Electronic Journal of Statistics
Vol. 18 (2024) 1887–1916
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2240

Transportation-based functional ANOVA and PCA for covariance operators

Valentina Masarotto[1], **Victor M. Panaretos**[2] and **Yoav Zemel**[2]

[1] *Mathematisch Instituut, Universiteit Leiden, The Netherlands,*
e-mail: v.masarotto@math.leidenuniv.nl

[2] *Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, Suisse,*
e-mail: victor.panaretos@epfl.ch; yoav.zemel@epfl.ch

Abstract: We consider the problem of comparing several samples of stochastic processes with respect to their second-order structure, and describing the main modes of variation in this second order structure, if present. These tasks can be seen as an Analysis of Variance (ANOVA) and a Principal Component Analysis (PCA) of covariance operators, respectively. They arise naturally in functional data analysis, where several populations are to be contrasted relative to the nature of their dispersion around their means, rather than relative to their means themselves. We contribute a novel approach based on optimal (multi)transport, where each covariance can be identified with a a centred Gaussian process of corresponding covariance. By means of constructing the optimal simultaneous coupling of these Gaussian processes, we contrast the (linear) maps that achieve it with the identity with respect to a norm-induced distance. The resulting test statistic, calibrated by permutation, is seen to distinctly outperform the state-of-the-art, and to furnish considerable power even under local alternatives. This effect is seen to be genuinely functional, and is related to the potential for perfect discrimination in infinite dimensions. In the event of a rejection of the null hypothesis stipulating equality, a geometric interpretation of the transport maps allows us to construct a (tangent space) PCA revealing the main modes of variation. As a necessary step to developing our methodology, we prove results on the existence and boundedness of optimal multitransport maps. These are of independent interest in the theory of transport of Gaussian processes. The transportation ANOVA and PCA are illustrated on a variety of simulated and real examples.

MSC2020 subject classifications: Primary 62R10, 60G15; secondary 60H25, 62J10.
Keywords and phrases: Coupling, functional data analysis, Fréchet mean, Gaussian measure, optimal transport, multimarginal transport, procrustes analysis, tangent space PCA, trace-class operator.

Received December 2022.

1. Introduction

Let $\{X_{i,1}\}_{i=1}^{n_1}, \ldots, \{X_{i,K}\}_{i=1}^{n_K}$ be K independent samples of i.i.d. random elements in a separable Hilbert space $\mathcal{H}$, posessing well-defined means $\{\mu_j\}_{j=1}^{K}$

and covariances $\{\Sigma_j\}_{j=1}^K$. We consider the problem of testing the hypothesis

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_K \tag{1}$$

on the basis of the observations $\{X_{i,j}\}$ and, if H_0 is rejected, the subsequent problem of describing the main mode(s) of variation of the K underlying covariances.

This problem arises very naturally in functional data analysis, i.e. when $\mathcal{H}$ is taken to be a function space (for instance $L^2[0,1]$ or a reproducing kernel Hilbert subspace thereof), and one is interested in discerning whether K different groups of functions manifest the same type of dispersion relative to their mean. For instance, the functions could be curves representing DNA minicircles (Panaretos et al. [30], Kraus and Panaretos [25], and Tavakoli and Panaretos [42]), where different groups correspond to different base-pair sequences, and one is interested in probing for a dependence of the mechanical properties on the base pair sequence; or they could be surfaces representing the log spectrograms of short spoken words by different speakers (as in Ferraty and Vieu [14]), and one may wish to see whether there is a difference in several groups of sounds; yet a further example may be in the analysis of age-dependent wheel-running activity curves in mice, where one may wish to see whether the level of activity across age had evolved under several generation selections (Cabassi et al. [6]). What is common to all these examples is that it is not the mean structure that is suspected to differ (or at least to capture the most interesting differences); in that case, the problems would fall under the well studied topic of functional analysis of variance (see, e.g., Benko et al. [2], Zhang [45], Cuesta-Albertos and Febrero-Bande [7], Górecki and Smaga [18]). Rather it is the fluctuations around the means μ_j, as encapsulated by the operators Σ_j, in what could be termed a functional covariance ANOVA.

Early contributions in this direction focus on the two-sample case, as in Benko et al. [3], Panaretos et al. [30], Fremdt et al. [16]. In particular, Benko et al. [3] propose a two–sample bootstrap based tests for some aspects of the spectrum of functional data, Panaretos et al. [30] consider the problem in a two-sample setting with Gaussian processes, and Fremdt et al. [16] extend to non-Gaussian. Kraus and Panaretos [25] provide resistant versions of two-sample tests, focussed on operators related to the covariance. Two-sample testing has been first extended to the K-sample case by Boente et al. [4], who propose a test based on the Hilbert–Schmidt distance between the estimated covariance operators of each population and where the critical values of the test statistics are calibrated via a bootstrap procedure. The common theme in these papers is that the covariances are contrasted with respect to the Hilbert–Schmidt metric, which corresponds to imbedding covariance operators in a larger linear space, whereas they are not closed under linear operations. Instead, covariance operators are trace-class non-negative operators, so rather than being seen as Hilbert–Schmidt operators, they are better represented as "squares" of such operators. For this reason, Pigoli et al. [33] considered the use of nonlinear metrics adapted to non-negative operators in the two-sample setting, generalising some of the work in

Dryden et al. [11] in finite dimensions. Cabassi et al. [6] extend their metric-based methodology to the K-sample case. Other contributions for the K-sample comparison are from Paparoditis and Sapatinas [31], who develop an empirical bootstrap methodology and prove its consistency when the test statistics is based on the Hilbert—Schmidt norm, and Kashlak et al. [23], who perform K-sample comparison via concentration inequalities based methods. Recently, Hlávka et al. [21] proposed a method to perform functional ANOVA based on empirical characteristic functionals. When comparing with Anderson [1], Paparoditis and Sapatinas [31] and Kashlak et al. [23], Cabassi et al. [6] report simulation results illustrating state-of-the-art performance of their method. We found that this holds true even when comparing against the recent work of Hlávka et al. [21]. In Hlávka et al. [21], a specific choice of the parameters covariance matrix yield similar conclusion to Cabassi et al. [6] when comparing covariances in a real-data example.

Pigoli et al. [33] paid particular attention to the *Procrustes* distance, which generalises a metric used to compare unlabelled shapes into a metric between covariance operators. Heuristically, the Procrustes metric aims to compare roots of two operators in the Hilbert-Schmidt distance, a natural choice since covariances are characterised as squares of Hilbert-Schmidt operators. However, there is an ambiguity as to which precise root one ought to use, and the Procrustes distance corrects for that by optimising over the square root orbits. Indeed, later work by Pigoli et al. [34] reports that the Procrustes metric offered the most natural framework to compare trace-class operators, in that it uses a map from the space of covariance operators to the linear space of Hilbert-Schmidt operators. Masarotto et al. [27] carried out a deeper study of the Procrustes metric and established a fruitful connection with the Wasserstein metric between Gaussian processes. On the one hand, this allowed them to provide a complete geometrical description of the space of covariances under the Procrustes metric, including basic results about Fréchet means; on the other hand, it established intriguing parallels with the theory of optimal transportation, offering potentially new avenues and tools for the analysis of covariance operators (see also the discussion of Pigoli et al. [34] by Panaretos [29]).

In this paper, we use precisely this novel transportation perspective to introduce a new ANOVA test, and then exploit the corresponding geometry to construct a tangent space PCA that respects the nature of the covariance operators, by representing covariance operators as transport maps. Specifically, we view the testing problem through the lens of optimally multicoupling the Gaussian processes $\{N(0,\Sigma_1),\dots,N(0,\Sigma_K)\}$, thus translating the task of testing the hypothesis H_0 in Equation (1) into that of testing whether the optimal multicoupling is "trivial". To do this we first prove that the optimal multicoupling can always be *deterministically*[1] produced by means of *bounded linear* transport maps $\{\mathbf{t}_j\}$, and this regardless of the validity of the null hypothesis. These two results are of independent interest, and are stated as Theorem 2.2. Then, given these results, we translate the task of testing the hypothesis in Equation (1)

[1]Rather than stochastically, by means of a probability measure on $\mathcal{H}^K$.

into the equivalent task of testing the hypothesis

$$H_0' : \mathbf{t}_1 - \mathscr{I} = \ldots = \mathbf{t}_K - \mathscr{I} = 0 \tag{2}$$

for $\mathscr{I}$ the identity, and so the hypothesis (2) can now be tested by means of a norm-based (e.g., operator, Hilbert–Schmidt, or nuclear) test statistic, establishing a direct analogy with classical ANOVA. The $\Delta_j = \mathbf{t}_j - \mathscr{I}$ can heuristically also be viewed as "roots" of the original covariances, albeit free of any unitary ambiguity. Though the test is motivated by the 1-to-1 correspondence between covariances and Gaussian processes, it relies in no way on a Gaussian assumption, and will be valid on any location-scatter family – indeed, it admits an interpretation purely in terms of the Procrustean geometry on covariances. Our simulation experiments indicate that the new test dominates state-of-the-art competitors, with dramatic gains in power, particularly against more challenging local alternatives. This is also explained by means of theory, and is seen to be a genuinely functional effect, with connection to the Hajek–Feldman condition. See Section 6 for more details. In terms the computation, the quantity $(K^{-1}\sum_{j=1}^{K}\mathbf{t}_j - \mathscr{I})$ in finite dimension is precisely the negative gradient of the Fréchet (sum-of-squares) functional Zemel and Panaretos [44, Theorem 1] and its computation is stable and feasible because it relies on the fast nature of the steepest descent algorithm in the space of covariances endowed with the Procrustes metric.

When the mull hypothesis is rejected, it is natural as a second step to wish to describe the variation manifested by the covariances $\{\Sigma_j\}$, or indeed obtain a parsimonious representation thereof. We show how the maps $\{\mathbf{t}_j\}$ can then readily be employed to do just that, via a tangent principal component analysis. In particular, the $\Delta_j = \mathbf{t}_j - \mathscr{I}$ can be interpreted as the logarithms of the $\{\Sigma_j\}$ at their Procrustes–Fréchet mean. The corresponding tangent space admits a Hilbertian structure with respect to a modified Hilbert–Schmidt inner product, which we use to produce a tangent space fPCA, and then retract the principal components back onto the covariance space to allow visualisation via geodesics. To the best of our knowledge, this is the first instance of a functional PCA on covariance operators that respects their intrinsic geometric features as trace-class positive operators. We describe the computational steps required to do so in Section 3, and illustrate the usefulness of the procedure on simulated data as well as a linguistic data set. All analyses were performed using R Statistical Software R Core Team [35]. The proposed methodology has been made available in the R package "fdWasserstein" (Masarotto and Masarotto [26]). The next paragraph collects the notational conventions employed throughout the paper. Proofs of our theoretical results are given in Section 6.

2. Methodology

2.1. Basic setting and notation

As stated in the introduction, we will be interested in exploring the variation in a finite collection of covariances $\{\Sigma_j\}_{j=1}^{K}$ on a real separable Hilbert space $\mathcal{H}$,

equipped with the inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, and corresponding norm $\| \cdot \| : \mathcal{H} \to [0, \infty)$. Since $\mathcal{H}$ will in principle be infinite-dimensional, we will need to review some basic definitions and notation, which can be more subtle.

Given a bounded linear operator $A : \mathcal{H} \to \mathcal{H}$, we will denote its trace (when defined) by $\mathrm{tr}A$ or $\mathrm{tr}(A)$, its adjoint by A^*, and its inverse by A^{-1}. The inverse may not be defined, or defined only on a (dense) subspace of $\mathcal{H}$. The *range* of A will be denoted by $\mathrm{range}(A) = \{Av : v \in \mathcal{H}\}$ whereas the *kernel* of A will be denoted by $\ker(A) = \{v \in \mathcal{H} : Av = 0\}$. We will say that a (possibly unbounded) operator A is *self-adjoint* if $\langle Au, v \rangle = \langle u, Av \rangle$ for all u, v in the domain of definition of A; if A happens to also be bounded, then this is equivalent to the condition that $A = A^*$.

A *non-negative* operator is a self-adjoint, possibly unbounded operator A such that $\langle Au, u \rangle \geq 0$ for all u in the domain of A. If in addition A is compact, then there exists a unique non-negative operator whose square equals A, which will be denoted by either $A^{1/2}$ or $\sqrt{A}$. The inverse square root $(A^{1/2})^{-1}$ is denoted by $A^{-1/2}$. For any bounded operator A, A^*A is non-negative. The identity operator on $\mathcal{H}$ will be denoted by $\mathscr{I}$. The operator, Hilbert–Schmidt and trace (nuclear) norms will respectively be

$$\left|\!\left|\!\left| A \right|\!\right|\!\right|_\infty = \sup_{\|h\|=1} \|Ah\|, \quad \left|\!\left|\!\left| A \right|\!\right|\!\right|_2 = \sqrt{\mathrm{tr}\,(A^*A)}, \quad \left|\!\left|\!\left| A \right|\!\right|\!\right|_1 = \mathrm{tr}\left(\sqrt{A^*A}\right)$$

and can be ordered from coarser to finer as follows

$$\left|\!\left|\!\left| A \right|\!\right|\!\right|_\infty \leq \left|\!\left|\!\left| A \right|\!\right|\!\right|_2 \leq \left|\!\left|\!\left| A \right|\!\right|\!\right|_1.$$

When $\left|\!\left|\!\left| A \right|\!\right|\!\right|_2 < \infty$ we say that A is Hilbert–Schmidt and when $\left|\!\left|\!\left| A \right|\!\right|\!\right|_1 < \infty$ we say that A is *nuclear* or *trace-class*.

Summarising, in this setting, covariances are linear operators from $\mathcal{H}$ into $\mathcal{H}$, that are *self-adjoint*, *non-negative*, and *trace-class*. As such, a covariance operator Σ on $\mathcal{H}$ can be considered as the "square" of a Hilbert–Schmidt operator: if $\left|\!\left|\!\left| B \right|\!\right|\!\right|_2 < \infty$ then B is certainly bounded, and B^*B defines a valid covariance operator.

It therefore becomes clear that covariance operators are non-linear objects, and though they can be contrasted by means of any of the three norms $\left|\!\left|\!\left| \cdot \right|\!\right|\!\right|_\infty$, $\left|\!\left|\!\left| \cdot \right|\!\right|\!\right|_2$, or $\left|\!\left|\!\left| \cdot \right|\!\right|\!\right|_1$, it may be preferable to find a means of comparison that respects this non-linear nature. In finite dimensions, this is done by means of some form of *linearisation*, i.e. the use of a transformation that substitutes a covariance pair (Σ_1, Σ_2) to be considered by an operator that can be contrasted to zero by means of one of the norms $\left|\!\left|\!\left| \cdot \right|\!\right|\!\right|_r$, $r = 1, 2, \infty$. For instance, in classical two-sampled covariance tests, two covariances Σ_1 and Σ_2 have been contrasted by means of quantities such as

$$\left|\!\left|\!\left| \Sigma_1\Sigma_2^{-1} - \mathscr{I} \right|\!\right|\!\right|_r \quad \& \quad \left|\!\left|\!\left| 2\Sigma_2(\Sigma_1 + \Sigma_2)^{-1} - \mathscr{I} \right|\!\right|\!\right|_r$$

assuming that the inverses exist (e.g., Roy [40], Kiefer and Schwartz [24], Giri [17]. In non-Euclidean statistics, covariances (Σ_1, Σ_2) have been contrasted by

means of

$$\left\| \log(\Sigma_1) - \log(\Sigma_2) \right\|_r \quad \& \quad \left\| \log(\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}) \right\|_r,$$

again, assuming that the inverses exist[2] (Dryden et al. [11]).

In infinite dimensions, however, these criteria will generally fail to be well-defined. For example, the inverse of Σ_2 will be unbounded, and there is no guarantee that $\Sigma_1\Sigma_2^{-1}$ will be bounded, except if Σ_1 and Σ_2 share some special relation. Similarly, the logarithm of a covariance operator will typically be unbounded, and unless there is a specific relation between Σ_1 and Σ_2, the logarithmic criteria will fail to be well defined. This is one of the main reasons why much of the literature on covariance operators has focussed on bypassing their nonlinear nature, and comparing them directly, e.g. by means of $\left\|\Sigma_1 - \Sigma_2\right\|_2$ (Panaretos et al. [30], Fremdt et al. [16], Boente et al. [4]).

A first step in obtaining linearisations that would yield contrasts respecting the nature of covariances, while being well-defined in infinite dimensions was made by Pigoli et al. [33] Since covariances are "squares" of Hilbert–Schmidt operators, they considered contrasting the square roots of in the Hilbert–Schmidt distance

$$\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2.$$

Observing that one could nevertheless choose roots other than the (unique) positive roots, by means of the fact that $\Sigma_2^{1/2}U(\Sigma_2 U)^* = \Sigma_2$ they arrived at the *Procrustes metric*

$$\Pi(\Sigma_1, \Sigma_2) = \inf_{U^*U=\mathscr{I}} \left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}U\right\|_2$$

which lifts the unitary ambiguity by optimising over unitary matrices, and is well defined in both finite and infinite dimensions. Indeed they use this metric to develop a two-sample test for covariance comparison. Masarotto et al. [27] further developed several key properties of this metric and its geometry, interpreting it via the optimal transportation of Gaussian processes as the L^2-Wasserstein distance between two Gaussian measures $N(0, \Sigma_1)$ and $N(0, \Sigma_2)$ on $\mathcal{H}$,

$$\Pi(\Sigma_1, \Sigma_2) = \inf_{X_i \sim N(0,\Sigma_i)} \mathbb{E}\|X_1 - X_2\|_2^2 = \left\|\Sigma_1\right\|_1 + \left\|\Sigma_2\right\|_1 - 2\mathrm{trace}\left\{\sqrt{\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}}\right\}.$$

The key observation in this paper is that the optimal transport theory developed in Masarotto et al. [27] can be directly leveraged in order to provide natural notions of "roots" (or linearisations) that:

- are unequivocally defined without any unitary ambiguity;
- are efficiently computable;
- that offer remarkable power when used in a covariance ANOVA;

[2] If more covariances $\{\Sigma_i\}$ are to be simultaneously compared, for instance in a covariance ANOVA, one could use the same contrasts, replacing Σ_1 with Σ_i and Σ_2 by the arithmetic average $n^{-1}\sum_i \Sigma_i$.

- can be used in order to obtain a natural PCA, when the equality of covariances is rejected.

These are defined via the notion of an optimal multicoupling, and are introduced in the next Section.

2.2. Optimal multicoupling and transport maps

As already stated, our strategy for testing $H_0 : \Sigma_1 = \cdots = \Sigma_K$ is to view the covariance operators through the lens of optimal multicoupling of Gaussian processes. Specifically, we observe that the collection of covariances $\{\Sigma_1, \ldots, \Sigma_K\}$ can be bijectively identified with a collection of centred Gaussian measures $\{N(0, \Sigma_1), \ldots, N(0, \Sigma_K)\}$ on the Hilbert space $\mathcal{H}$. Denote these measures as $\{\gamma_1, \ldots, \gamma_K\}$. Equality of the covariance operators thus holds true, if and only if the measures $\{\gamma_j\}$ coincide. Viewing the measures $\{\gamma_j\}$ as the marginals of a joint measure π on $\mathcal{H}^K$, one can ask what are the possible forms of π. This set of possible joint measures π is always non-empty (it always contains the product measure), and is called the set of *multicouplings* of $\{\gamma_1, \ldots, \gamma_K\}$. An *optimal multicoupling* is a multicoupling π^* such that the marginals are as tightly coupled as possible in a pairwise mean-square sense, in that it minimizes the functional

$$F(\pi) = \frac{1}{2K^2} \sum_{i,j=1}^{K} \int_{\mathcal{H}^K} \|x_i - x_j\|^2 \pi(dx_1, \ldots, dx_K).$$

Said differently, π^* is the joint distribution of collection of K Gaussian processes on $\mathcal{H}$, say $(Z_1, \ldots, Z_K)$, such that $Z_j \sim N(0, \Sigma_j)$ marginally for all $j \leq K$, while $\sum_{i<j} \mathbb{E}\|Z_i - Z_j\|^2$ is minimized. Existence of finite second moments of Gaussian measures (a consequence of Fernique's [13] theorem) implies that $F(\pi)$ is finite for any multicoupling π. It can be shown that an optimal multicoupling of Gaussians always exists (Masarotto et al. [27]). We say that such an optimal multicoupling π^* is manifested by (deterministic) transport maps if the collection $(Z_1, \ldots, Z_K) \sim \pi^*$ can be generated by taking a single process Z, and a collection of deterministic maps $\mathbf{t}_j : \mathcal{H} \to \mathcal{H}$ such that

$$(Z_1, \ldots, Z_K) \stackrel{d}{=} (\mathbf{t}_1(Z), \ldots, \mathbf{t}_K(Z)).$$

In other words, an optimal multicoupling π^* is generated by deterministic maps if it is supported on the graph of a vector-valued function from $\mathcal{H}$ to $\mathcal{H}^K$. It is a priori unclear whether a deterministic multicoupling exists in general, and if it does, whether the maps $\mathbf{t}_j$ are bounded. but it is not hard to see that it will exist under the null H_0 and that it will be "trivial":

Lemma 2.1. *The equality $\Sigma_1 = \cdots = \Sigma_K$ holds true if and only if the (unique) optimal multicoupling of $(\gamma_1, \ldots, \gamma_K)$ can be achieved by transport maps satisfying $\mathbf{t}_1 = \cdots = \mathbf{t}_K$.*

The maps $\mathbf{t}_j$ are called transport maps because they can be thought of as "transporting" the (unspecified) law of Z to that of Z_j. The lemma suggests that we can detect departures from the hypothesis $\{H_0 : \Sigma_1 = \cdots = \Sigma_K\}$ by focussing on departures from the hypothesis $\{H_0' : \mathbf{t}_1 = \cdots = \mathbf{t}_K\}$. But to even speak of departures from H_0', we must be assured that such maps exist even under the alternative regime, and this existence is not a priori guaranteed (see Conjecture 17 and the discussion in Section 12 of [27]). Furthermore, to quantify the extent of departures from the null, we need to make sure that the multicoupling maps not only exist, but are bona fide bounded linear operators over all of $\mathcal{H}$, and can thus be contrasted by appropriate norms.

Our main theoretical result, in the form of the following theorem, shows that a multicoupling can always be realised by means of bounded deterministic maps, a result that is of independent interest in optimal transport in its own right.

Theorem 2.2. *Let $\{\gamma_1, \ldots, \gamma_K\}$ be an arbitrary finite collection of Gaussian measures on $\mathcal{H}$ with mean zero. Then there exists an optimal multicoupling of $\{\gamma_j\}_{j=1}^K$ manifested by deterministic transport maps $\mathbf{t}_j : \mathcal{H} \to \mathcal{H}$ that are bounded non-negative linear operators satisfying $|||\mathbf{t}_j|||_\infty \leq K$, for all $j \leq K$.*

Although the optimal coupling π^* is typically unique, its representation in terms of the maps is not. For instance, if π^* is manifested as the law of $(\mathbf{t}_1(Z), \ldots, \mathbf{t}_K(Z))$, it may also be represented as $(2\mathbf{t}_1(Z/2), \ldots, 2\mathbf{t}_K(Z/2))$. It is natural to take $Z \sim N(0, \overline{\Sigma})$, where $\overline{\Sigma}$ is a centre, i.e., the Fréchet mean of $\Sigma_1, \ldots, \Sigma_K$ with respect to the Procrustes metric. This choice forces the maps $\mathbf{t}_j$ to have mean identity (see (4) below), and in particular they must be the identity under the null (1), so that (2) holds. Using this convention, the existence and boundedness result in Theorem 2.2 opens the way for a testing procedure: the deviations $\Delta_j = \mathbf{t}_j - \mathscr{I}$ are all self-adjoint and bounded, but no longer restricted to be non-negative. When H_0 is valid, Lemma 2.1 implies that $\Delta_j = \mathbf{t}_j - \mathscr{I} = 0$ for all j. Under the alternative, at least one Δ_j is non-zero. We can thus replace the null hypothesis

$$H_0 : \Sigma_1 = \cdots = \Sigma_K$$

by the equivalent hypothesis

$$H_0' : \Delta_1 = \cdots = \Delta_K = 0$$

viewing the Δ_j as elements of a linear space, and reducing the original testing problem to a more traditional linear functional ANOVA setting. Since the Δ_j are guaranteed to be bounded (Theorem 2.2), they can certainly be contrasted to 0 using the operator norm. However, one can devise even more powerful procedures by measuring the size of Δ_j in a stronger norm, such as the the Hilbert–Schmidt norm, or even the trace norm. In the finite-dimensional setting this choice of norm will typically not make much difference, since all norms are equivalent. But in the infinite-dimensional case, a finer norm will detect subtle departures from the null. For instance, if $K = 2$ and $\Sigma_2 = \delta^2\Sigma_1$ for some $\delta \geq 0$, then one can show that for $j = 1, 2$, $|||\Delta_j|||_\infty = |1 - \delta|/(1 + \delta) \leq 1$ while

$\|\Delta_j\|_2 = \infty$ unless $\delta = 1$. Similarly, if the Δ_j are Hilbert–Schmidt but not trace class, their trace norm will be infinite, promising to furnish high power even against very local alternatives. This genuinely functional phenomenon is not unlike the possibility of perfect discrimination of Gaussian processes (Feldman [12], Hájek [19], Rao and Varadarajan [39]; see Section 6 for a more detailed discussion). It is demonstrated empirically in our later simulations. If there are only $K = 2$ populations, then in view of (4), $\Delta_2 = -\Delta_1$ and the test statistic is $\|\Delta_1\|_r$. When comparing more than two covariance operators, the criteria $\|\Delta_j\|_r$ will need to be combined into a single criterion (e.g., by taking their supremum over j or by summation).

How does one concretely construct a deterministic multicoupling $\{\mathbf{t}_j\}$, and hence the $\{\Delta_j\}$ in practice? In proving Theorem 2.2 we establish the existence and boundedness of the maps

$$\mathbf{t}_j = \Sigma^{-1/2}(\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}\Sigma^{-1/2}, \quad j = 1, \dots, K, \tag{3}$$

where Σ is a Fréchet mean of $\{\Sigma_j\}_{j=1}^K$ with respect to the procrustes metric Π, i.e., a minimiser of the sum-of-squares functional $\Gamma \mapsto \sum_{j=1}^K \Pi^2(\Gamma, \Sigma_j)$ over the space of trace-class covariances. Moreover, the $\mathbf{t}_j$ are centred around the identity in that

$$\frac{1}{K}\sum_{j=1}^K \mathbf{t}_j = \mathscr{I}. \tag{4}$$

The Fréchet mean Σ is unique when at least one of the Σ_j is injective (or more generally, if the kernel of at least one Σ_j is contained in the kernels of all other Σ_j); see Masarotto et al. [27, Proposition 10]. Its algorithmic construction is discussed in detail in Section 3.

Once the multicoupling (3) has been constructed, and a norm $\|\cdot\|_r$ ($r = 1, 2, \infty$) has been chosen, the null hypothesis can be tested by measuring a combined deviation of the Δ_j from zero using that norm, and calibrating the typical values of such deviations under the null. This is discussed in the next subsection. As discussed in Masarotto et al. [27], the optimal multicouping has an elegant geometrical interpretation in terms of the manifold geometry of the Procrustes distance — this will later be exploited in Section 2.4 in order to construct a functional PCA of the covariance operators.

2.3. Transportation-based functional ANOVA of covariances

Assume now that we have K independent groups of functional data $\{X_{ij},\ j = 1, \dots, n,\ i = 1, \dots, K\}$, each having covariance Σ_i (the procedure can be easily adapted to different group sizes). Without loss of generality, the data are assumed to be zero mean. Based on the discussion in the previous paragraph, we can test the equality of covariance operators by means of testing the hypothesis

$$H_0' : \underbrace{\mathbf{t}_1 - \mathscr{I}}_{=\Delta_1} = \cdots = \underbrace{\mathbf{t}_K - \mathscr{I}}_{=\Delta_K} = 0$$

where

$$\mathbf{t}_j = \Sigma^{-1/2}(\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}\Sigma^{-1/2}, \quad j = 1, \ldots, K,$$

and Σ is a Fréchet mean of $\{\Sigma_j\}_{j=1}^K$ with respect to the procrustes metric Π. At the level of our sample, we have access to empirical versions of $\{\hat{\Sigma}_j\}$, constructed on the basis of the samples of size n from each group. These could simply be the empirical covariances within each group (under a complete observation assumption), or some smoothed estimator (for instance the empirical covariance of smoothed versions of the $\{X_{ij}\}$ (Ramsay and Silverman [37]), or PACE-type estimators (Yao et al. [43])). Whichever the case may be, the $\hat{\Sigma}_j$ are finite dimensional, of rank $q \leq n$. In case a smoothing technique is used, we assume that it is such that the $\hat{\Sigma}_j$ share a common range, and can thus be represented as $q \times q$ positive matrices, via a common (tensor product) basis. For tidiness, we use the same notation for $\hat{\Sigma}_j$ and its $q \times q$ matrix representation in the common basis. It is clear that this basis can be chosen so that at least one of these matrices is of full rank q.

In this case, there exists a unique empirical Fréchet mean $\hat{\Sigma}$,

$$\hat{\Sigma} = \underset{\mathbb{R}^{q\times q}\ni\Gamma\succeq 0}{\arg\min} \sum_{j=1}^{K} \Pi^2(\hat{\Sigma}_j, \Gamma)$$

and this can be computed from $\{\hat{\Sigma}_1, \ldots, \hat{\Sigma}_K\}$ using steepest descent (see Section 3). This gives rise to empirical versions of the $\mathbf{t}_j$,

$$\hat{\mathbf{t}}_j = \hat{\Sigma}^{-1/2}(\hat{\Sigma}^{1/2}\hat{\Sigma}_j\hat{\Sigma}^{1/2})^{1/2}\hat{\Sigma}^{-1/2}, \quad j = 1, \ldots, K,$$

and corresponding empirical deviations from the identity

$$\hat{\Delta}_j = \hat{\mathbf{t}}_j - I_{q\times q}.$$

The testing procedure is now based on the test statistic

$$T_r = \sum_{j=1}^{K} \left|\!\left|\!\left| \Delta_j \right|\!\right|\!\right|_r^2,$$

where $r \in \{1, 2, \infty\}$ (the performance under the different choices of r is investigated in the simulation section). We avoid making any concrete parametric assumptions, and instead calibrate the test statistic by means of permutations. The typical permuted value will be calculated according to the following steps:

- Reassign the $n \times K$ curves $\{X_{i,j}\}$ into K groups of equal size. Call these new groups $\{X^*_{i,j}\}$.
- Construct the empirical covariance $\hat{\Sigma}^*_j$ for the jth group $\{X^*_{i,j}\}_{i=1}^n$, $j = 1, \ldots, K$.
- Compute empirical Fréchet mean $\hat{\Sigma}^*$ of $\{\hat{\Sigma}^*_1, \ldots, \hat{\Sigma}^*_K\}$.

- Construct $\hat{\mathbf{t}}_j^* = (\hat{\Sigma}^*)^{-1/2}\sqrt{(\hat{\Sigma}^*)^{1/2}\hat{\Sigma}_j^*(\hat{\Sigma}^*)^{1/2}}(\hat{\Sigma}^*)^{-1/2}$ and compute

$$T_r^* = \sum_{j=1}^{K} \left\| \hat{\mathbf{t}}_j^* - I_{q\times q} \right\|_r^2 = \sum_{j=1}^{K} \left\| \hat{\Delta}_j^* \right\|_r^2.$$

Repeating these steps for all possible re-assignments yields the distribution for the permuted statistics T_r^*, which can be used to generate a p-value for T_r under the null hypothesis. As usual, an exact such p-value can become prohibitive for large K, and in practice we resort to Monte Carlo sampling of permutations. Note that similar steps allow for the implementation of a bootstrap-type procedure, simply by randomly permuting indices with replacement. However, we opt for the permutation approach since the exchangeability of the permutation labels under H_0 guarantees the (near) exactness of the K-sample permutation test (Pesarin and Salmaso [32]), and we do not pursue the bootstrap approach further.

Remark 2.3. [The role of Gaussianity] It may be worth highlighting that the test procedure does not assume the $\{X_{ij}\}$ to be Gaussian process. Gaussianity (or lack of it) does not play a role in the calibration of the test, which is done via permutation. It only serves to motivate the transport-based measure of dissimilarity that comprises our test statistic, by way of the 1-to-1 correspondence between covariances and centred Gaussian measures. But once this measure of dissimilarity has been defined, it can be interpreted in its own right, without any reference to Gaussian measures.

2.4. Transportation-based functional tangent PCA of covariances

When the null hypothesis of equality between covariances is rejected, the analyst may wish to explore whether the detected differences are carried by some interpretable main modes of variation. The transport maps $\mathbf{t}_j$ (or their empirical versions) can be used to this aim. When these exist, the differences $\Delta_j = \mathbf{t}_j - \mathscr{I}$ admit an elegant geometric interpretation as the logarithms of the operators Σ_j at the Fréchet mean Σ, under the manifold-like geometry induced by the Procrustes metric $\Pi(\cdot,\cdot)$ on the space of (trace-class) covariance operators. Specifically, Masarotto et al. [27] show that it admits a tangent space with respect to geometry induced by Π that is characterised as

$$\mathrm{Tan}_\Sigma = \overline{\left\{Q : Q = Q^*,\ \left\|\Sigma^{1/2}Q\right\|_2 < \infty\right\}}$$

where the closure is with respect to the inner product

$$\langle Q_1, Q_2\rangle_\Sigma = \mathrm{trace}(Q_1\Sigma Q_2).$$

When Σ is injective, this is a bona fide inner product, that is, $\langle Q, Q\rangle = 0 \iff Q = 0$. However, the Fréchet mean need not be injective even if all Σ_j are so, and it is not clear that the Σ_j's can be lifted to the tangent space. Nevertheless,

our new result in the form of Theorem 2.2 guarantees that the maps Δ_j do exist as bounded self-adjoint operators, and indeed the 1-form

$$\operatorname{trace}(\Delta_i \Sigma \Delta_j) \leq \left\| \Sigma^{1/2} \Delta_i \right\|_2 \left\| \Sigma^{1/2} \Delta_j \right\|_2 = \Pi(\Sigma_i, \Sigma) \Pi(\Sigma_j, \Sigma) < \infty$$

is well-defined, regardless of the injectivity of Σ by means of the formulae for $\mathbf{t}_i$ and Π. Consequently, the finite-dimensional span of $\{\Delta_1, \ldots, \Delta_K\}$ admits a Hilbertian structure when equipped with the inner product $\langle \cdot, \cdot \rangle_\Sigma$, and for all practical purposes can be used to carry out a PCA[3] based on the spectral decomposition of the non-negative operator

$$\mathscr{K} = \frac{1}{K} \sum_{j=1}^{K} \Delta_j \otimes_\Sigma \Delta_j = \frac{1}{K} \sum_{j=1}^{K} \left(\mathbf{t}_j - \mathscr{I} \right) \otimes_\Sigma \left(\mathbf{t}_j - \mathscr{I} \right),$$

where $(A \otimes_\Sigma B)C = \langle B, C \rangle_\Sigma A$. Notice that the latter constitutes precisely the empirical covariance of the collection $\{\Delta_j\}_{j=1}^K$, because

$$\sum_{j=1}^{K} \Delta_j = 0,$$

by Equation (4). Once the principal components are constructed, the main modes of covariance variation can be visualised by retracting appropriate subspaces of the tangent space back to the space of covariance operators (see also Fletcher et al. [15] for the study of principal geodesics analysis in a general Riemannian symmetric space). Specifically, if E_1 is the eigenoperator associated with the largest eigenvalue of $\mathscr{K}$, this retraction takes the form

$$t \mapsto (\mathscr{I} + tE_1)\Sigma(\mathscr{I} + tE_1), \qquad t \in [-\epsilon, \epsilon],$$

which is a geodesic for sufficiently small $\epsilon > 0$. This principal geodesic is the visualisation of the main mode of variation of $\{\Sigma_j\}_{j=1}^K$ near their Fréchet mean Σ.

A subtlety here is that the PCA is to be carried out on a Hilbert space endowed with an inner product other than the standard Hilbert–Schmidt inner product. This different choice of inner product affects both the formal definition and the computational evaluation of the principal components. The defining maximisation problem yielding the first principal component is now

$$\begin{aligned} \underset{\|B\|_\Sigma = 1}{\arg\max} \langle \mathscr{K} B, B \rangle_\Sigma &= \underset{\left\| \Sigma^{1/2} A \right\|_2 = 1}{\arg\max} \langle \mathscr{K} \Sigma^{1/2} A, \Sigma^{1/2} A \rangle_2 \\ &= \underset{\operatorname{trace}(A \Sigma A) = 1}{\arg\max} \operatorname{trace}(\mathscr{K} \Sigma^{1/2} A^2 \Sigma^{1/2}). \end{aligned}$$

[3]Such a PCA can be interpreted as a tangent space PCA with respect to a *Procrustean metric tensor*

$$\langle Q_1, Q_2 \rangle_\Gamma = \operatorname{trace}(Q_1 \Gamma Q_2), \quad \Gamma \in \mathscr{L} = \left\{ \underset{\Gamma \succeq 0}{\arg\min} \sum_{j=1}^{K} \alpha_j \Pi^2(\hat{\Sigma}_j, \Gamma) : \alpha_j > 0 \,\&\, \sum_{j=1}^{K} \alpha_j = 1 \right\}$$

over the barycentric locus $\mathscr{L}$ of the operators $\{\Sigma_j\}_{j=1}^K$.

Nevertheless, this change of inner product poses no essential difficulty, and has indeed considered before by Silverman [41] in the case of Sobolev inner products, and generalised by Ocaña et al. [28]. Further details are given in Section 3.

3. Computational implementation

In the next Section we will work with K independent groups of functional data $\{X_{ij},\ j = 1,\dots,K,\ i = 1,\dots,n_j\}$, each group of sample size n_j and with covariance operator Σ_j, $j = 1,\dots,K$. Unless otherwise stated, all curves are curves are simulated from a multivariate Gaussian process and sampled on an equispaced grid on $\Omega = [0,1]$. The sample size and the grid points vary across applications, therefore, in practice, we only have access to estimated empirical covariances $\widehat{\Sigma}_1,\dots,\widehat{\Sigma}_K$. In our case, $\widehat{\Sigma}_1,\dots,\widehat{\Sigma}_K$ are obtained from the smoothed versions of the $\{X_{ij}\}$ as traditional sample covariance functions, through the command `var.fd` in the `R` package `fda` (Ramsay and Silverman [36], Ramsay et al. [38]). If a smoothed version of the $\{X_{ij}\}$'s is not available, a PACE-type estimator can be used (Yao et al. [43]). All the functions needed to apply transport ANOVA and transport tangent PCA have been made available in the R package "fdWasserstein" (Masarotto and Masarotto [26]).

3.1. Transport ANOVA

Once estimators $\widehat{\Sigma}_1,\dots,\widehat{\Sigma}_K$ are at our disposal, our transport ANOVA requires their Fréchet mean

$$\overline{\Sigma} = \arg\min_{\Sigma} \sum_{j=1}^{K} \Pi^2(\Sigma, \widehat{\Sigma}_j)$$

and the transport maps contrasted with the identity

$$\Delta_j = \mathbf{t}_j - \mathscr{I} = \overline{\Sigma}^{-1/2}(\overline{\Sigma}^{1/2}\widehat{\Sigma}_j\overline{\Sigma}^{1/2})^{1/2}\overline{\Sigma}^{-1/2} - \mathscr{I}, \quad j = 1,\dots,K.$$

When $\widehat{\Sigma}_j$ commute ($\widehat{\Sigma}_j\widehat{\Sigma_i} = \widehat{\Sigma_i}\widehat{\Sigma}_j$ for all i,j), the $\overline{\Sigma}$ has the explicit form

$$\overline{\Sigma}^{1/2} = K^{-1}\left[\widehat{\Sigma}_1^{1/2} + \cdots + \widehat{\Sigma}_K^{1/2}\right].$$

However, there is no reason that these commutativity should hold. For general covariances, $\overline{\Sigma}$ has no closed form formula, but it can be approximated by the iterative procedure described in [27, Section 8]. It can be interpreted as steepest descent in the Procrustes space of covariances, and in finite dimensions provably approximates $\overline{\Sigma}$ and Δ_j to arbitrary precision. It is carried out as follows:

- Let $\Sigma^0 : \mathcal{H} \to \mathcal{H}$ be an injective covariance, serving as the initial point.
- Denote the current iterate at step k as Σ^k.
- For each j compute the optimal maps from Σ^k to each of the prescribed operators $\widehat{\Sigma}_j : \mathcal{H} \to \mathcal{H}$, namely

$$\mathbf{t}_{\Sigma^k}^{\widehat{\Sigma}_j} = (\Sigma^k)^{-1/2}[(\Sigma^k)^{1/2}\widehat{\Sigma}_j(\Sigma^k)^{1/2}]^{1/2}(\Sigma^k)^{-1/2}.$$

- Define their average $T_k = K^{-1}\sum_{j=1}^{K} \mathbf{t}_{\Sigma^k}^{\widehat{\Sigma}_j}$, which is itself a non-negative operator on $\mathcal{H}$.
- Set the next iterate to $\Sigma^{k+1} = T_k \Sigma^k T_k$.

In practice the algorithm will stop after, say, k iterations, Σ^k will be our numerical approximation for $\overline{\Sigma}$ and $\mathbf{t}_{\Sigma^k}^{\widehat{\Sigma}_j} - \mathscr{I}$ will approximate $\mathbf{t}_j - \mathscr{I}$.

In terms of the manifold-like geometry of covariances under the Procrustes metric (see Section 2.4), the algorithm starts with an initial guess of the Fréchet mean; it then lifts all observations to the tangent space at that initial guess via the log map, and averages linearly on the tangent space; this linear average is then retracted onto the manifold via the exponential map, providing the next guess, and iterates. The quantity $T_k - \mathscr{I}$ is precisely the *negative gradient* of the Fréchet (sum-of-squares) functional, which is the reason why this is steepest descent.

The test statistic, a linear combination (or the maximum) of powers of $\left\|\left|\Delta_j\right|\right\|_2^2$, is readily computable from the spectral decomposition.

3.2. Transport tangent PCA

Once the empirical Fréchet mean $\overline{\Sigma}$ of $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K$ and the $\widehat{\Delta}_j$'s are computed (see the previous section), we can perform functional PCA on the collection $\{\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K\}$ by their tangent space representation $\widehat{\Delta}_1, \ldots, \widehat{\Delta}_K$ (see Section 2.4). As explained there, there are some subtleties involved in this PCA, since we are working with a different inner product than the standard Hilbert–Schmidt one. However, Ocaña et al. [28] proved that PCA with respect to the tangent space inner product is equivalent to the PCA performed with the Hilbert–Schmidt inner product on suitably transformed data, thus allowing a framework to interpret standard Euclidean PCA in the Procrustes geometry. More precisely, let $\langle \cdot, \cdot \rangle_{HS}$ be the Hilbert–Schmidt inner product and $\langle \cdot, \cdot \rangle_{\Sigma}$ be the Wasserstein one at the tangent space at $\bar{\Sigma}$, that is $\langle A, B \rangle_{\Sigma} = \mathrm{tr}(A\Sigma B)$. We follow the steps by Ocaña et al. [28], using that there is a unique operator $\mathscr{T}$ characterised by

$$\langle A, B \rangle_{\Sigma} = \langle \mathscr{T}(A), B \rangle_{HS} = \mathrm{tr}([\mathscr{T}(A)]^* B).$$

which in our case we take to be the multiplication from the right by Σ (so $\mathscr{T}(A) = (A\Sigma^{1/2})\Sigma^{1/2}$ is trace class and has an adjoint). is computed as the PCA of $[\mathscr{T}^{1/2}(\mathscr{X}_i)]_{i=1}^n$ with Hilbert–Schmidt norm, in the sense that the eigenvalues (i.e. the variances) remain the same and the eigenfunctions with respect to $\langle \cdot, \cdot \rangle_{\Sigma}$ are $\mathscr{T}^{1/2}$ applied to the eigenfunctions with respect to $\langle \cdot, \cdot \rangle_{HS}$. In our specific case, $\mathscr{T}^{1/2}(\mathscr{X}) = \mathscr{X}\Sigma^{1/2}$, and the PCA on the tangent space is carried out as follows:

- Multiply $\Delta_j = \mathbf{t}_j - \mathscr{I}$ from the right by $\Sigma^{1/2}$.
- Find the spectral decomposition of the empirical operator $\tilde{\mathscr{K}} = K^{-1}\sum \Delta_j \Sigma^{1/2} \otimes \Delta_j \Sigma^{1/2}$, defined on the space of Hilbert–Schmidt operators with respect to the Hilbert–Schmidt norm.

- Multiply (from the right) the eigenfunctions of $\tilde{\mathscr{K}}$ by $\Sigma^{-1/2}$ to obtain the eigenfunctions of $\mathscr{K}$.

4. Numerical experiments

We now demonstrate the efficacy of the proposed methods through a variety of simulated examples. Simulations are broadly categorized into two subsets: functional ANOVA and tangent space PCA. We initially explore the behavior in a scenario where the Fréchet mean is known and the transport-based functional ANOVA test is applied to covariances $\{\Sigma_j = T_j \Sigma T_j\}_{j=1}^K$ obtained via perturbation according to the generative model described in Masarotto et al. [27, Section 10]. To allow comparability of our method with the existing literature, we subsequently consider another simulation scenario, directly taken from Cabassi et al. [6], where the simulated covariance operators are perturbation of the male and female subjects in the Berkeley growth data set (Jones and Bayley [22]). In both scenarios, generative model and Berkeley data, we compare the functional ANOVA based on Transport Maps with the permutation test of Cabassi et al. [6], the concentration inequality method by Kashlak et al. [23] and the functional ANOVA method based on empirical characteristic functionals of Hlávka et al. [21]. Cabassi et al. [6] provide result showing that their method is state-of-the-art, and to the best of our knowledge, no other alternative procedures besides the ones listed exist. It is evident from Figure 1 and 2 that our testing method over-powers other methods, reaching nearly perfect power even in the presence of small differences. After validating the performance of transport-based functional ANOVA, in Section 4.2 we make use of the generative model scenario to perform tangent space Principal Component. Finally, in 5 we test the performance of our method on the classic phoneme dataset (Ferraty and Vieu [14]), which consists of 4509 log-periodograms of 5 different phonemes. The data is available at `https://hastie.su.domains/ElemStatLearn/`. Real data analysis consolidates the strength of our method with respect to the competitors, as it can be seen in Section 5.1. If the null hyphothesis of equality among covariance operators is rejected, Section 5.2 shows how tangent space PCA can be a successful tool in understanding dataset variability.

4.1. Simulation experiments

In order to avoid propagation of error, it is convenient to formulate a simulation setup in which the Fréchet mean is known exactly and does not need to be approximated (Section 3). It is easy to construct such examples in the commutative case, but we shall not do so, as this case is overly restrictive and unrealistic. We thus appeal to the generative model in Masarotto et al. [27, Section 10], which states that if a collection of nonnegative maps $T_1, \ldots, T_K$ has mean identity, then any covariance operator Σ is the Fréchet mean of $\{\Sigma_j = T_j \Sigma T_j\}_{j=1}^K$, and the maps $\mathbf{t}_j$ in (3) must equal T_j (on the closed range of Σ).

To construct the collection $\{T_1, \ldots, T_K\}$ we proceed as follows. Let $\mathcal{H} = L^2[0,1]$ and for $f, g \in \mathcal{H}$ their tensor product $f \otimes g$ is the operator

$$(f \otimes g)(h) = \langle f, h \rangle g = \left(\int_0^1 f(t)h(t)dt \right) g \in \mathcal{H}, \qquad h \in \mathcal{H}.$$

If t is the parameter of the functions, we write $f(t) \otimes g(t)$ to mean $f \otimes g$. With this notation set

$$T_j = k^{-1} \sum_{n=1}^{\infty} \delta_n^{(j)} \sin(2n\pi t - \theta^{(j)}) \otimes \sin(2n\pi t - \theta^{(j)}), \quad j \in \{1, \ldots, K\}, \quad \delta_n^{(j)} \overset{iid}{\sim} \chi_k^2, \tag{5}$$

where $\delta_n^{(j)}$ are independent of $\theta^{(j)}$, and $k > 0$. This construction guarantees that $\mathbb{E}[T_j] = \mathbb{E}[\mathbb{E}[T_j|\theta^{(j)}] = \mathscr{I}$ regardless of the distribution of $\theta^{(j)}$. The parameter k controls the concentration of $T^{(j)}$ around the identity; when k is large, the law of large numbers entails that $\delta_n^{(j)}$ is close to 1. Of course, a given realisation of $T_1, \ldots, T_K$ will not average precisely to the identity, but will average approximately to the identity if K and k are not too small. The parameter $\theta_i \geq 0$ on the other hand, serves as indicators on how far we are from commutativity. On this note, a parametric model can be assumed for the θ_i. We chose the θ_i to be sampled from a von Mises distribution with mean 0 and measure of concentration $1/\sigma$, with the degenerate case of $\sigma \to \infty$ yielding commutativity.

We then generate n_j Gaussian curves $X_{i,j}$, $i \in \{1, \ldots, n_i\}$ with mean zero and covariance $\Sigma_j = T_j \bar{\Sigma} T_j^{\star}$, $j \in \{1, \ldots, K\}$, $j \in \{1, \ldots, K\}$. Inspired from Kashlak et al. [23], the "population" Fréchet mean was chosen to be a matrix with eigenvalue decay rate $O(n^{-4})$:

$$\bar{\Sigma} = U \left[\sum_{n=1}^{\infty} n^{-4} \sin(2n\pi t) \otimes \sin(2n\pi t) \right] U^* \tag{6}$$

where U is a randomly generated orthogonal operator. To simulate a functional case and have enough information to display the decay of the spectrum, we chose for the matrices a relative high size of 50×50. The power is estimated from 500 replications. The number of permutations is 200. At each replication, we generate two optimal maps T_1 and T_2 via the generative model, and two corresponding covariances $\Sigma_1 = T_1 \Sigma T_1^*$ and $\Sigma_2 = T_2 \Sigma T_2^*$, with Σ given by equation (6). For each Σ_i, $i = 1, 2$, we sample 40 observations of a Gaussian process with mean-zero and covariance Σ_i. The empirical covariance computed from these observations will yield a replica of Σ_i. We repeat this as to obtain k_1 replicas of Σ_1, and k_2 replicas of Σ_2, for a total of $k_1 + k_2$ covariances divided into two groups of size k_1 and k_2 respectively. The values of the pair (k_1, k_2) are (1,2), (1,3), (1,7) and (4,4). This procedure is repeated for several values of the Von Mises parameter, namely $\sigma^{-1} = (0.1, 1, 5, 10)$.

We compare the power of our procedure with that of the K-sample permutation test in Cabassi et al. [6], Kashlak et al. [23], Hlávka et al. [21]. The idea in Cabassi et al. [6] is to perform a series of partial 2-sample tests for each pair

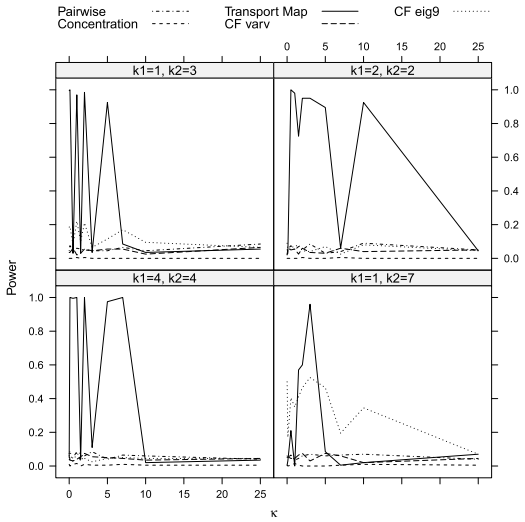

FIG 1. *Empirical power of different tests as a function of the dispersion parameter k in* (5).

of groups, and combine the pairwise test statistics through the non-parametric combination algorithm of Pesarin and Salmaso [32]. The pairwise test statistics are $T_{ij} = d(\Sigma_i, \Sigma_j)$, where Σ_i and Σ_j are the sample covariance operators of the corresponding groups, and d is a metric on the operators space. The method of Cabassi et al. [6] is general, as any distance d can be used as test statistic. It is shown to be more powerful than competing method such as Kashlak et al. [23], and is implemented in the R-package `fdcov` (Cabassi and Kashlak [5]). In our comparisons, we have used the square root distance $\left\|\Sigma_i^{1/2} - \Sigma_j^{1/2}\right\|_2$, which led to the best performance according to Cabassi et al. [6]. Using the Procrustes distance instead of the square root distance had similar performance but increased computational cost. Hlávka et al. [21] propose a a test statistics of Cramér-von Mises type with the distance of the empirical characteristics functionals (ECFs) $\int |\phi_1(\omega) - \phi_2(\omega)|^2 \mathrm{d}Q(\omega)$, where ϕ_i is the ECF of the sample and Q is some probability measure on the dual space of $\mathcal{H}$. The performance of the proposed test by Hlávka et al. [21] relies heavily on the choice of the measure Q. We consider a Gaussian measure characterized by the choice of two different covariance operators: the sample covariance matrix on one hand and an approximation of the inverse of the pooled sample covariance matrix computed from the first 9 eigenvectors on the other. These two cases appear to be performing the best in Hlávka et al. [21]. We include both choices in our simulation study because the first yields overall better performance, but the latter gains power in difficult cases, like when most operators are equals. After performing the global test, in case the null hyphothesis H_0 is rejected, one can investigate pairwise differences with post-analysis comparison, as in Cabassi et al. [6], Pesarin and Salmaso [32].

Figure 1 shows the empirical power of all procedures. The x-axis represents the value of the dispersion parameter k in (5), while the y-axes displays the empirical power. It is evident from the figure that our procedure over-powers

that of Cabassi et al. [6], Kashlak et al. [23] as well as of Hlávka et al. [21] for both variations considered. To better understand Figure 1, it is important to note that the transport perturbation given by the generative model would not result in a monotonic effect on the power curve. This is due to the intrinsic nature of the generative model as we are not sampling curves directly, but we are constructing transport operators which have a periodic pattern. The effect of different variances σ of the Von Mises distribution will be affected by such periodicity. The only predictable effect we could foresee is that for very large k (i.e. very small variance) the Von Mises random deviates are very concentrated around 0. This would yield commutativity and, consequently, a very low power. Figure 1 displays both the periodical nature of the model, as well as the decrease in power when we approach commutativity.

The success of our test is a genuinely *functional* phenomenon and indeed, when the data are truncated, the differences are not so overwhelming. To understand this better notice that, since χ^2_k is an unbounded random variable we have $\|\|T_j\|\|_\infty = \infty$ almost surely, and our test based on $\|\|T_j - \mathscr{I}\|\|_2 = \infty$ consequently rejects the null hypothesis. If the series is truncated at a finite level n_0, then $\|\|T_j\|\|_1 \sim k^{-1}\chi^2_{k,n_0}$ and so $P(\|\|T_j\|\|_1 > R)$ decays to zero exponentially as R and/or k increase; a fortriori the same holds for $\|\|T_j\|\|_2$ and $\|\|T_j\|\|_\infty$. The procedure of Cabassi et al. [6] effectively puts very small weights on what happens at the tails (n large), whereas our procedure is able to detect departures from the null even when they only occur at very high frequencies. We refer to Section 6 for more insight on the power of the functional ANOVA test.

One may argue that the generative model setup in (5) artificially favours our testing procedure, as it guarantees $\|\|\Delta_j\|\| = \infty$. We therefore consider another simulation scenario, directly taken from Cabassi et al. [6], in order to compare the two methods on the same ground. Here $n = 20$ curves generated from a mean-zero Gaussian process with suitably chosen covariances are evaluated on a equipaced grid of 31 points on $[0, 1]$. Such curves are assumed to come from K populations, and the covariance operators of each population is obtained via perturbations of some given, known covariances Σ_f and Σ_m. These are computed with the `fda` R-package as the covariance operators of the smoothed growth curves of male and female subjects in the Berkeley growth data set (Jones and Bayley [22]). [22] The perturbations take two different forms:

1. *geodesic perturbations:* $K_1 < K$ of the groups have covariance operator

$$\Sigma(\gamma) = [\Sigma_m^{1/2} + \gamma(\Sigma_f^{1/2} R - \Sigma_m^{1/2})][\Sigma_m^{1/2} + \gamma(\Sigma_f^{1/2} R - \Sigma_m^{1/2})]^*$$

   with R the operator minimising the procrustes distance and $\gamma \in [0, 5]$. The other $K_2 = K - K_1$ groups have covariance operator Σ_m.
2. *additive perturbations:* $K_1 < K$ of the groups have covariance operator $\Sigma(\gamma) = (1 + \gamma)\Sigma_m$, $\gamma \in [0, 5]$. The other $K_2 = K - K_1$ groups have covariance operator Σ_m.

The number of permutation is again 200. The power is estimated from a total of 500 replications. The test-statistics employ the Hilbert-Schmidt norm ($r = 2$).

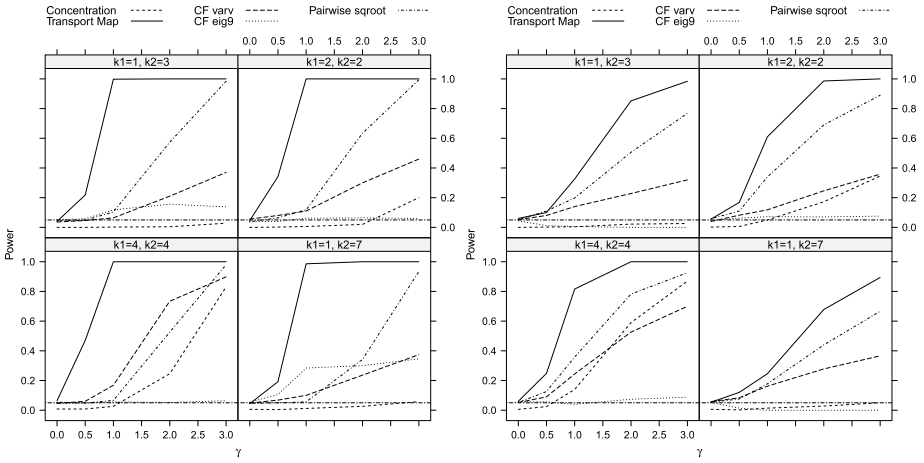

FIG 2. *Empirical power of ours and Cabassi et al. [6], Kashlak et al. [23], Hlávka et al. [21]'s method in the Gaussian case, as a function of the perturbation parameter γ. Left: geodesic perturbations; right: additive perturbations.*

The probabilities of false positive (I type error) are estimated using all the available replications when $\gamma = 0$. Figure 2 compares the power of our transport test with that of Cabassi et al. [6], Kashlak et al. [23], Hlávka et al. [21] on these synthetic data. The x-axis gives the value of the γ parameter, while the y-axes displays the empirical power. It is seen that our method is more powerful in all scenarios considered. Moreover, in case of geodesic perturbations, we achieve near perfect power, as opposed to the other tests that have little to nearly no power, for small values of γ, i.e. against local alternatives. Furthermore, notice that without knowing the null distribution is not possible to use the calibration procedure of Kashlak et al. [23], which is too conservative and does not respect the nominal level of 0.05 under H_0. In terms of computational cost, we have compared the runtime between our procedure, and those of [5] and [21]. Computation of a p-value this setting is around 5 seconds for our method and that of [5], and around 2 seconds for [21]. Arguably the difference is immaterial in practice.

To illustrate that the test does not rely in any way on the assumption of Gaussianity, as explained in Remark 2.3, we run the test in exactly the same settings as 2 but on t-student processes. As it is visible from the plots, we retain a higher power even in the non-Gaussian case.

4.2. Tangent space PCA

We validate the PCA framework described in Section 3.2 on a synthetic datasets which is inspired by the theoretical generative model (5) and which yields N covariances well separated in K groups. The aim is to see whether PCA is able to differentiate between the groups. The operators $\Sigma_1, \ldots, \Sigma_K$ are obtained

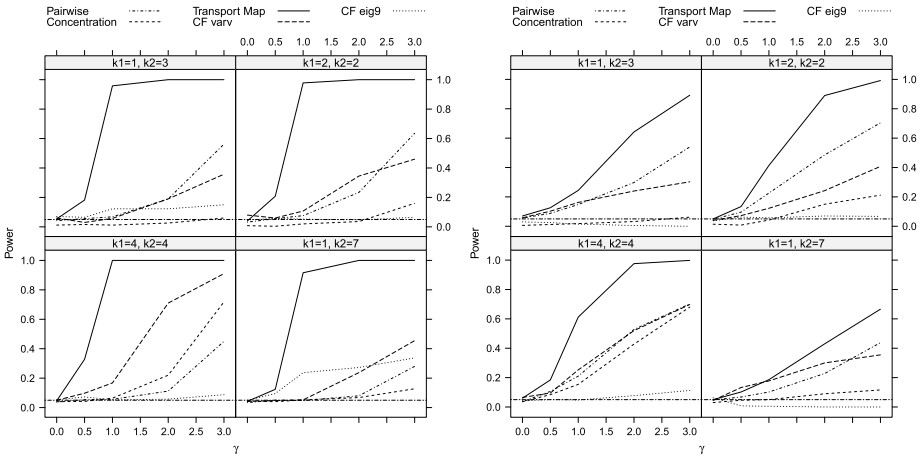

FIG 3. *Empirical power of ours and Cabassi et al. [6], Kashlak et al. [23], Hlávka et al. [21]'s method in the Gaussian case, as a function of the perturbation parameter γ. Left: geodesic perturbations; right: additive perturbations.*

as a conjugation perturbation of some known Fréchet mean by the generated "optimal" maps $T_1, \ldots, T_K$

$$T_i = \sum_n \delta_n^{(i)} \sin(2n\pi t - \theta^{(i)}) \sin(2n\pi t - \theta^{(i)})$$

where the $\delta_n^{(j)}$ are drawn from a χ^2 distribution and $\theta^{(i)}$ are sampled from a von Mises distribution of mean 0 and measure of concentration $1/\sigma$. The Fréchet mean is chosen to be $\bar{\Sigma} = U\Lambda U^*$ as in Kashlak et al. [23], with U being a randomly generated unitary operator, and Λ a $d \times d$ diagonal matrix with eigenvalue decay of $O(d^{-4})$, d being the dimension of the matrices.

As the generative model yields optimal maps which are small perturbations of the identity, the dimension of the matrices used to approximate the operators needs to be large, otherwise the estimation errors would overwhelm the intrinsic variability of the sample. The dimension is chosen to be 200, the measure of concentration to be 1 and the number of groups K to be $K = 3$. For each of the Σ_j, $j = 1, 2, 3$, we generate 100 samples of 50 Gaussian curves each. We then estimate the empirical covariance of these curves, obtaining a sample of $N = 300$ covariances. Results of the PCA are shown in Table 1 and Figures 4 and 5. The Figures show that the different groups are clearly identified.

TABLE 1
Importance of each PC, first experiment with the generative model.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	0.5810	0.1545	0.0966	0.0371	0.0276	0.0216
Proportion of Variance	0.8989	0.0635	0.0248	0.0037	0.0020	0.0012
Cumulative Proportion	0.8989	0.9624	0.9873	0.9909	0.9930	0.9942

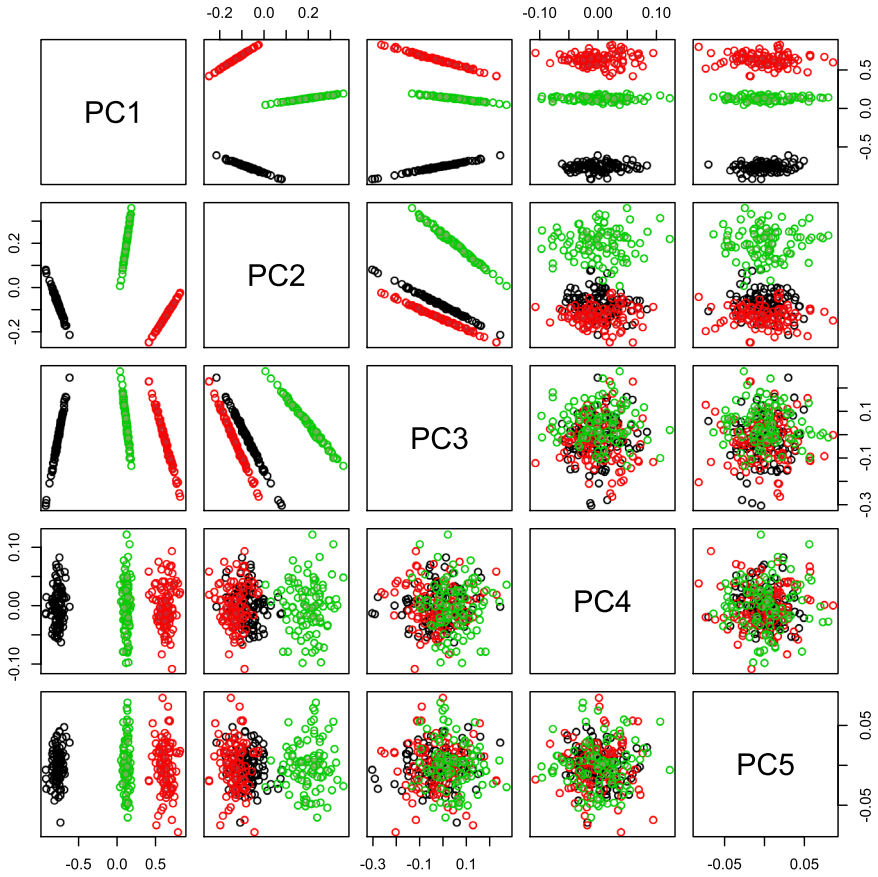

FIG 4. *PCA scores, first experiment with the generative model. Colours correspond to the three maps generated from the model.*

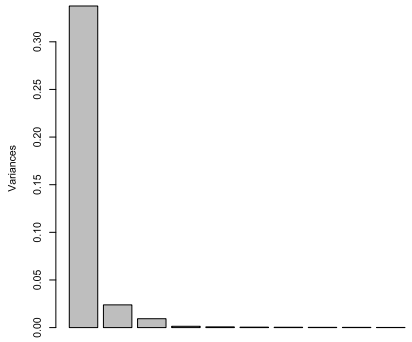

FIG 5. *Eigenvalues screeplot, first experiment with the generative model*

5. Data analysis: phoneme periodograms

In this section, we illustrate our method on the phoneme data set considered in Hastie et al. [20]. The dataset consists of 4509 log-periodograms of length 256 each, computed from continuous speech frames of 50 male speakers. Each speech frame is 32msec long, sampled at a rate of 16kHz and represents one of the following five phoneme: "aa" (as in "dark", nasal a), "ao" (as in "water"), "iy" (as in "she"), "sh" (as in "she"), "dcl" (as in "dark", "british" d). Each phoneme j gives rise to a covariance operator Σ_j. We use this sample of $K = 5$ covariances to generate K populations of n Gaussian processes, on which inference will be performed.

5.1. ANOVA

In order to perform ANOVA on the phoneme dataset, we extract the log-periodograms corresponding to the phonemes "aa", "ao", "iy". We limit the test to these three phonemes because of their similarity, which makes it harder to distinguish them and allows for better discrimination among different procedures. If we include all 5 phonemes, the difference between vowels and consonants sound is so stark that all tests have very high power.

To sample under H_0, we sample $3n$ log-periodograms for the "iy" phoneme which are then randomly assigned to three groups, each of size n. To sample under the alternative H_1, we sample n log-periodograms for each phoneme. We repeat the test for $n = 25$ and $n = 50$, and for 500 replications and 200 permutations. Again we compare both with Cabassi et al. [6], Kashlak et al. [23] and Hlávka et al. [21]. We limit the comparison with Hlávka et al. [21] to the test statistics using the sample covariance matrix, as it is the scenario that gives consistently better results. We extend their two-sample testing to 3-samples by considering all pairwise comparisons and using the maximum test statistics. When interpreting the results, it is important to treat the outputs carefully, since the procedure of Kashlak et al. [23] was unable to produce a result in a small number of cases, as the computation of the distance using SVD failed, while in some cases of Hlávka et al. [21], both the test statistics under the null and the p-values are identically 0.

Table 2 shows the comparison between the test on smoothed phonema log-periodograms. Since the different phonemes have different mean functions, observations must be centered around the sample mean of each group, implying that the right type I error probability under H_0 might not be respected. Regardless, the transport test delivers a level very close to the nominal 0.05, especially when $n = 50$ (which is still relatively low compared to the 256 points where the curves are sampled). The tests of Cabassi et al. [6] and Hlávka et al. [21] also reache an acceptable nominal level under H_0. This is not the case for Kashlak et al. [23] making impossible the comparison. However, Cabassi et al. [6] show that their test outperforms that of Kashlak et al. [23]. It is worth mentioning that the test of Hlávka et al. [21] responds very well to variation of the mean,

TABLE 2
Comparison of the empirical power of the three different testing methods on the `phoneme` *dataset, when applied on the phonemes "aa", "ao" and "iy".*

	n	Pairwise	Concentration	Transport Maps	ECF-varv
H_0	25	0.086	0.000	0.068	0.2
	50	0.050	0.000	0.046	0.5
H_1	25	0.271	0.300	0.470	0.01
	50	0.670	0.944	0.994	0.1

because it takes into account the full distribution thanks to the characteristic functionals. In the phoneme dataset, the mean log-periodogram captures most of the variability. Thus, if one was to consider variation with respect to both first and second moment simultaneously, the method of Hlávka et al. [21] would show a significant increase in power. However, for the scope of this work, we are interested specifically in second order variation. When centering with respect to the mean, the test based on the transport maps greatly outperforms the competing methods under the alternative hypothesis.

5.2. PCA

In this section, we illustrate the use of tangent space PCA of covariance operators by applying it to the phoneme dataset described in Hastie et al. [20]. The collection of curves corresponding to each phoneme gives rise to a sample covariance operator, for a total of five covariances. The five empirical covariances are lifted to the tangent space via the log map centered at their Fréchet mean $\bar{\Sigma}$. Successively they are scaled by $\bar{\Sigma}^{1/2}$ as explained in section 3.2. Standard PCA can now be run on these quantities. Figure 6 (left) shows the results of applying PCA on `phoneme` data. We see clearly that tangent space PCA captures very well the difference among the phonemes, as each sillable is isolated in at least one plot. The colours are as follows: "sh" black, "iy" red, "dcl "green", "aa" blue, "ao" cyan, more precisely:

1. The first PC captures (part of) the difference between "aa, ao and iy" (vowels) and "dcl and sh" (consonants)
2. The second PC captures (part of) the difference between "dcl" and "sh" (two consonants).
3. The third PC captures (part of) the difference between "aa and ao" and "iy" (separating the two similar sounding vowels from the third more different one).
4. The fourth PC captures (part of) the difference between "aa" and "ao" (separating the last two remaining, and very similar, sounds).
5. Since, the order in the y-axes is the order of magnitude of the eigenvalues the analysis suggests also the importance of the differences between the operators. As intuition dictates the difference between vowels and consonants is nearly four times more pronounced than the difference between the sounds "aa" and "ao".

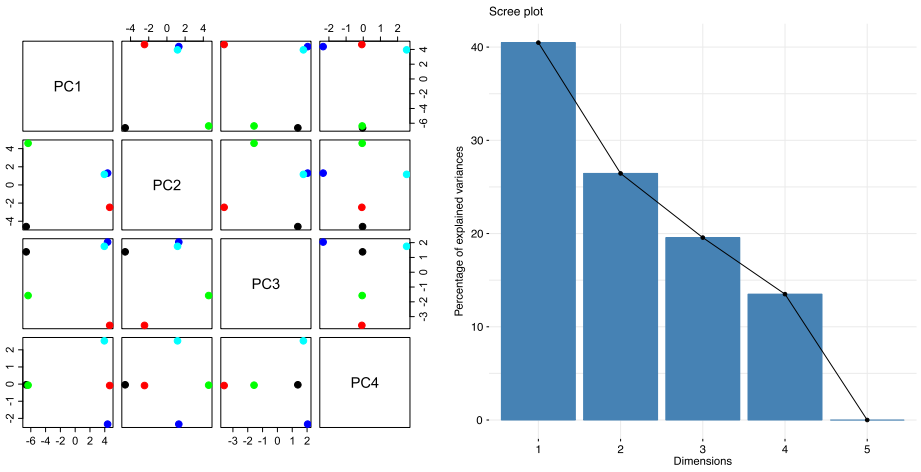

FIG 6. *Left: PCA scores, as computed from the phoneme dataset. The colours are as follows: "sh" black, "iy" red, "dcl "green", "aa" blue, "ao" cyan. Right: screeplot of eigenvalues, phoneme dataset*

The screeplot (Figure 6, right) shows that four PCs explain the full variance of the data, which is obvious as we have only five data points. The fourth PC is quite important and explains 13% of the variance.

In order to test our methodology in a more realistic situation, we artificially enlarge our sample of covariances by subsampling the original data. Specifically, from each of the five phonemes we subsample $B = 50$ of the corresponding log-periodograms to obtain a new estimator of the covariance operator of that phoneme. We do this $G = 12$ times so that in total we have a sample of $5G = 60$ covariances, divided into five groups, and the covariances in a group should be close to each other. We then carry out the PCA on these 60 covariances The results are showed in Figure 6 (right). Again, we can see that each phonema is isolated in at least one plot. Figure 7 shows the comparison of the PCA scores both in Euclidean and in Wasserstein distance. It is seen that the PCA based on the Procrustes tangent space distance is much more successful in distinguishing the covariances of different phonemes.

6. Concluding remarks

This paper introduces a framework that allows the comparison of several population of stochastic processes with respect to their covariance structure. We contributed a new methodology that exploits the theory of optimal (multi)transport and demonstrate how taking such a stand point allows to develop: (a) a testing procedure which outperforms the state of the art and (b) the first instance of tangent space principal component analysis of covariance operators. A fundamental ingredient of our approach and the main theoretical contribution of this paper is the proof that Gaussian measures can always be multicoupled

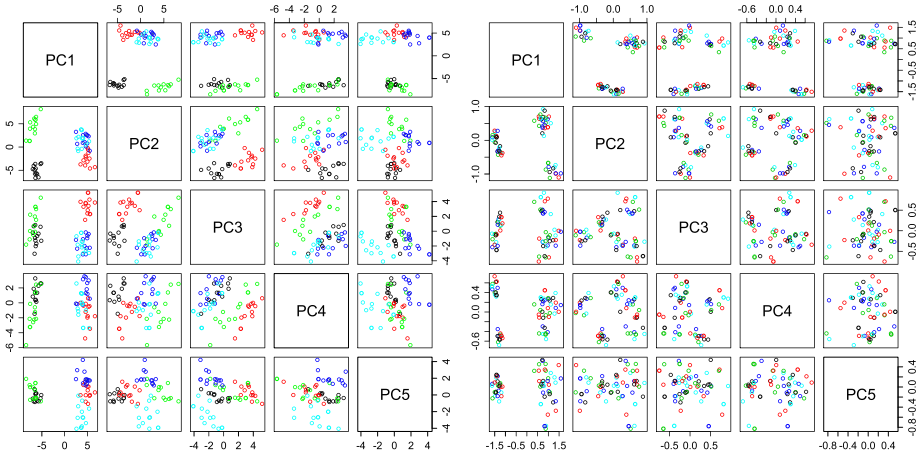

FIG 7. *PCA of the 60 covariance operators, based on the Procrustes tangent space distance (left) and the Hilbert–Schmidt distance (right)*

through bounded non-negative linear operators. The existence and boundedness result is elemental to both the testing procedure, and the PCA: the test statistic compares these coupling maps to the identity in norms of various strenghts, whereas the PCA uses the deviations of these couplings from the identity as a basis for eigenanalysis.

Specifically concerning the testing procedure, an appealing aspect of the presented methodology is that it harnesses a genuinely functional effect, in order to manifest exceptionally powerful performance under wide classes of alternatives – – alternatives for which previous tests would not perform nearly as well.

The genuinely functional effect arises under alternatives where the optimal coupling maps are *not* Hilbert–Schmidt (they are merely guaranteed to be bounded). At the population level, this corresponds to $\left|\!\left|\!\left| T_j - \mathscr{I} \right|\!\right|\!\right|_2 = \infty$, yielding very large values of the empirical test statistic.

Such a situation arises when departures from the null happen "across the whole spectrum" and not just in its bulk. These situations are not a theoretical curiosity —they can indeed be very common, as the following example will illustrate.

Consider the two-sample setting, and suppose that Σ_1 is the covariance operator of standard Brownian motion on $[0,1]$, whereas $\Sigma_2 = \sigma^2\Sigma_1$ is the covariance of standard Brownian motion on [0,1] scaled by the positive scalar $\sigma > 0$. These two covariance operators commute with Fréchet mean having covariance $(1+\sigma)^2\Sigma_1/4$, and corresponding transport maps $\mathbf{t}_1 = [2\sigma/(1+\sigma)]\mathscr{I}$ and $\mathbf{t}_2 = [2/(1+\sigma)]\mathscr{I}$. Thus Δ_j are both bounded, but they are not Hilbert–Schmidt unless $\sigma = 1$, leading to $\left|\!\left|\!\left| \Delta_2 \right|\!\right|\!\right|_k = \infty$, for any $k \geq 1$. The Wasserstein distance itself, however, equals $\left|\!\left|\!\left| \Sigma_1^{1/2} - \Sigma_2^{1/2} \right|\!\right|\!\right|_2 = (1-\sigma)\left|\!\left|\!\left| \Sigma_1^{1/2} \right|\!\right|\!\right|_2 = (1-\sigma)/\sqrt{12}$ and thus becomes arbitrarily small as σ nears one (recall that in the commutative case, the Wasserstein distance becomes the Hilbert–Schmidt distance of the corresponding positive roots).

The functional effect taking place is related to the Hajek–Feldman alternative, which has been exploited to obtain perfect discrimination of Gaussian process differing in their mean within an FDA context (see Delaigle and Hall [9]), by similarly exploiting the fact that a certain norm diverges under the alternative.

To see the connection with our setting, assume $\Sigma_j = \sum_n \lambda_{j,n}\varphi_n \otimes \varphi_n$ $(j = 1, 2)$ have the same eigenfunctions and thus commute. Then zero-mean Gaussian measures $N(0, \Sigma_1)$ and $N(0, \Sigma_2)$ are equivalent if and only if $\sum_n (r_n - 1)^2$ converges, where $r_n = \lambda_{2,n}/\lambda_{1,n}$. Indeed, summability implies that $r_n \to 1$ so that $\Sigma_1^{1/2}$ and $\Sigma_2^{1/2}$ have the same range and

$$\left\|\left|(\Sigma_1^{-1/2}\Sigma_2^{1/2})(\Sigma_1^{-1/2}\Sigma_2^{1/2})^* - \mathscr{I}\right|\right\|_2 = \sum_{n=1}^{\infty}(r_n - 1)^2$$

is finite, as required (see e.g., Da Prato and Zabczyk [8, Theorem 2.25]). The Fréchet mean and transport maps are

$$\overline{\Sigma} = \sum_{n=1}^{\infty}\left[\frac{\sqrt{\lambda_{1,n}} + \sqrt{\lambda_{1,n}}}{2}\right]^2 \varphi_n \otimes \varphi_n, \quad \mathbf{t}_1 = \frac{2}{1 + \sqrt{r_n}}\varphi_n \otimes \varphi_n,$$
$$\mathbf{t}_2 = \frac{2}{1 + \sqrt{r_n^{-1}}}\varphi_n \otimes \varphi_n,$$

and simple algebra shows

$$\left\|\left|\Delta_1\right|\right\|_2 < \infty \iff \sum_{n=1}^{\infty}(r_n - 1)^2 \iff \left\|\left|\Delta_2\right|\right\|_2 < \infty.$$

Thus, in the commutative case, the population level test statistic is finite if and only if the Gaussian measures are equivalent. Whether or not this is the case depends on how differences persist across the whole spectrum, rather than just in the bulk.

Proofs of formal statements

Proof of Lemma 2.1. If $\Sigma_1 = \cdots = \Sigma_K$, then the unique optimal multicoupling is given by the maps $\mathbf{t}_j(z) = z$ and the process $Z \sim N(0, \Sigma_1)$. Conversely, if a multicoupling of $(\gamma_1, \ldots, \gamma_K)$ is achieved as the law of $(\mathbf{t}_1(Z), \ldots, \mathbf{t}_K(Z))$ for some process Z and some maps satisfying $\mathbf{t}_1 = \cdots = \mathbf{t}_K$, then γ_i, the law of $\mathbf{t}_i(Z)$, is the same for all i, i.e., $\gamma_1 = \cdots = \gamma_K$, and so $\Sigma_1 = \cdots = \Sigma_K$. Uniqueness of the optimal multicoupling follows from the first sentence in the proof. □

For the proof of Theorem 2.2, we need the following result from Douglas [10].

Lemma A.1. *Let $0 \le A \le B$ be bounded operators, where $A \le B$ means that $B - A$ is non-negative. Then there exists a bounded operator G with $\left\|\left|G\right|\right\|_\infty \le 1$ such that $A^{1/2} = B^{1/2}G$ and $\ker G^* \supseteq \ker B$.*

Proof of Theorem 2.2. Let $\overline{\Sigma}$ be any Fréchet mean of $\Sigma_1, \ldots, \Sigma_K$ and define $Q_i = (\overline{\Sigma}^{1/2}\Sigma_i\overline{\Sigma}^{1/2})^{1/2} \geq 0$. The fixed point equation for Fréchet means ([27, Proposition 16]) yields the inequality

$$Q_i \leq \sum_{j=1}^{K} Q_j = K\overline{\Sigma}.$$

By Lemma A.1 there exists an operator G with range included in the closed range of $\overline{\Sigma}$ such that $Q_i^{1/2} = \overline{\Sigma}^{1/2}G$, $\|G\|_\infty \leq \sqrt{K}$ and we may write $G = \overline{\Sigma}^{-1/2}Q_i^{1/2}$. If we can identify G^* with $Q_i^{1/2}\overline{\Sigma}^{-1/2}$, then we can conclude that

$$\mathbf{t}_i = \mathbf{t}_{\overline{\Sigma}}^{\Sigma_i} = \overline{\Sigma}^{-1/2}(\overline{\Sigma}^{1/2}\Sigma_i\Sigma^{1/2})^{1/2}\overline{\Sigma}^{-1/2} = GG^*$$

is well-defined and bounded, with operator norm bounded by K. Let $y = \overline{\Sigma}^{1/2}z$ and notice that for all x

$$\begin{aligned}\langle G^*y, x\rangle = \left\langle \overline{\Sigma}^{1/2}z, Gx\right\rangle &= \left\langle \overline{\Sigma}^{1/2}z, \overline{\Sigma}^{-1/2}Q_i^{1/2}x\right\rangle \\ &= \left\langle z, Q_i^{1/2}x\right\rangle = \left\langle Q_i^{1/2}\overline{\Sigma}^{-1/2}y, x\right\rangle.\end{aligned}$$

Thus $G^* = Q_i^{1/2}\overline{\Sigma}^{-1/2}$ on range$(\overline{\Sigma}^{1/2})$. Since G^* is bounded the equality extends to the closure of the range, which is $(\ker\overline{\Sigma})^\perp$. On $\ker\overline{\Sigma}$ both operators are identically zero.

We have thus established the existence of deterministic optimal maps from the Fréchet mean $\overline{\Sigma}$ to each of the operators Σ_j. Now if $Z \sim N(0, \overline{\Sigma})$, then $\pi = (\mathbf{t}_1(Z), \ldots, \mathbf{t}_K(Z))$ is a multicoupling of the corresponding Gaussian measures, and the optimality of π follows from Zemel and Panaretos [44, Proposition 2]. □

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

Research supported in part by a Swiss National Science foundation grant to V. M. Panaretos

References

[1] Anderson, M. J. (2006), 'Distance-based tests for homogeneity of multivariate dispersions', *Biometrics* **62**(1), 245–253. MR2226579

[2] Benko, M., Härdle, W. and Kneip, A. (2009*a*), 'Common functional principal components.', *Ann. Stat.* **37**(1), 1–34. MR2488343
[3] Benko, M., Härdle, W. and Kneip, A. (2009*b*), 'Common functional principal components', *The Annals of Statistics* **37**(1), 1–34. MR2488343
[4] Boente, G., Rodriguez, D. and Sued, M. (2018), 'Testing equality between several populations covariance operators', *Annals of the Institute of Statistical Mathematics* **70**(4), 919–950. MR3830292
[5] Cabassi, A. and Kashlak, A. (2016), 'fdcov: Analysis of covariance operators', *R package version* **1**(0).
[6] Cabassi, A., Pigoli, D., Secchi, P. and Carter, P. A. (2017), 'Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology', *Electronic Journal of Statistics* **11**(2), 3815–3840. MR3714299
[7] Cuesta-Albertos, J. and Febrero-Bande, M. (2010), 'A simple multiway anova for functional data', *Test* **19**(3), 537–557. MR2746001
[8] Da Prato, G. and Zabczyk, J. (2014), *Stochastic equations in infinite dimensions*, Cambridge university press. MR3236753
[9] Delaigle, A. and Hall, P. (2012), 'Achieving near perfect classification for functional data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(2), 267–286. MR2899863
[10] Douglas, R. G. (1966), 'On majorization, factorization, and range inclusion of operators on hilbert space', *Proceedings of the American Mathematical Society* **17**(2), 413–415. MR0203464
[11] Dryden, I. L., Koloydenko, A. and Zhou, D. (2009), 'Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging', *The Annals of Applied Statistics* pp. 1102–1123. MR2750388
[12] Feldman, J. (1958), 'Equivalence and perpendicularity of gaussian processes.', *Pacific Journal of Mathematics* **8**(4), 699–708. MR0102760
[13] Fernique, X. (1970), 'Intégrabilité des vecteurs gaussiens', *CR Acad. Sci. Paris Serie A* **270**, 1698–1699. MR0266263
[14] Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer. MR2229687
[15] Fletcher, P. T., Lu, C., Pizer, S. M. and Joshi, S. (2004), 'Principal geodesic analysis for the study of nonlinear statistics of shape', *IEEE transactions on medical imaging* **23**(8), 995–1005.
[16] Fremdt, S., Steinebach, J. G., Horváth, L. and Kokoszka, P. (2013), 'Testing the equality of covariance operators in functional samples', *Scandinavian Journal of Statistics* **40**(1), 138–152. MR3024036
[17] Giri, N. (1968), 'On tests of the equality of two covariance matrices', *The Annals of Mathematical Statistics* **39**(1), 275–277. MR0223013
[18] Górecki, T. and Smaga, Ł. (2015), 'A comparison of tests for the one-way anova problem for functional data', *Computational Statistics* **30**(4), 987–1010. MR3433439
[19] Hájek, J. (1958), 'A property of j-divergences of marginal probability distributions', *Czechoslovak Mathematical Journal* **8**(3), 460–463. MR0099712
[20] Hastie, T., Buja, A. and Tibshirani, R. (1995), 'Penalized discriminant

analysis', *The Annals of Statistics* **23**(1), 73–102. MR1331657

[21] Hlávka, Z., Hlubinka, D. and Koňasová, K. (2022), 'Functional anova based on empirical characteristic functionals', *Journal of Multivariate Analysis* **189**, 104878. MR4384120

[22] Jones, H. E. and Bayley, N. (1941), 'The berkeley growth study', *Child development* pp. 167–173.

[23] Kashlak, A. B., Aston, J. A. and Nickl, R. (2019), 'Inference on covariance operators via concentration inequalities: k-sample tests, classification, and clustering via rademacher complexities', *Sankhya A* **81**(1), 214–243. MR3982196

[24] Kiefer, J. and Schwartz, R. (1965), 'Admissible bayes character of t2-, r2-, and other fully invariant tests for classical multivariate normal problems', *The Annals of Mathematical Statistics* pp. 747–770. MR0175245

[25] Kraus, D. and Panaretos, V. M. (2012), 'Dispersion operators and resistant second-order functional data analysis', *Biometrika* **99**(4), 813–832.

[26] Masarotto, V. and Masarotto, G. (2024), *fdWasserstein: Application of Optimal Transport to Functional Data Analysis*. R package version 1.0. URL `https://CRAN.R-project.org/package=fdWasserstein`.

[27] Masarotto, V., Panaretos, V. M. and Zemel, Y. (2018), 'Procrustes metrics on covariance operators and optimal transportation of gaussian processes', *Sankhya A* pp. 1–42. MR3982195

[28] Ocaña, F. A., Aguilera, A. M. and Valderrama, M. J. (1999), 'Functional principal components analysis by choice of norm', *Journal of Multivariate Analysis* **71**(2), 262–276. MR1735111

[29] Panaretos, V. M. (2018), 'Discussion of the statistical analysis of acoustic phonetic data: exploring differences between spoken romance languages'.

[30] Panaretos, V. M., Kraus, D. and Maddocks, J. H. (2010), 'Second-order comparison of gaussian random functions and the geometry of dna minicircles', *Journal of the American Statistical Association* **105**(490), 670–682. MR2724851

[31] Paparoditis, E. and Sapatinas, T. (2016), 'Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data', *Biometrika* **103**(3), 727–733. MR3551795

[32] Pesarin, F. and Salmaso, L. (2010), 'The permutation testing approach: a review', *Statistica* **70**(4), 481–509.

[33] Pigoli, D., Aston, J. A., Dryden, I. L. and Secchi, P. (2014), 'Distances and inference for covariance operators', *Biometrika* **101**(2), 409–422. MR3215356

[34] Pigoli, D., Hadjipantelis, P. Z., Coleman, J. S. and Aston, J. A. (2018), 'The statistical analysis of acoustic phonetic data: exploring differences between spoken romance languages', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(5), 1103–1145. MR3873703

[35] R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

[36] Ramsay, J. and Silverman, B. (2005*a*), 'Springer series in statistics'.

[37] Ramsay, J. and Silverman, B. W. (2005*b*), *Functional Data Analysis*, Springer, New York. MR2168993

[38] Ramsay, J., Wickham, H., Ramsay, M. J. and deSolve, S. (2021), 'Package 'fda''.

[39] Rao, C. R. and Varadarajan, V. (1963), 'Discrimination of gaussian processes', *Sankhyā: The Indian Journal of Statistics, Series A* pp. 303–330. MR0183090

[40] Roy, S. N. (1953), 'On a heuristic method of test construction and its use in multivariate analysis', *The Annals of Mathematical Statistics* pp. 220–238. MR0057519

[41] Silverman, B. W. (1996), 'Smoothed functional principal components analysis by choice of norm', *The Annals of Statistics* **24**(1), 1–24. MR1389877

[42] Tavakoli, S. and Panaretos, V. M. (2016), 'Detecting and localizing differences in functional time series dynamics: a case study in molecular biophysics', *Journal of the American Statistical Association* **111**(515), 1020–1035. MR3561926

[43] Yao, F., Müller, H.-G. and Wang, J.-L. (2005), 'Functional data analysis for sparse longitudinal data', *J. Amer. Statist. Assoc.* **100**, 577–590. URL http://dx.doi.org/10.1198/016214504000001745. MR2160561

[44] Zemel, Y. and Panaretos, V. M. (2019), 'Fréchet means and procrustes analysis in wasserstein space', *Bernoulli* **25**(2), 932–976. MR3920362

[45] Zhang, J.-T. (2013), *Analysis of variance for functional data*, Chapman and Hall/CRC.