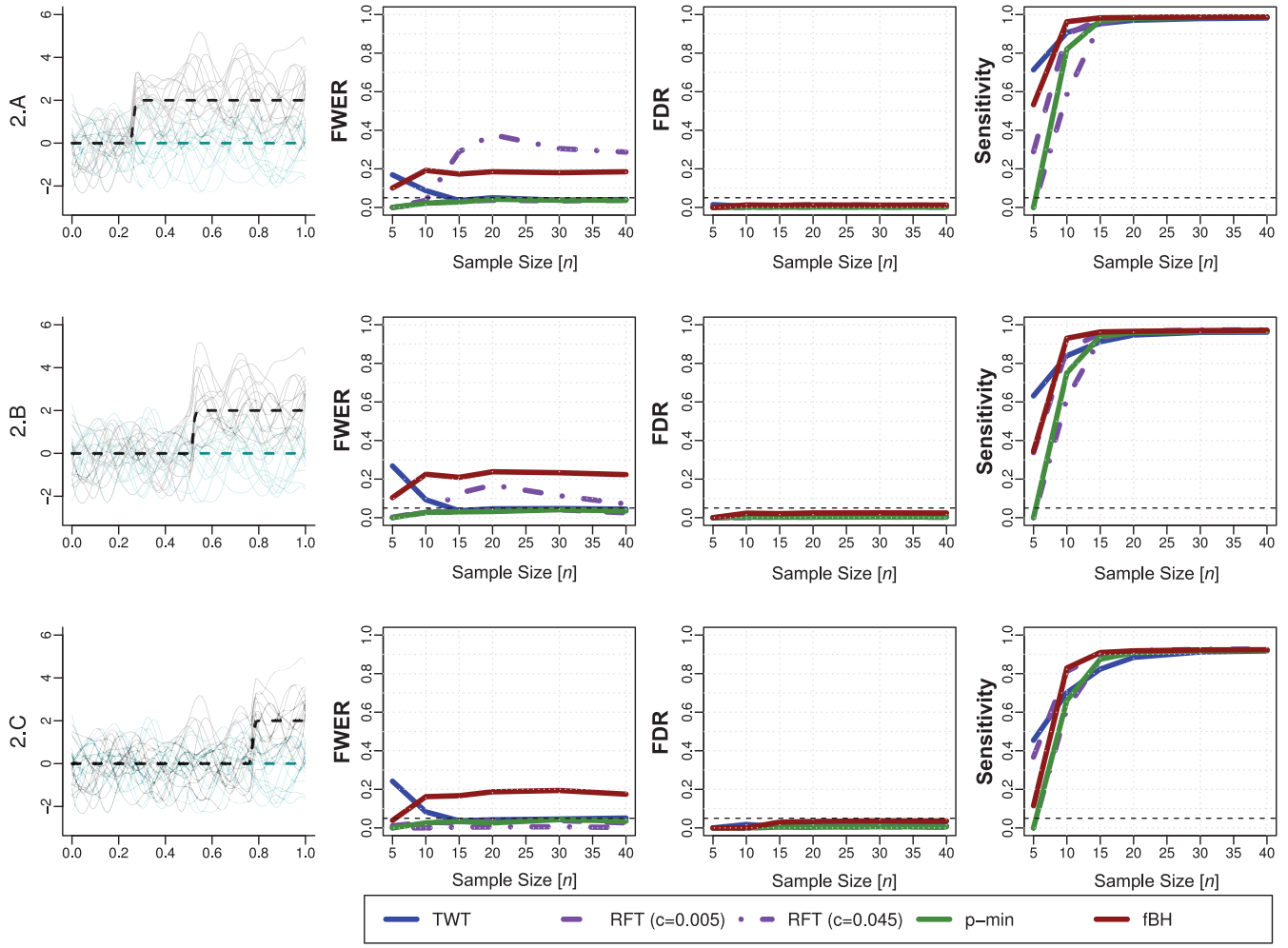


FIGURE 2 Results for simulation study 2 with effect size $d = 2$ and constant standard deviation function. Examples of $n = 15$ sample functions from both populations (distinguished by color) are presented together with their corresponding mean functions (first column). Effect of increased sample size n on the estimated FWER (second column), FDR (third column), and sensitivity (fourth column) for the compared methods in the three scenarios with different portions of $\mathcal{D}_0$ and $\mathcal{D}_1$. Scenarios 2.A, 2.B, and 2.C correspond to 25%, 50%, and 75% of the domain corresponding to $\mathcal{D}_0$, respectively. Line colors correspond to different methods, while line types correspond to different values of parameter c in the RFT method. The dashed horizontal line corresponds to the nominal level $\alpha = 0.05$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

6.2 | Simulation study 2: Comparison with alternative methods

In this study, we compare the performance of TWT, being a member of our framework, with some alternative methods presented in the literature. We consider a method introduced in Cox and Lee (2008) aiming at control of the FWER using the permutational distribution of the minimum p-value (p-min). We also consider two methods controlling the FDR: the functional Benjamini–Hochberg (fBH) method introduced in Olsen *et al.* (2021) and the method proposed in Perone Pacifico *et al.* (2004) based on random field theory (henceforth denoted RFT). The RFT method includes a parameter $c \in (0, \alpha)$ which, while keeping the FDR control at level α, affects the power of the resulting procedure. In the simulation studies, we compare the performance of the method for two distinct values of this parameter (0.1α and 0.9α). Here we present the effect of varying the size of $\mathcal{D}_0$ and $\mathcal{D}_1$, with the effect size $d = 2$ and constant variance.

Figure 2 shows that in this scenario as n increases the strong FWER control is attained asymptotically by all methods except fBH, while the FDR is controlled by all methods. The RTF method is sensitive to the choice of the parameter c, and even though the FDR is always controlled, the FWER control is not guaranteed for the higher value of c for smaller samples. The p-min method controls the FWER in all cases. However, recent studies (Mrkvička

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

et al., 2022) have shown that for high-dimensional data the power of the method decreases drastically. After reaching FWER control, TWT shows a similar sensitivity as p-min. Additional scenarios are considered in Web Appendix E, where we study the effect of variance heterogeneity and effect size. In general, we see the expected effect of the signal-to-noise ratio on all of the methods and the main conclusions remain unchanged.

7 | REAL DATA APPLICATIONS

7.1 | Knee kinematic data

Our simulation study of methods within the unified framework is complemented with the analysis of one-dimensional kinematic data, elucidating how the detected regions can differ when different methods are applied. The results together with a discussion are presented in Web Appendix G.

7.2 | Analysis of diffusion magnetic resonance imaging data

In what follows, we compare the detected regions of the methods presented in the second simulation study on diffusion magnetic resonance imaging (MRI) data. A brain image is a complex spatial domain since it is a subspace of $\mathbb{R}^3$ with a complex shape. In this application, the complex domain is defined by the voxels (three-dimensional pixels of the imaged brain) that are intersected by the so-called corpus callosum (CC), which is the set of axons connecting the two hemispheres of our brain. The CC axons form a bundle that defines a two-dimensional manifold of $\mathbb{R}^3$ (see Web Figure 10 for an example).

The CC axons are intrinsically an anisotropic environment since axons can be broadly viewed as cylinders. In particular, in this study we focus on fractional anisotropy (FA), an index measuring the degree of anisotropy along brain tracts, which has been widely adopted as a proxy for quantifying axonal damage (Horsfield and Jones, 2002; Assaf and Pasternak, 2008). FA is typically quantified with two approaches: the first proposed approach is a single-tensor model (STM; Basser *et al.*, 1994) consisting of a single anisotropic component, and a more complex approach is a multicompartment model (MCM; Panagiotaki *et al.*, 2012) incorporating an additional isotropic component related to free water.

Here we propose to demonstrate that improving upon STM by using MCM does result in a significantly lower population variance of FA. To achieve this goal, we processed diffusion MRI data of 30 healthy subjects from the Human Connectome Project (Van Essen *et al.*, 2013) to obtain a reconstruction of the CC of each subject using both STM and MCM. We chose the CC because its reconstruction is relatively easy. Finally, we defined the common domain of the CC as the set of all the voxels of size 1.25 mm^3 that were intersected by the CCs of all the 30 healthy subjects, which provided us with a domain of 950 voxels in three dimensions lying on a two-dimensional manifold. For more details on STM and MCM models and the method we used for fitting them, see Web Appendix H.

To test the stability of FA, we hypothesize that the population variance should be lower when using the more complex MCM over the STM. We therefore perform a paired one-tailed permutation test using the variance ratio as the test statistic. Domain selection is of paramount importance in brain applications where we need spatial localization of the differences. We can achieve domain selection via TWT based on a discretized unadjusted p-value function evaluated on the CC voxels. For completeness, we also included all methods evaluated in the second simulation study, that is, p-min, RFT with the two choices $c = 0.1\alpha$ and $c = 0.9\alpha$, and fBH. All methods were performed on the same discrete evaluation of the data on 950 voxels, and p-values were evaluated using 5000 permutations.

Figure 3 reports the regions of the brain where significant differences are observed by the considered methods at $\alpha = 0.05$. First of all, note that the p-min method does not detect any significant difference. This is due to the drastic decrease in the power of this method for high-dimensional data (Mrkvička *et al.*, 2022). Only when increasing the number of permutations to 10,000, the method starts detecting some differences. We would expect to obtain more significant differences when increasing the number of permutations, at a cost of a significant increase in computational time. TWT detects instead a large region, which is comparable with respect to the one detected by fBH, and it is substantially larger than the one detected by the RFT method with both choices of c. This latter result could be related to the lower power of the RFT method with respect to fBH also observed in simulation study 2. Finally, note that even though the regions detected with TWT-adjusted p-values and unadjusted p-values are very similar, TWT is performing a substantial adjustment of p-values, which can be seen in Web Figure 11.

The TWT approach identifies two symmetric areas (one in each brain hemisphere) where the FA variance cannot be claimed to be significantly lower in the MCM with respect to the STM. This is very interesting from a neurological perspective because these two areas are precisely the regions where the CC tract crosses with two other well-known tracts, namely the superior longitudinal fasciculus and the pyramidal tract. This shows that in these regions

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

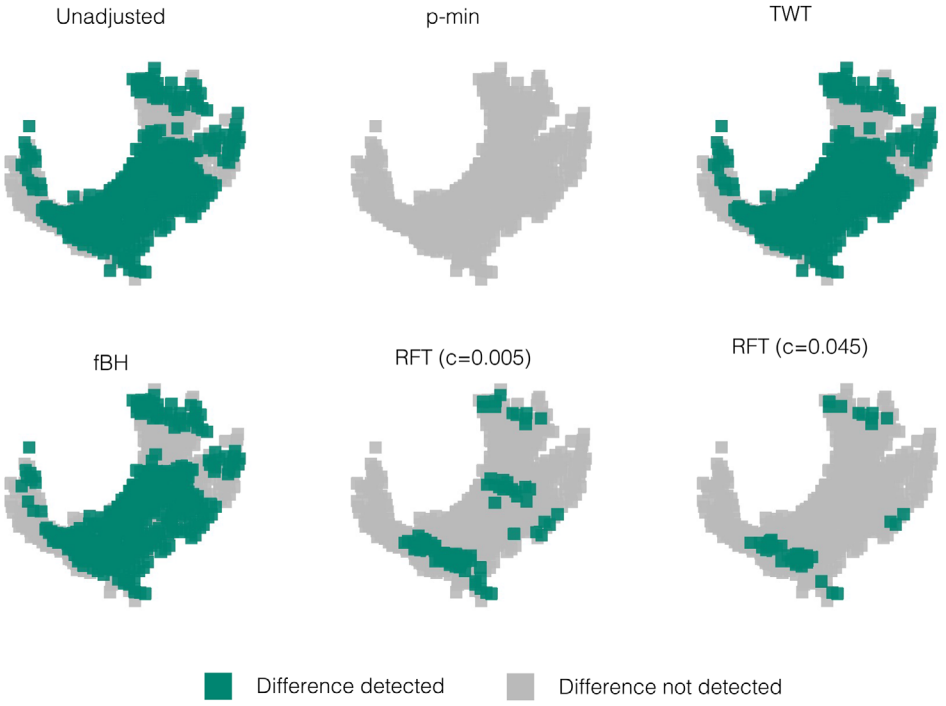


FIGURE 3 Voxels where the null hypothesis of equality of the variances of the two populations is rejected (green) and not rejected (gray) according to the different methods. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

the introduction of the free water–related isotropic component is not sufficient to reduce the population variance. Hence, the addition of a second anisotropic component possibly would be needed to model the additional tracts.

The running time was about 4 min for TWT, 3 min for each of the RFT methods, and 7 min for p-min; the timing was evaluated on a 2.6-GHz Quad-core i7 processor, with 16 GB 2133 MHz LPDDR3 RAM and 512 Gb SSD hard drive.

8 | CONCLUSIONS

In this paper, we introduce a general framework for local inference for functional data, where subsetwise test procedures on the functional data perform domain selection while controlling the FWER. We investigate the properties of the test procedures (methods) within the framework. The test procedures are based on two types of families (of subsets of the domain): *predefined*, appearing in the existing literature, and *data driven* proposed in this paper. We show that some serious practical limitations of the methods based on the predefined families can be overcome with the data-driven families. The possibility of selecting significant regions in a possibly complex domain, while retaining asymptotic FWER control restricted to a family generated by the predefined data resolution, is presented and illustrated in two application-focused examples.

ACKNOWLEDGMENTS

This work was supported by the Swedish Research Council (grant numbers 2016-02763 and 340-2013-5203). We are grateful to the associate editor and a reviewer for their valuable input. We are also grateful to Charlotte K. Häger for providing the kinematic data in Section 7.1. Data in Section 7.2 were provided by the Human Connectome Project (Van Essen *et al.*, 2013).

DATA AVAILABILITY STATEMENT

Knee kinematic data supporting the findings of this paper (Section 7.1) are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

MRI data supporting the findings of this paper (Section 7.2) are openly available in the Human Connectome Project, WU-Minn Consortium (principal investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Data are available at http://www.humanconnectomeproject.org/data/.

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ORCID

Konrad Abramowicz https://orcid.org/0000-0002-9040-6674
Alessia Pini https://orcid.org/0000-0001-9235-3062
Lina Schelin https://orcid.org/0000-0001-7917-5687
Sara Sjöstedt de Luna https://orcid.org/0000-0003-1591-5716
Aymeric Stamm https://orcid.org/0000-0002-8725-3654
Simone Vantini https://orcid.org/0000-0001-8255-5306

REFERENCES

Abramowicz, K., Häger, C.K., Pini, A., Schelin, L., Sjöstedt de Luna, S. and Vantini, S. (2018) Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics*, 45, 1036–1061.

Assaf, Y. and Pasternak, O. (2008) Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *Journal of Molecular Neuroscience*, 34, 51–61.

Basser, P.J., Mattiello, J. and LeBihan, D. (1994) MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66, 259–67.

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

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

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

Crainiceanu, C.M., Staicu, A.-M., Ray, S. and Punjabi, N. (2012) Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine*, 31, 3223–3240.

Degras, D. (2017) Simultaneous confidence bands for the mean of functional data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9, e1397.

Fan, J. and Zhang, J.T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society Series B*, 62, 303–322.

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

Holmes, A.P., Blair, R.C., Watson, J.D.G. and Ford, I. (1996) Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, 16, 7–22.

Horsfield, M. and Jones, D. (2002) Applications of diffusion-weighted and diffusion tensor MRI to white matter diseases— a review. *NMR in Biomedicine*, 15, 570–577.

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

Liebl, D. and Reimherr, M. (2020) Simultaneous inference for function-valued parameters: a fast and fair approach. In Aneiros, G., Horová, I., Hušková, M. and Vieu, P. (Eds.) *Functional and High-Dimensional Statistics and Related Fields*. Cham: Springer International Publishing, pp. 153–159.

Logan, B.R. and Rowe, D.B. (2004) An evaluation of thresholding techniques in fMRI analysis. *NeuroImage*, 22, 95–108.

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

Mrkvička, T., Myllymäki, M., Kuronen, M. and Narisetty, N.N. (2022) New methods for multiple testing in permutation inference for the general linear model. *Statistics in Medicine*, 41, 276–297.

Naouma, H. and Pataky, T.C. (2019) A comparison of random-field-theory and false-discovery-rate inference results in the analysis of registered one-dimensional biomechanical datasets. *PeerJ*, 7, e8189.

Olsen, N.L., Pini, A. and Vantini, S. (2021) False discovery rate for functional data. *TEST*, 30, 784–809.

Panagiotaki, E., Schneider, T., Siow, B., Hall, M., Lythgoe, M. and Alexander, D. (2012) Compartment models of the diffusion MR signal in brain white matter: a taxonomy and comparison. *Neuroimage*, 59, 2241–2254.

Park, S.Y., Staicu, A.-M., Xiao, L. and Crainiceanu, C.M. (2017) Simple fixed-effects inference for complex functional models. *Biostatistics*, 19, 137–152.

Pataky, T.C., Abramowicz, K., Liebl, D., Pini, A., de Luna, S.S. and Schelin, L. (2021) Simultaneous inference for functional data in sports biomechanics. *AStA Advances in Statistical Analysis* . Published online. https://doi.org/10.1007/s10182-021-00418-4

Perone Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2004) False discovery control for random fields. *Journal of the American Statistical Association*, 99, 1002–1014.

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

Pini, A. and Vantini, S. (2017) Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29, 407–424.

Rathnayake, L.N. and Choudhary, P.K. (2016) Tolerance bands for functional data. *Biometrics*, 72, 503–512.

Reiss, P.T., Huang, L. and Mennes, M. (2010) Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics*, 6, 28.

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

Telschow, F.J. and Schwartzman, A. (2022) Simultaneous confidence bands for functional data using the Gaussian kinematic formula. *Journal of Statistical Planning and Inference*, 216, 70–94.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K. and WU-Minn HCP Consortium, (2013) The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80, 62–79.

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

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M. and Nichols, T.E. (2014) Permutation inference for the general linear model. *NeuroImage*, 92, 381–397.

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

SUPPORTING INFORMATION

Web Appendices, and Figures referenced in Sections 4–7 are available with this paper at the Biometrics website on Wiley Online Library. R code implementing the proposed TWT method is available at github https://github.com/astamm/fdatest. R code for reproducing the simulated results is also available at the Biometrics website on Wiley Online Library.

How to cite this article: Abramowicz, K., Pini, A., Schelin, L., Sjöstedt de Luna, S., Stamm, A., Vantini, S. (2023) Domain selection and familywise error rate for functional data: A unified framework. *Biometrics*, 79, 1119–1132. https://doi.org/10.1111/biom.13669

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13669 by Umea University, Wiley Online Library on [04/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License