Electronic Journal of Statistics
Vol. 11 (2017) 3815–3840
ISSN: 1935-7524
DOI: 10.1214/17-EJS1347

Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology

Alessandra Cabassi

MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK
e-mail: ac2051@cam.ac.uk

Davide Pigoli

Department of Mathematics, King's College London, London, UK
e-mail: davide.pigoli@kcl.ac.uk

Piercesare Secchi

Department of Mathematics, Politecnico di Milano, Milan, Italy
e-mail: piercesare.secchi@polimi.it

and

Patrick A. Carter

School of Biological Sciences, Washington State University, Pullman, USA
e-mail: pacarter@wsu.edu

Abstract: In this paper, we generalize the metric-based permutation test for the equality of covariance operators proposed by Pigoli et al. (2014) to the case of multiple samples of functional data. To this end, the non-parametric combination methodology of Pesarin and Salmaso (2010) is used to combine all the pairwise comparisons between samples into a global test. Different combining functions and permutation strategies are reviewed and analyzed in detail. The resulting test allows to make inference on the equality of the covariance operators of multiple groups and, if there is evidence to reject the null hypothesis, to identify the pairs of groups having different covariances. It is shown that, for some combining functions, step-down adjusting procedures are available to control for the multiple testing problem in this setting. The empirical power of this new test is then explored via simulations and compared with those of existing alternative approaches in different scenarios. Finally, the proposed methodology is applied to data from wheel running activity experiments, that used selective breeding to study the evolution of locomotor behavior in mice.

MSC 2010 subject classifications: Primary 62G10, 62J15; secondary 62P10.

Keywords and phrases: Non-Euclidean metrics, non-parametric combination, post-hoc analysis, quantitative genetics.

Received January 2017.

1. Introduction

In recent years, an increasing number of applications has involved data that are best described as being functional. Examples can be found in medicine (West et al., 2007), neuroimaging (Jiang et al., 2009, Viviani et al., 2005), biology (Wu and Müller, 2010, Illian et al., 2009), finance (Laukaitis, 2008) and quality control (Colosimo and Pacella, 2010, Torres et al., 2011), to mention just a few fields.

These data asked for the development of new methodologies that take into account the properties of the functional data (see Ramsay and Silverman, 2005, Ferraty and Vieu, 2006 and Horváth and Kokoszka, 2012). Most recently, much attention has been devoted to inferential procedures for covariance operators of functional data. Panaretos et al. (2010) examined the testing of equality of covariance structures from two groups of functional curves generated from Gaussian processes and Fremdt et al. (2013) extended their approach to the case of non-Gaussian data. A similar asymptotic test after regularization of the pooled covariance operator is also presented in Ji and Ruymgaart (2008). These methods make use of test statistics based on the Karhunen–Loéve expansions of the covariance operators, thus exploiting the embedding of the space of covariance operators in the space of Hilbert–Schmidt operators, which is the infinite dimensional equivalent of embedding covariance matrices in the space of symmetric matrices. However, Pigoli et al. (2014) show that better results can be achieved by using metrics that take into account the non-Euclidean geometry of the space of covariance operators. The drawback is that explicit analytic distributions are not available for the test statistics based on these metrics and therefore the authors proposed to use a permutation approach to carry out the test.

The aim of this work is to extend this idea to the case of multiple samples of functional data. The testing of equality of several covariance operators has been first considered by Boente et al. (2014), that, in order to improve asymptotic approximations, proposed to apply a bootstrap procedure to calibrate the critical values of the test statistic obtained from the Hilbert–Schmidt norm of the differences between sample covariance operators. Paparoditis and Sapatinas (2016) investigated then the properties of an empirical bootstrap methodology, applicable to more than two populations, but its consistency has been proven only for test statistics based on the Hilbert–Schmidt norms and on the Karhunen–Loéve expansions of the covariance operators. More recently, Kashlak et al. (2016) applied concentration inequalities to the analysis of covariance operators. These allow to construct non-asymptotic confidence sets that can be used to make multiple-sample tests for the equality of covariances.

Since in the two-sample case the choice of the distance to define the test statistic has been shown to impact the inferential performance in many scenar-

ios (Pigoli et al., 2014), we propose here a more general approach that can be applied to test statistics defined through any valid distance between covariance operators. Previous works (Dryden et al., 2009; Pigoli et al., 2014) show that using distances that take into account the geometry of the space of covariance operators can benefit the statistical analysis. While we found out that this is the case in the simulation settings we consider in Section 3.1, different distances can be used if necessary, without any modification of the testing procedure. Moreover, an appropriate choice of the permutation strategy provides also pairwise tests between groups with a guaranteed control of the family-wise error rate. The proposed method has been implemented in R (R Core Team, 2016) and it has been made available in the R package "fdcov" (Cabassi and Kashlak, 2016).

Let us consider q samples of random curves. We assume that curves in sample i:

$$x_{i1}, \dots, x_{in_i} \in L^2(\Omega), \quad i = 1, \dots, q$$

are realizations of a random process with mean μ_i and covariance operator Σ_i. We would like to test the hypothesis

$$H_0 : \{\Sigma_1 = \Sigma_2 = \cdots = \Sigma_q\} \quad \text{against} \quad H_1 : \exists i \neq j \text{ s.t. } \Sigma_i \neq \Sigma_j.$$

Moreover, if the null hypothesis H_0 is rejected, we would like to identify which pairs of groups show a difference between covariance operators. To do this, we will rely on the non-parametric combination methodology introduced by Pesarin and Salmaso (2010) for multivariate permutation, which enables to combine many different partial tests in an overall global test. In our case, the idea is to combine all the pairwise comparisons between the q samples in order to obtain the p-value of the global test. Using this method, the post-hoc comparisons are straightforward: the global p-value and the partial p-values of the pairwise group comparisons are computed simultaneously. However, some care is required when jointly analyzing the latter, because a multiple testing problem arises. Thus, we suggest to use a step-down approach to control the family-wise error rate. The empirical power of the proposed test is evaluated through simulation studies and compared with those of previously proposed testing procedures.

Finally, we analyze the covariance operators of age-dependent wheel-running activity curves in mice (Morgan et al., 2003). These mice were from the 16-th generation of a large evolution experiment artificially selecting for high levels of wheel running activity (Swallow et al., 1998). Both the phenotypic and genotypic covariances in all functional biological traits, including growth curves and activity curves, are crucial because such covariances may constrain the evolution of the functional trait (Irwin and Carter, 2013, 2014). In this specific mouse experiment we wished to test the hypothesis that the phenotypic covariance structure of activity across age had evolved under 16 generations of selection relative to control lines of mice (Morgan et al., 2003). In addition, because there were 4 replicate selected lines and four replicate control lines, all derived from the same source population, we have the opportunity to test for the evolution of activity curves by genetic drift by comparing replicate lines within a given selection group. (see Koteja et al., 1999; Kane et al., 2008).

2. Testing equality of covariance operators

In this section, we describe the proposed strategy to test the equality of covariance operators across multiple groups, which allows for the use of the most appropriate metric for covariance operators in the problem at hand and, at the same time, for the investigation of pairwise difference between groups. First, we discuss a few possible choices of distance between covariance operators.

2.1. Metrics for covariance operators

Let x be a random function which takes values in $L^2(\Omega)$, $\Omega \subseteq \mathbb{R}$, such that $E(||x||^2_{L^2(\Omega)}) < +\infty$. The covariance operator Σ_x is defined, for $g \in L^2(\Omega)$, as $\Sigma_x g(t) = \int_\Omega c_x(s,t)g(s)\mathrm{d}s$, where

$$c_x(s,t) = \mathrm{cov}(x(s), x(t)) =$$

$$= E\left[(x(s) - E\left[x(s)\right])\left(x(t) - E\left[x(t)\right]\right)\right].$$

Then, Σ_x is a trace class, self-adjoint, compact operator on $L^2(\Omega)$ with non negative eigenvalues (see, e.g., Bosq, 2012, Section 1.5). Indeed, any compact operator T has a canonical decomposition that implies the existence of two orthonormal bases $\{u_k\}, \{v_k\}$ for $L^2(\Omega)$ such that $Tf = \sum_k \sigma_k \langle f, v_k\rangle u_k$, or, equivalently, $Tv_k = \sigma_k u_k$, where $\langle v, v\rangle$ indicates the inner product in $L^2(\Omega)$ and the non negative real numbers $\{\sigma_k\}_{k\in\mathbb{N}}$, are called the singular values of T. If the operator is self-adjoint, there exists an orthonormal basis $\{v_k\}$ such that $Tf = \sum_k \lambda_k \langle f, v_k\rangle v_k$, or, equivalently, $Tv_k = \lambda_k v_k$ and the sequence $\{\lambda_k\} \in \mathbb{R}$ is called the sequence of eigenvalues for T. A compact operator T is said to be trace class if the trace $\mathrm{tr}(T) = \sum_k \langle Te_k, e_k\rangle < +\infty$ for every orthonormal basis $\{e_k\}$. A compact operator T is said instead to be Hilbert–Schmidt if its Hilbert–Schmidt norm is bounded, i.e., $||T||^2_{\mathrm{HS}} = \mathrm{tr}(T'T) < +\infty$, where T' denotes the adjoint operator of T. The Hilbert–Schmidt norm is a generalization of the Frobenius norm for finite-dimensional matrices.

It is then possible to embed the space of covariance operators in the space of Hilbert–Schmidt operators and use the Hilbert–Schmidt distance $||\Sigma_1 - \Sigma_2||_{HS}$ to measure the distance between two covariance operators Σ_1 and Σ_2. However, this is an extrinsic metric based on the above embedding and thus ignores the geometry of the space of covariance operators, such as the trace class property and the non negativity constraints on the eigenvalues. The same is true for the other distances based on p-Schatten norms, such as the nuclear distance or the spectral distance (see Pigoli et al., 2014). Pigoli et al. (2014) show that when the covariance operator is the object of interest for the statistical analysis, taking into account the property of the space leads to tests with higher empirical power. This motivated the introduction of new metrics such as the square root distance and the Procrustes distance. These are examples of distances that are instead based on the additional structure available for covariance operators and the simulation studies reported in Section 3.1 do indeed confirm that this family of distances provide a better performance.

Let Σ be a self-adjoint trace class operator, there exists a Hilbert–Schmidt self-adjoint operator

$$(\Sigma)^{1/2} f = \sum_k \lambda_k^{1/2} \langle f, v_k \rangle v_k,$$

where λ_k are eigenvalues and v_k eigenfunctions of Σ. The square root distance between two covariance operators Σ_1 and Σ_2 is therefore defined as

$$d_R(\Sigma_1, \Sigma_2) = ||\Sigma_1^{1/2} - \Sigma_2^{1/2}||_{\text{HS}}.$$

The square root distance is based on the mapping of the two operators Σ_1 and Σ_2 from the space of covariance operators to the space of Hilbert–Schmidt operators, through the square root map. This is a particular choice among a family of such maps that transform the covariance operator Σ to a Hilbert–Schmidt operator L so that $\Sigma = LL'$. It is easy to see that L is defined up to a unitary operator R, since $(LR)(LR)' = LRR'L' = LL' = \Sigma$. In the case of square root distance, L is arbitrarily chosen to be symmetric. Therefore, a natural generalization consists in following a Procrustes approach to minimize the distance between two equivalence classes in the square root space. Pigoli et al. (2014) define the square of the Procrustes reflection size-and-shape distance between two covariance operators Σ_1 and Σ_2 as

$$d_P(\Sigma_1, \Sigma_2)^2 = \inf_{R \in O(L^2(\Omega))} ||L_1 - L_2 R||^2_{\text{HS}} =$$

$$= \inf_{R \in O(L^2(\Omega))} \text{trace}((L_1 - L_2 R)'(L_1 - L_2 R)),$$

where L_i are such that $\Sigma_i = L_i L_i'$, for $i = 1, 2$, and $O(L^2(\Omega))$ is the space of unitary operators on $L^2(\Omega)$. It can be seen that the square root and the Procrustes distances are indeed well defined distances, their property being ultimately induced by the Hilbert–Schmidt norm used in their definition.

2.2. Non-parametric combination

In this section we describe how it is possible to test the global hypothesis that all the covariance operators are equal across the groups by combining pairwise group comparisons which are based on the two-sample permutation test described in Pigoli et al. (2014). This approach will allow us to use any metric in the definition of the test statistics without making any assumption on the data generating process.

Let us assume we have q independent groups of functional data

$$x_{i1}, \ldots, x_{in_i} \in L^2(\Omega), \quad i = 1, \ldots, q.$$

and they are independent and identically distributed samples from a random process with distribution P_i, mean μ_i and covariance operator Σ_i. In the following, we denote with $\mathbf{x}_i$ the vector of observations $(x_{i1}, \ldots, x_{in_i})$ from group i.

We would like to test if the covariance operators are all equal. The global null hypothesis can be viewed as an intersection of partial null hypotheses and the global alternative hypothesis as the union of the corresponding alternative hypotheses, i.e.

$$H_0 : \bigcap_{i \neq j} H_0^{ij} \text{ against } H_1 : \bigcup_{i \neq j} H_1^{ij}, \text{ where } H_0^{ij} : \{\Sigma_i = \Sigma_j\} \text{ and } H_1^{ij} : \{\Sigma_i \neq \Sigma_j\}.$$

The idea is to combine the $k = q(q-1)/2$ two-sample tests for each pair of groups in a global test, using the non-parametric combination algorithm of Pesarin and Salmaso (2010).

Let $T_{ij} = d(S_i, S_j)$ be the test statistic of our choice, associated to the partial test H_0^{ij} of groups i and j respectively, where S_i, S_j are sample covariance operators of the corresponding groups and $d(\cdot, \cdot)$ is some distance between covariance operators. In particular, in this work we consider the square root, Procrustes and Hilbert–Schmidt distances defined in Section 2.1. Let us define by $\mathbf{T} = (T_{1,2}, T_{1,3}, \ldots, T_{q-1,q})$, the vector of all partial test statistcs T_{ij}, with $1 \leq i < j \leq q$.

The partial tests $H_0^{ij} : d(\Sigma_i, \Sigma_j) = 0$ against $H_1^{ij} : d(\Sigma_i, \Sigma_j) \neq 0$ marginally satisfy the assumptions required for the test (i.e. they are marginally unbiased, consistent and significant for large values) for any of the distances presented in Pigoli et al. (2014). Therefore, the considered algorithms can be applied to any functional dataset using the vector of test statistics $\mathbf{T}$.

The partial test statistics in $\mathbf{T}$ are combined by a function Ψ that must satisfy the properties indicated by Pesarin and Salmaso (2010):

1. Ψ is non-decreasing in each argument,
2. If one or more arguments are zero, Ψ attains its supremum value $\bar{\Psi}$, possibly not finite.
3. For all $\alpha > 0$, the critical value T_Ψ^α of Ψ is assumed to be finite and strictly smaller than $\bar{\Psi}$.

Also, the curves must be centred around the sample mean of each group, because exchangeability of the observations is required in order to apply permutations.

We indicate by $x_{ij}^{(0)}$ the observations centred around the sample mean of the group m_i, by $\boldsymbol{x}_i^{(0)}$ the vector of centred observations of group i and by $S_i^{(0)}$ the associated sample covariance operator. Similarly, we indicate by $\boldsymbol{\pi}^{(b)}$ the b-th permutation of the data labels and so the superscript (b) indicates the centred dataset, permuted according to $\boldsymbol{\pi}^{(b)}$.

We obtain the following algorithm:

Algorithm 2.1 (Multiple-sample permutation test for the equality of covariance operators)**.**
Let $x_{ij}, i = 1, \ldots, q, j = 1, \ldots, n_i$ be the considered dataset.

1. *Let $x_{ij}^{(0)} = x_{ij} - m_i$, where m_i is the sample mean of $\mathbf{x}_i$, for all $i = 1, \ldots, q$, $j = 1, \ldots, n_i$.*

2. *Let* $\mathbf{T}^{(0)}$ *be the* k*-dimensional vector containing the pairwise distances between the sample covariance operators of the centred groups* $\mathbf{x}_i^{(0)}$ *and* $\mathbf{x}_j^{(0)}$, $d(S_i^{(0)}, S_j^{(0)})$, *for all* $1 \leq i < j \leq q$.
3. *For* $b = 1, \ldots, B$, *consider a random permutation* $\boldsymbol{\pi}^{(b)}$ *of the data labels and compute the* k*-dimensional vector* $\mathbf{T}^{(b)}$ *containing the distances between the sample covariance operators of the groups of the permuted dataset,* $d(S_i^{(b)}, S_j^{(b)})$, *for all* $1 \leq i < j \leq q$. $\{\mathbf{T}^{(b)}\}_{b=1}^{B}$ *is a random sampling from the permutational distribution of the random vector* $\mathbf{T}$.
4. *Let*
$$\hat{p}_{ij}(d) = \frac{\sum_b \mathbb{1}\big[d(S_i^{(b)}, S_j^{(b)}) \geq d\big]}{B}$$
be consistent estimates of $p_{ij}(d) = \mathbb{P}\big(d(S_i^{(b)}, S_j^{(b)}) \geq d\big)$, $d \in \mathbb{R}, d \geq 0$.
5. *Compute the estimated partial* p*-values of the test as* $\hat{p}_{ij} = \hat{p}_{ij}(d(S_i^{(0)}, S_j^{(0)}))$.
6. *Combine the* $\hat{p}_{ij}$ *through the combining function* Ψ *to obtain the observed global test statistic* $T_\Psi^{(0)} = \Psi(\hat{p}_{1,2}, \hat{p}_{1,3}, \ldots, \hat{p}_{q,q-1})$.
7. *For* $b = 1, \ldots, B$, *compute the* b*-th combined test statistic as*
$$T_\Psi^{(b)} = \Psi(\hat{p}_{1,2}^{(b)}, \hat{p}_{1,3}^{(b)}, \ldots, \hat{p}_{q-1,q}^{(b)}), \text{ where } \hat{p}_{ij}^{(b)} = \hat{p}_{ij}\big(d(S_i^{(b)}, S_j^{(b)})\big).$$
8. *Compute the estimate of the* p*-value of the combined test*
$$\hat{p}_\Psi = \frac{\sum_b \mathbb{1}[T_\Psi^{(b)} \geq T_\Psi^{(0)}]}{B}.$$
9. *If* $\hat{p}_\Psi \leq \alpha$, *reject* H_0.

Proposition 1. *If we make the additional assumptions that, when* n *goes to infinity, then so also do the sample sizes of all groups and that the number* B *of Monte Carlo iterations goes to infinity while* k *and* α *remain fixed, then it is possible to prove that the test we obtain is strongly consistent and unbiased for the overall null hypothesis* H_0 *against the alternative* H_1.

This is a direct consequence of Theorems 2, 4.3.1 and 3, 4.3.2 of Pesarin and Salmaso (2010).

2.3. Synchronized permutation tests

Step 3. of Algorithm 2.1 requires to generate a certain number of permutations of the original dataset. When data belong to multiple groups, different strategies can be used to generate the permuted samples. In Solari et al. (2009), three different ways of permuting data are proposed.

The simplest idea is to perform permutations involving the whole dataset, so-called *pooled* permutations. This means that a permutation $\boldsymbol{\pi} : \{1, \ldots, \sum_i n_i\} \to \{1, \ldots, \sum_i n_i\}$ is applied to the whole data vector $\mathbf{X}' = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$. That is,

defining by $P^{\boldsymbol{\pi}}$ the permutation matrix associated to permutation $\boldsymbol{\pi}$, such that the (i,j)-th element of the matrix $P^{\boldsymbol{\pi}}_{i,j} = 1$ if $j = \boldsymbol{\pi}(i)$ and $P^{\boldsymbol{\pi}}_{i,j} = 0$ otherwise, then the permuted dataset is

$$\mathbf{X}^* = P^{\boldsymbol{\pi}}\mathbf{X}.$$

This can be done because, under H_0, the observations of all groups are exchangeable. So the first permuted group of observations $\mathbf{x}^*_1$ is composed by the first n_1 elements of $\mathbf{X}^*$, the second one, $\mathbf{x}^*_2$, by the following n_2 elements, and so on, and the test statistics T^*_{ij} are computed using these permuted groups. However, this strategy does not allow to test also the partial hypotheses, since each pairwise comparison based on T^*_{ij} involves not only the observations belonging to the pair of considered groups, but also those of the other groups. Therefore, the resulting global p-value is correct, but the partial p-values would not be accurate when doing post-hoc comparisons.

The second proposal is to apply *paired* permutations, that is, while comparing the i-th and j-th groups, the inference is made on each paired vector $\mathbf{X}'_{ij} = (\mathbf{x}_i, \mathbf{x}_j)$ independently. In other words, for each pair (i,j) we define a permutation $\boldsymbol{\pi}_{ij} : \{1,2,\ldots,n_i+n_j\} \to \{1,2,\ldots,n_i+n_j\}$ and permute each paired vector independently:

$$\mathbf{X}^*_{ij} = P^{\boldsymbol{\pi}_{ij}}\mathbf{X}_{ij}$$

The result would be opposite than the one obtained with pooled permutations: the partial tests are exact, just like in the two-sample case, but the global test is not reliable since this method does not take into account the dependencies between the marginal tests.

Therefore, we want paired permutations to be done not independently but jointly. At the same time, we would like to keep the partial comparisons separate, so as to be able to do post-hoc comparison with no additional computational effort. Then, if the design is balanced, i.e. $n_1 = \cdots = n_q = \bar{n}$, a further possibility is to apply *synchronized* permutations, exchanging the same number ν of units between each pair of blocks. As shown by Solari et al. (2009), applying synchronized permutations allows both maintaining the dependencies among partial tests and involving the observations of each comparison at the same time. In particular, here we choose to apply *constrained* synchronized permutations, that is to exchange units in the same original position within each block. This can be achieved simply by appliying the same permutation $\boldsymbol{\pi} : \{1,2,\ldots,2\bar{n}\} \to \{1,2,\ldots,2\bar{n}\}$ to each paired vector $\mathbf{X}_{ij}$:

$$\mathbf{X}^*_{ij} = P^{\boldsymbol{\pi}}\mathbf{X}_{ij}.$$

In conclusion, if the groups are balanced, it is better to use synchronized permutations, since they allow to produce an approximated distribution for each partial test statistic T_{ij}, similarly to the two-sample case, at the same time as the approximated distribution for the global test statistic T_Ψ. In all other settings, pooled permutations can be used and only the global hypothesis will be considered.

2.4. Post-hoc analysis

After performing the global test, if the null hypothesis H_0 is rejected in favour of the alternative H_1, it is often of interest to find out which of the data samples led to this conclusion. One of the advantages of the non-parametric combination methodology is that partial p-values are computed at the same time of the global one. Therefore, the post-hoc comparisons can be done with a small computational effort. We investigate here the methods that allow to control the family-wise error rate, in order to simultaneously assess which of the partial null hypotheses H_0^{ij} are rejected.

First, we recall the resampling step-down method proposed by Westfall and Young (1993). The idea is that, rather than adjusting all p-values according to the minimum p-value distribution, one should only adjust the minimum p-value using this distribution and then adjust the remaining p-values according to smaller and smaller sets of p-values. The effect of using restricted sets of p-values is to make the adjusted p-values smaller, thereby improving the power of the method.

Let the ordered partial p-values have indexes $r_1, \ldots, r_k$ so that $\hat{p}_{(1)} = \hat{p}_{r_1}$, $\hat{p}_{(2)} = \hat{p}_{r_2}, \ldots, \hat{p}_{(k)} = \hat{p}_{r_k}$. The step-down adjusted p-values are defined sequentially as follows:

$$\tilde{p}_{(1)} = \mathbb{P}\left(\min_{j \in \{r_1, \ldots, r_k\}} \hat{p}_j \leq \hat{p}_{(1)} | H_0 \right)$$

$$\tilde{p}_{(i)} = \max \left\{ \tilde{p}_{(i-1)}, \mathbb{P}\left(\min_{j \in \{r_i, \ldots, r_k\}} \hat{p}_j \leq \hat{p}_{(i)} | H_0 \right) \right\}, \quad i = 2, \ldots, k.$$

The use of max operator insures that the order of the adjusted p-values is the same as that of the original p-values. Westfall and Young (1993) proved that this procedure controls the family-wise error rate in the strong sense.

Pesarin and Salmaso (2010) showed that the resampling method proposed by Westfall and Young (1993) is equivalent to iteratively use the non-parametric combination with the Tippett combining function Ψ_{Tippett} (Birnbaum, 1954):

Algorithm 2.2 (Step-down method for the Tippett combining function)**.**

Let $p_{(1)}, \ldots, p_{(k)}$ be the increasing ordered p-values corresponding to the set of partial hypotheses.

1. $\tilde{p}_{(1)} = \Psi_{Tippett}(p_{(1)}, \ldots, p_{(k)}) = \min(p_{(1)}, \ldots, p_{(k)})$,
   - *If $\tilde{p}_{(1)} \leq \alpha$, reject the corresponding hypothesis $H_0^{(1)}$ and continue;*
   - *Otherwise retain the hypotheses $H_0^{(1)}, \ldots, H_0^{(k)}$ and stop.*
2. *For $i = 2, \ldots, k$, $\tilde{p}_{(i)} = \Psi_{Tippett}(p_{(i)}, \ldots, p_{(k)})$*
   - *If $\tilde{p}_{(i)} \leq \alpha$, reject also $H_0^{(i)}$ and continue;*
   - *Otherwise retain the hypotheses $H_0^{(i)}, \ldots, H_0^{(k)}$ and stop.*

Furthermore, Lehmann and Romano (2006) presented a similar step-down method, that uses the test statistics T_{ij} instead of the p-values p_{ij}. This method is equivalent to the one based on the Tippett combining function but allows to avoid the computations of the permutational distributions of the partial p-values. For example, let us suppose that the individual tests H_0^{ij} are based on test statistics T_{ij} with large values indicating evidence against the partial null hypotheses. Let $K = \{H_0^{ij}, 1 \leq i < j \leq q\}$ be the set of all the partial test hypotheses and $\bar{K}$ a subset of K, $\bar{K} \subseteq K$. First of all, we have to define the critical value of the combined test of all the hypotheses contained in $\bar{K}$ at level $\alpha \in [0,1]$ so that the family-wise error rate is controlled in the strong sense. Many definitions are possible, as long as the properties indicated in Lehmann and Romano (2006), Theorem 9.1.3 are verified. We choose to use the definition given in Solari et al. (2009), where the critical value of $\bar{K}$ at level α is defined as the m-th smallest value among the permutation distributions of $T_{\bar{K}} = \max_{H_0^{ij} \in \bar{K}} T_{ij}$

$$c_{\bar{K}}(\alpha) = \left\{ \max_{H_0^{ij} \in \bar{K}} T_{ij}^{(b)}, b = 1, \ldots, B \right\}_{(m)},$$

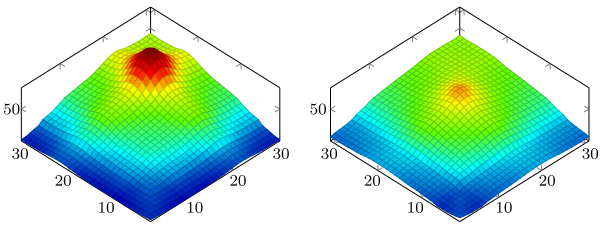
where $m = B - \lfloor B\alpha \rfloor$. For this reason we will refer to this as the step-down method for the $\max T$ combining function. The algorithm is defined as follows:

Algorithm 2.3 (Step-down method for the $\max T$ combining function)**.**

Let $T_{(1)} = T_{r_1} \geq \ldots \geq T_{(k)} = T_{r_k}$ denote the observed ordered test statistics where $r_1, \ldots, r_k$ are such that $T_{r_1} \geq T_{r_2} \geq \ldots \geq T_{r_k}$ and let $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(k)}$ be the corresponding hypotheses.

1. *Let $K_1 = K$,*
   - *If $T_{r_1} \geq c_{K_1}(\alpha)$ reject $H_0^{(1)}$ and continue;*
   - *Otherwise retain the hypotheses $H_0^{(1)}, \ldots, H_0^{(k)}$ and stop.*
2. *For $i = 2, \ldots, k$, let K_i be the set of hypotheses not previously rejected, i.e. $K_i = \{H_0^{(i)}, \ldots, H_0^{(k)}\}$,*
   - *If $T_{r_i} \geq c_{K_i}(\alpha)$ reject $H_0^{(i)}$ and continue;*
   - *Otherwise retain the hypotheses $H_0^{(i)}, \ldots, H_0^{(k)}$ and stop.*

Lastly, when using another combining function, it is possible to use the closed testing procedure of Marcus et al. (1976). This method is based on the idea that one may reject any hypothesis H_0^{ij}, while controlling the family-wise error rate, when the test of H_0^{ij} itself is significant and the test of every intersection of partial hypotheses that includes H_0^{ij} is significant. Hence, $\tilde{p}_{ij}$ is the maximum of all the p-values of the partial hypotheses containing H_0^{ij}. This method has two major drawbacks: it requires a greater number of computations and it is very conservative. However, it is a useful tool when the use of the Tippett or $\max T$ combining functions is not suitable.



(a) Male (Σ_1). (b) Female (Σ_2).

FIG 1. *Covariance operators of the subjects in the Berkeley growth study dataset.*

3. Simulation studies

3.1. Synthetic datasets

We generate synthetic datasets as follows. All the curves are generated on an equispaced grid of 31 points on $\Omega = [0, 1]$ and the sample size of each group is $\bar{n} = 20$. Unless otherwise stated, curves are simulated from a multivariate Gaussian process. We consider q different groups (with q varying across simulation studies) and for all q groups the mean function is equal to $\sin(x)$, $x \in [0, 1]$. The covariance operator of each group varies according to the test case. Let Σ_1 and Σ_2 be the sample covariance operators of the male and female subjects in the Berkeley growth study dataset described in Ramsay and Silverman (2005), rescaled to $[0, 1]$. We obtained this covariance operators from the functional dataset available in the R package `fda` (Ramsay et al., 2014), without any preprocessing. A surface representation of Σ_1 and Σ_2 can be seen in Figure 1.

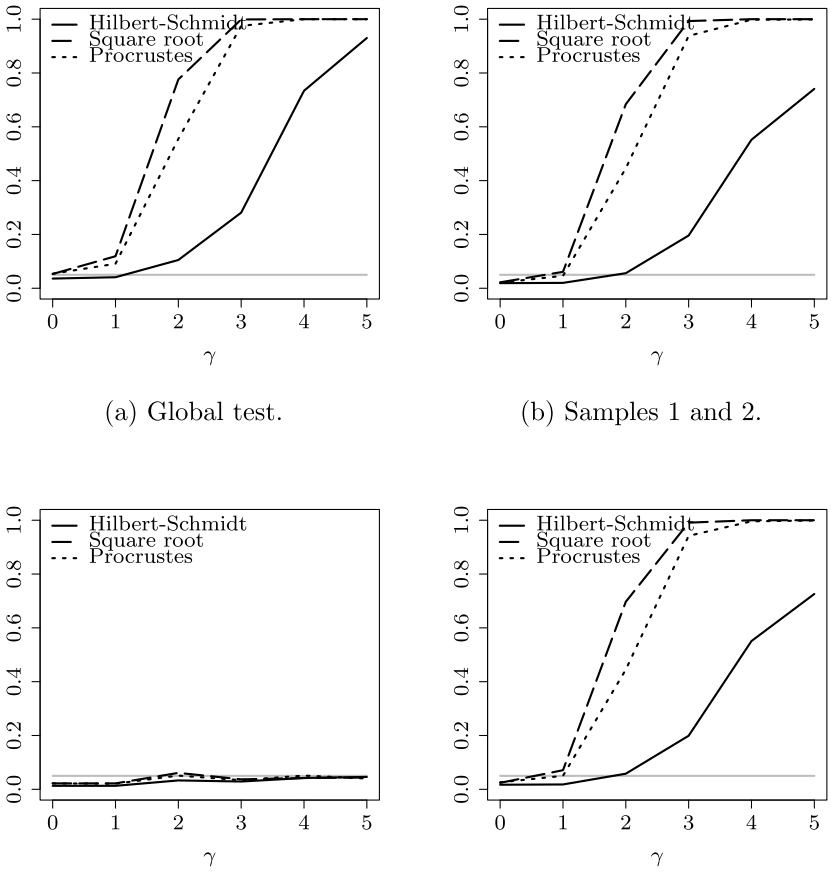
We then consider two forms for the expression of the covariance operators of some of the groups under the alternative:

First test case Some of the groups have covariance operator $\Sigma(\gamma) = [(\Sigma_1)^{1/2} + \gamma\{(\Sigma_2)^{1/2}\tilde{R} - (\Sigma_1)^{1/2}\}][(\Sigma_1)^{1/2} + \gamma\{(\Sigma_2)^{1/2}\tilde{R} - (\Sigma_1)^{1/2}\}]'$ where $\tilde{R}$ is the operator minimizing the Procrustes distance between Σ_1 and Σ_2 (Pigoli et al., 2014) and $\gamma \in [0, 5]$ is a parameter which controls how far this covariance operator is from Σ_1.

Second test case Some of the groups have covariance operator $\Sigma(\gamma) = (1 + \gamma)\Sigma_1$, $\gamma \in [0, 5]$.

The two test cases represent two different ways in which the null hypothesis can be violated. The second case pertains to a difference in the total variation between groups, while the first test case presents also a difference in shape between covariance operators.

Each permutation test is performed with $B = 1000$ iterations of the Monte Carlo Algorithm 2.1 and is repeated for 1000 replicates of the simulated dataset. In the following, we use this simulated data to evaluate the empirical size and the empirical power of the proposed test when using different distances between covariance operators. All the functions needed to apply the permutation tests



(a) Global test.

(b) Samples 1 and 2.

(c) Samples 2 and 3.

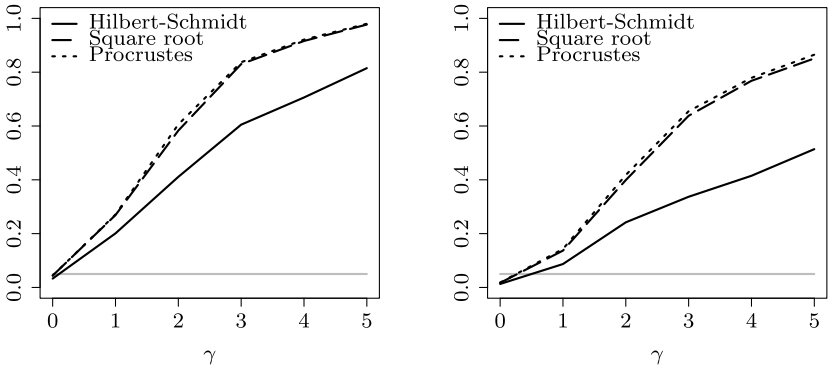
(d) Samples 1 and 3.

FIG 2. *Empirical power of synchronized permutation global and partial tests applied to the first test case using* $\max T$ *combining function. p-values have been adjusted using the* $\max T$ *step-down procedure.*

to these simulated data have been made available in the R package "fdcov" (Cabassi and Kashlak, 2016).

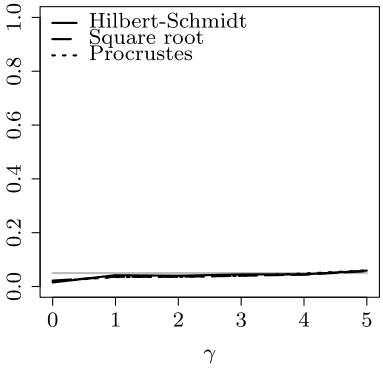
3.2. Empirical size and power of the test

We consider first a simulation with $q = 3$ groups, where the first group has covariance operator Σ_1 and the others two covariance operators $\Sigma(\gamma)$. Figures 2 and 3 show the empirical power of the global and partial tests done using the synchronized permutations, the $\max T$ combining function and the Procrustes, square root and Hilbert–Schmidt distance, for the first and second test cases respectively. It is evident that the test has greater empirical power when using

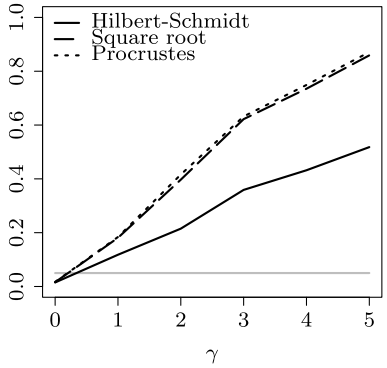


(a) Global test.

(b) Samples 1 and 2.



(c) Samples 2 and 3.

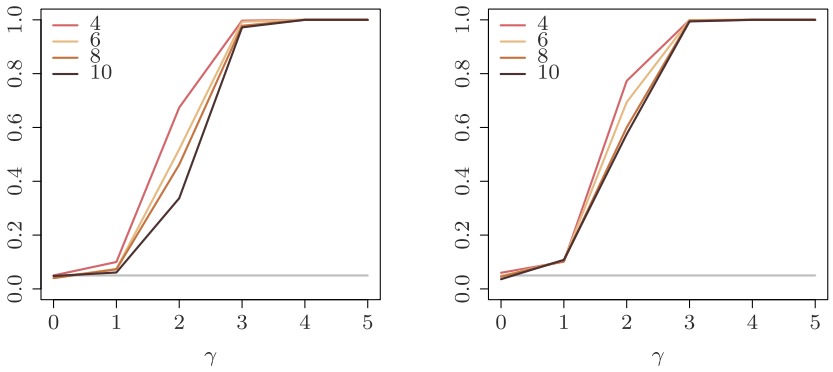


(d) Samples 1 and 3.

FIG 3. *Empirical power of synchronized permutation global and partial tests applied to the second test case using the* $\max T$ *combining function. p-values have been adjusted using the* $\max T$ *step-down procedure.*

Procrustes and square root distances, with the latter being in this case preferable due to the lower computational cost. Moreover, the global test appears to have the correct level for all the distances while the partial tests are conservative for $\gamma = 0$, as expected, and the proportion of rejection for the partial test between the second and third group (which have equal covariances) is close to or less than 5% for all values of γ.

We want now to explore how the performance of the test changes when the number of groups increases. Figure 4(a) shows the empirical power of the global test using the square root distance when the number of groups goes from 4 to 10, always with the first group with covariance operator Σ_1 and all the others with covariance operator $\Sigma(\gamma)$, with γ varying from 1 to 5. It is possible to see



(a) One group has fixed covariance Σ_1, the others have covariance Σ_γ.

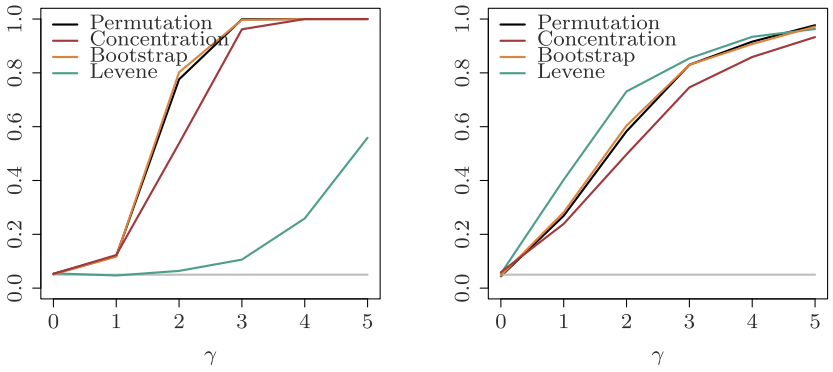
(b) Half of the groups have fixed covariance Σ_1, the others covariance Σ_γ.

FIG 4. *Empirical power of synchronized permutation global tests applied to the first test case using the* $\max T$ *combining function, with 4, 6, 8 and 10 data samples.*

that the level of the test is respected for all numbers of groups while the empirical power tends to decrease when the number of groups increases. This is due to the fact that only q partial tests out of $q(q-1)/2$ are bringing information about the violation of the null and they form a smaller and smaller proportion of all the partial tests when q increases. If we instead have half of the groups with covariance operator Σ_1 and half with covariance operator $\Sigma(\gamma)$, the loss of empirical power when q increases is smaller, as shown by the empirical power curves reported in Figure 4(b). This is because of the larger proportion of false partial hypothesis.

3.3. Comparison with the other existing tests

We compare now the proposed method, using the square root distance and the $\max T$ combining function, with some alternative approaches to test for the difference between covariance operators. We consider first a generalization of the Levene's test (Anderson, 2006) which is sensitive only to the difference in total variation between groups and it is implemented using the permutational analysis of variance. Paparoditis and Sapatinas (2016) introduced an empirical bootstrap approach based on Hilbert–Schmidt distance (or alternatively, on other test statistics based on the Karhunen–Loéve expansions of the covariance operators). In the interest of a fair comparison, we apply here the same procedure to the test statistics based on the square root distance. It should be noted however that the theoretical properties of this modified procedure still need to be studied. Finally, we consider the test based on the concentration inequalities method of Kashlak et al. (2016).

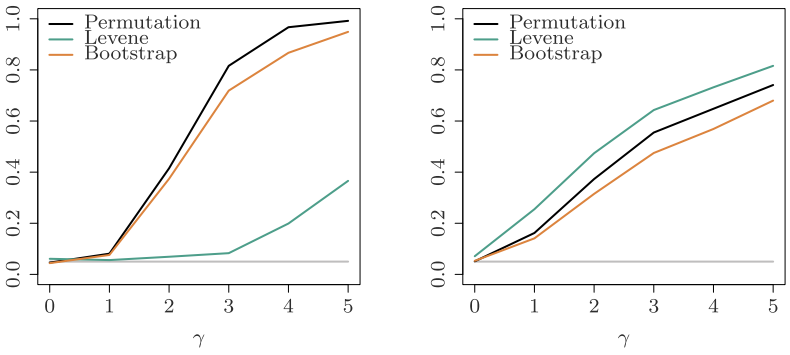


(a) First test case. (b) Second test case.

FIG 5. *Empirical power of synchronized permutation, Levene's, empirical bootstrap and concentration inequalities-based global tests applied to the first (left) and second (right) test cases. Data are sampled from a Gaussian process. The results shown were obtained using the combining function* $\max T$ *and the p-values have been adjusted with the step-down procedure.*

Figure 5 shows the empirical power of the generalisation of Levene's test (Anderson, 2006), the empirical bootstrap by Paparoditis and Sapatinas (2016) and the concentration inequalities method of Kashlak et al. (2016) for the two test cases, compared to the results obtained using the proposed permutation test. Here data are simulated from $q = 3$ groups with the first group having covariance operator Σ_1 and the other two covariance operators $\Sigma(\gamma)$. It appears that the permutation test and the empirical bootstrap have approximately the same empirical power in both test cases. On the contrary, Levene's test performs very differently. As expected, it outperforms the others in the second test case, where it captures very well the differences in scale, but it is dramatically less powerful in the first test case, where the difference between the covariance operator is mostly in shape. The non-asymptotic test of Kashlak et al. (2016) is slightly less powerful than the permutation test and the empirical bootstrap but it has the advantage of being much less computationally expensive than the resampling-based methods.

We want also to explore what happens when data are generated from a non-Gaussian distribution. We simulated data from a multivariate t distribution with 4 degrees of freedom and correlation matrix implied by the covariance operator Σ_1 for the first group and $\Sigma(\gamma)$ for the other two groups. Here it is not possible to apply the non-asymptotic test of Kashlak et al. (2016), because calibration parameters are not yet available when data are not Gaussian. Figure 6 shows the empirical power for the permutation test, the empirical bootstrap test and the Generalized Levene's test. Here the permutation test appears to perform slightly better then the bootstrap, while Levene's test is again performing very well in the second tests case but not in the first. Overall, the empirical power of all tests is lower than in the Gaussian case, but they respect the nominal level.



(a) First test case. (b) Second test case.

FIG 6. *Empirical power of synchronized permutation, Levene's and empirical bootstrap global tests applied to the first (left) and second (right) test cases. Data are generated using a multivariate t-Student. The results shown were obtained using the combining function* $\max T$ *and the p-values have been adjusted with the step-down procedure.*

4. Application to evolutionary biology

In this section we apply the proposed permutation test to a dataset of interest in evolutionary biology. We focus here on the phenotypic covariance function, i.e. the covariance between observed function-valued biological traits. While the interest of evolutionary biology may ultimately be on the genetic covariance, i.e. the proportion of covariance which is caused by genes, studying differences in the phenotypic covariance is also crucial because it may constrain the evolution of the functional trait (Irwin and Carter, 2013, 2014). The main question of interest here is whether there is a difference in the covariance operator of a function-valued trait (Kingsolver et al., 2001; Stinchcombe et al., 2012) among experimental lines of mice with known differences in evolutionary histories.

4.1. Dataset

Data were collected from aging house mice (*Mus domesticus*) that were members of the 16th generation of a selective breeding experiment for increased voluntary wheel-running behavior (Swallow et al., 1998). This experiment produced four replicate lines selected for the total number of wheel revolutions run on days 5 and 6 of a six day exposure to running wheels that occurred when the mice were six to eight weeks of age, and four replicate control lines that were randomly bred each generation (see Swallow et al., 1998, for additional details). At generation 16, a total of 360 mice were used to establish an aging colony (Morgan et al., 2003; Bronikowski et al., 2006). Half of the mice in the colony were from the four high-selected lines and half were from the four control lines that were randomly bred with respect to running behavior, and half of each selection group was housed with running wheels (active mice) and half was housed in cages without

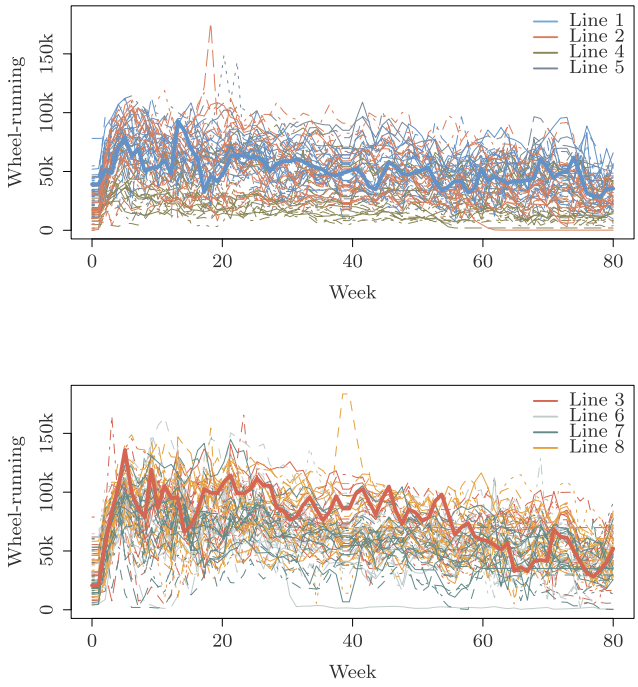


FIG 7. *Voluntary wheel running activity dataset, raw data, control and selected lines.*

wheels (sedentary mice). One male from one of the control lines died of unknown causes during the early stages of the experiment. Each week every mouse was measured for body mass and food consumed, and each active mouse had the total number of wheel revolutions run that week recorded (see Morgan et al., 2003; Bronikowski et al., 2006, for more details).

Herein we examine only data from the active mice from both selection groups from the first 80 weeks of the experiment, as reported by Morgan et al. (2003). The variables in the dataset are: a unique id number for each mouse and id of fullsib family (the group of first-degree relatives which, on average, share 50% of their genes) from which the mouse was drawn; the age and sex of the mouse; the line number (lines 1, 2, 4, 5 are control lines, the others are selected lines); the week of wheel measure and the number of revolutions run during the week. Some of these variables have been collected in view of identifying the genetic component of the covariance and will not be used in the present analysis. Total activity, measured as number of revolutions run in a given day, can be decomposed into the product of mean velocity and duration of activity. Thus, increased total activity levels could be accomplished by an increase in mean velocity, an increase in the amount of time spent running, or a combination of both. We will not consider here the family relationship between mice, although they would be an important part of any subsequent genetics analysis.

The raw data are presented in Figure 7. Each line connects the number of revolutions run by each mouse during the first 80 weeks of the experiment

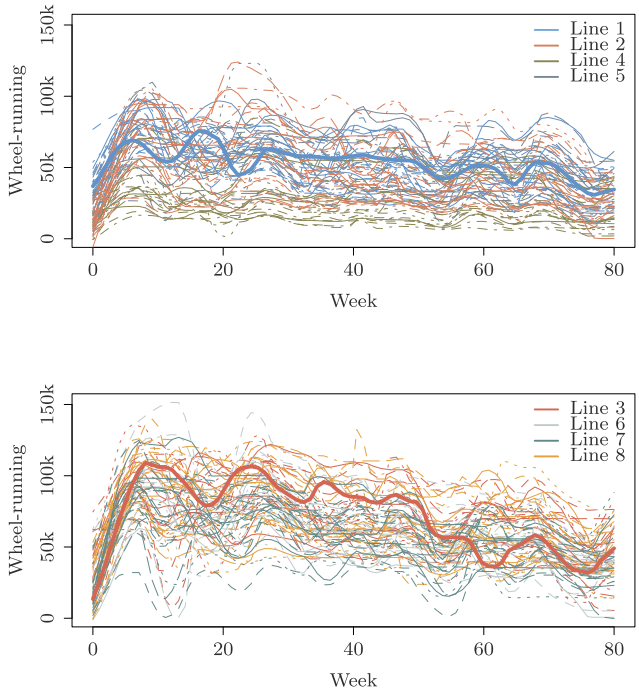


FIG 8. *Voluntary wheel running activity dataset after smoothing and alignment, control and selected lines.*

(Morgan et al., 2003). Mice identified by ID numbers 90183 and 90224 are taken as examples of the selected and control lines respectively. The corresponding wheel-running functions have been highlighted in each figure. The first one is a male belonging to family number 29 from line 1 (control), while the second one is a female belonging to family number 11 from line 3 (selected).

At several times during the experiment, data collection was skipped for one or two weeks. In these cases, the data collected after the skipped week(s) was divided by number of weeks, giving multiple weeks in a row with the same value. This is easily seen in Figure 7 at weeks 38, 39, 40, when the values are constant for each mouse, because the wheel revolutions recorded for week 40 were divided by 3 and assigned to weeks 38 and 39 as well as 40. The weeks in which this occurred are: 34; 35; 38; 39; 40; 50; 51; 72; 73.

We regularized data using cubic smoothing splines. In particular, we used the the routine `spline.smooth()` of the R package "stats" (R Core Team, 2016). Since individual mice can have their own biological clock, curves are aligned to remove phase variability (Ramsay and Silverman, 2005), via the elastic analysis described in Tucker et al. (2013) and implemented in the R package "fdasrvf" (Tucker, 2016). Figure 8 shows the smooth and aligned wheel-running activity curves.

4.2. Missing observation

In the voluntary wheel running activity dataset, all groups (experimental lines) are composed of 20 mice. However, one of the mice died of unknown causes during the early stages of the experiment and therefore one group has only 19 observations. For this reason, in order to apply the synchronized permutations, we have to prove that the presence of a missing observation does not affect the inference. Note that we could have used pool permutations if we were only interested in the global hypothesis, but the desire to carry out comparisons between lines as well motivated us to justify the extension of synchronised permutations to this setting.

Following the guidelines given by Pesarin and Salmaso (2010), we give a new formulation of the test that takes into account the presence of missing data. Thanks to this, we are able to prove that it is possible to apply the proposed test to an unbalanced dataset with one missing observation, under certain assumptions on the process that generates the missing observations.

Consider again a functional dataset of the form

$$\mathbf{X} = \{x_{ij},\ i = 1, \ldots, q,\ j = 1, \ldots, n_i\},$$

that consists of $q \geq 2$ samples of size $n_i \geq 2$. The groups are related to q levels of a treatment and the data x_{ij} are supposed to be independent and identically distributed with distributions $P_i \in \mathcal{P}$, $i = 1, \ldots, q$. In order to take into account that, for whatever reason, some of the data are missing, Pesarin and Salmaso (2010) suggested to consider the inclusion indicator associated to the considered dataset, that is

$$\mathbf{O} = \{o_{ij},\ i = 1, \ldots, q,\ j = 1, \ldots, n_i\},$$

where $o_{ij} = 1$ if x_{ij} has been observed and collected, $o_{ij} = 0$ otherwise. We denote with $\mathbf{o}_i$ the vector of observation indicators $(o_{i1}, \ldots, o_{in_i})$ from group i. This indicator represents the observed configuration in the dataset. Hence, the dataset can be seen as the pair of matrices $(\mathbf{X}, \mathbf{O})$. Therefore we would like to perform the following test:

$$H_0 : \{(\mathbf{x}_1, \mathbf{o}_1) \overset{d}{=} \ldots \overset{d}{=} (\mathbf{x}_q, \mathbf{o}_q)\} \quad \text{against} \quad H_1 : \{H_0 \text{ is not true}\}.$$

Thus, if we assume that data are jointly exchangeable under the null hypothesis with respect to the groups, we can, again, utilize a permutation test. Let us represent by P_i the joint multivariate distribution of $(\mathbf{x}_i, \mathbf{o}_i), i = 1, \ldots, q$ under the null hypothesis. Then it holds:

$$P_i = P_{\mathbf{o}_i} \cdot P_{\mathbf{x}_i|\mathbf{o}_i}.$$

The idea of Pesarin and Salmaso (2010) is to break down the null hypothesis in the following way:

$$H_0 : \{[\mathbf{o}_1 \overset{d}{=} \ldots \overset{d}{=} \mathbf{o}_q] \cap [\mathbf{x}_1 \overset{d}{=} \ldots \overset{d}{=} \mathbf{x}_q|\mathbf{O}]\} = \{H_0^{\mathbf{O}} \cap H_0^{\mathbf{X}|\mathbf{O}}\}.$$

Furthermore, we assume that the missing data are missing completely at random. In this case, we can condition with respect to the observed inclusion indicator and ignore $H_0^{\mathbf{O}}$, because $\mathbf{O}$ does not provide any information about treatment effects (Rubin, 1976). In other words, the partial hypotheses on $\mathbf{O}$ are true by assumption and the null hypothesis can be simplified:

$$H_0 = H_0^{\mathbf{X}|\mathbf{O}} = \{\mathbf{x}_1 \stackrel{d}{=} \ldots \stackrel{d}{=} \mathbf{x}_q|\mathbf{O}\}.$$

We indicate by $\mathbf{O}^*$ any permutation of $\mathbf{O}$, the permutational vector of inclusion indicators, and by $\boldsymbol{\kappa}^* = [\kappa_1^*, \ldots, \kappa_q^*]$ the corresponding vector of counts of valid observations in each group, where

$$\kappa_i^* = \sum_{j=1}^{n_i} o_{ij}^*, i = 1, \ldots, q.$$

Then we can group the set of all permutations of the dataset, according to the vectors of actual sample sizes of valid data $\boldsymbol{\kappa}^*$. Now, let $\mathbf{T}$ be the vector of partial test statistics based on functions of sample valid data; we denote its permutation distribution as $F[\mathbf{T}|(\mathbf{X}, \mathbf{O})]$, $\mathbf{T} \in \mathbb{R}^k$. Pesarin and Salmaso (2010) pointed out that, if the permutation sub-distributions of the partial test statistics are invariant with respect to the sub-groups induced by $\mathbf{O}^*$, then we can simply evaluate $F[\mathbf{T}|(\mathbf{X}, \mathbf{O})]$ ignoring the missing values. This implies that

$$F[\mathbf{T}|(\mathbf{X}, \mathbf{O})] = F[\mathbf{T}|(\mathbf{X}, \mathbf{O}^*)]$$

holds for every $\mathbf{T} \in \mathbb{R}^k$, for every permutation $\mathbf{O}^*$ of $\mathbf{O}$ and for all datasets $\mathbf{X}$. In the case of the tests for covariance operators, this is true because the test statistic T_Ψ of the global test is a combination of the partial test statistics of the pairwise comparisons between the groups. These, in turn, depend only on the distances between covariance operators and their permutations. We can suppose that, under the null hypothesis, the permutation distribution of the partial test statics T_{ij} depends essentially on the number κ_i^*, κ_j^* of summands. Thus, just like in the case of the multivariate analysis of variance studied in Pesarin and Salmaso (2010), the previous distributional equality is equivalent to

$$F[\mathbf{T}|(\mathbf{X}, \boldsymbol{\kappa})] = F[\mathbf{T}|(\mathbf{X}, \boldsymbol{\kappa}^*)], . \tag{1}$$

Hence, we would like our partial test statistics to be invariant with respect to $\boldsymbol{\kappa}^*$ and for all $\mathbf{X}$. Now, suppose that we are in the balanced case, i.e. $n_1 = \cdots = n_q = \bar{n}$ and one observation is missing in one of the groups, say group a, where $1 \leq a \leq q$. In the wheel-running dataset, for instance, $q = 8$, $\bar{n} = 20$ and one observation is missing in group 1. All the pairwise comparisons between groups i and j with $1 \leq i < j \leq q$ and $i, j \neq a$ are not affected by the problem of missing data since $\kappa_i^* = \kappa_j^* = \bar{n}$. As regarding the others, at each iteration of the algorithm, we could have $\kappa_a^* = \bar{n}$ and $\kappa_j^* = \bar{n} - 1$ or viceversa, depending on the permutation. However, since distances are symmetric, this two cases are permutationally equivalent under the null hypothesis and Equation (1) is always

satisfied. For this reason, we can apply the synchronised permutations as usual. At each iteration of Algorithm 2.1 the sample covariance of each permuted group is computed only with the available data. This is more complicated when the number of missing data becomes greater than one, since the vector $\boldsymbol{\kappa}^*$ of actual sample sizes can assume other values.

4.3. Hypothesis testing

We can finally apply the test to the smoothed and aligned wheel-running activity curves. The aim of the analysis is to check if the covariance operators of the eight groups of mice are the same and, if this is not the case, to identify which lines have different covariances. This is necessary for two reasons. First, the covariance operator is in itself of biological interest for exploring which type of variability is environmental in nature and which is due to genetic components. Second, inference on the mean functions often requires the assumption of equality of covariance operator and it is important to be able to check this assumption.

We want then to test the hypothesis

$$H_0 : \{\Sigma_1 = \cdots = \Sigma_8\} \quad \text{against} \quad H_1 : \{\text{at least one of the equalities is not true}\}.$$

To this end, we use the Monte Carlo Algorithm 2.1 to obtain an estimate of the permutation test proposed in Section 2.2. We have shown in the previous section that synchronized permutations can be used, even if one of the observations is missing. We use here the square root distance between covariance operators as partial test statistic and we choose the $\max T$ combining function. We set the number of iterations B to 1000. We choose to use the square root distance because it showed the best performance in the simulation study described in the previous section and it is also computationally less expensive then the Procrustes distance. However, the analysis carried out with the Procrustes distance would have led to the same conclusions.

The p-values of the partial tests between each pair of lines, adjusted with the step-down method, are reported in Figure 9. The p-value of the global test (< 0.001) indicates that there is strong evidence to reject the null hypothesis. This is due mainly to the first group of mice for which some partial null hypotheses are rejected (i.e., the differences between the covariance operator of line 1 and the covariance operators of these others lines are significant) and, when using the $\max T$ combining function, we reject H_0 even if only one of the partial tests is rejected.

The results of this test are somewhat surprising. First, no differences were detected between the selected lines and the control lines. Under the type of directional selection applied (Swallow et al., 1998) there is at least a theoretical expectation that genetic variances and covariances would evolve between selected and unselected populations (e.g. Falconer and Mackay, 1996); indeed, work on the 31st generation of mice from this same selection experiment demonstrated some evolution of the genetic variances of wheel running over the first 6 days of wheel running (Careau et al., 2015). Second, the results suggest that

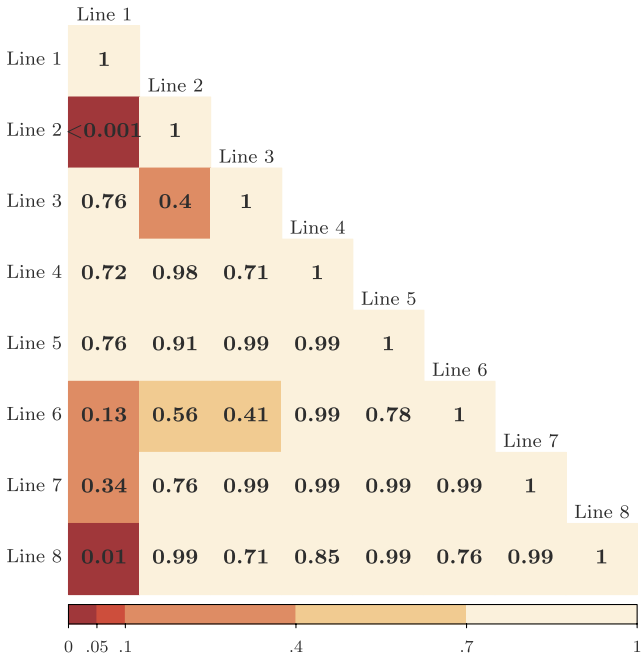


FIG 9. *Partial p-values of the synchronized permutation test on the covariance operators of the aligned data. For each* $1 \leq j \leq i \leq q$, $i \neq j$, *the value reported in row* i, *column* j *corresponds to the adjusted p-value of the test* $H_0 : \{\Sigma_i = \Sigma_j\}$ *against* $H_1 : \{\Sigma_i \neq \Sigma_j\}$. *The global p-value of the test is the minimum of the partial p-values and therefore is less than 0.001.*

line 1 randomly differs from one other control line and one other selected line. Such random differences in biological populations can be caused by genetic drift occurring during the selection experiment or by founder effects when the original base population was randomly subdivided into eight lines. Indeed, the trait that was actually under selection (wheel running on days 5 and 6 of a 6 day exposure) and underlying physiological traits (e.g., basal metabolic rate) already demonstrated the effects of drift and/or founder effects in Swallow et al. (1998) and Kane et al. (2008). The results presented herein are strongly suggestive of similar processes influencing the phenotypic covariance structure of wheel running across age, which presents interesting possibilities for additional biological experiments examining the impacts of constraints on functional traits (Irwin and Carter, 2013, 2014).

5. Conclusions and further developments

We extended the application of hypothesis tests that take into account the geometry of the space of covariance operators to the case of multiple groups, using a permutation approach. In particular, synchronized permutations allow us to

make inference also on the pairwise comparison between groups while controlling the family-wise error rate. We illustrate via simulation studies that the proposed test has the correct effect size and indeed the square root distance and the Procrustes distance lead to higher empirical power in the multiple groups comparison as well. While we have shown that the method can be applied in the case of a missing observation, a more general treatment of the case of unbalanced design and missing data is scope for future works. However, even in case of more general unbalanced design is still possible to apply the proposed method using pooled permutations, although only the global hypothesis can be tested in this way. It is worth to notice that other methods (such as bootstrap) are focused on the global hypothesis as well and a satisfactory treatment of partial hypothesis for unbalanced designs still needs to be devised.

We have also shown that the empirical power for the global test is comparable to those obtained using bootstrap approximation in the Gaussian case and slightly better in the non-Gaussian case. It is worth to notice that, while simulation results shows the bootstrap approach to be promising as well for the global test, its property has not yet rigorously studied for test statistics based on metric different from the Hilbert–Schmidt distance and this is an interesting direction for future research.

The application of the procedure to the mice voluntary wheel running activity curves shows that, while a difference between covariance operators is indeed present, this is not caused by selection itself. Instead it would appear that random biological processes such as genetic drift or founder effects are influencing the covariance operators of the phenotypic curves. This is an important result that suggests further investigation of this trait and demonstrates the importance of random processes during evolution. We did not consider here the family relationships between mice. However, in finite populations under selection, it may be possible for the family relationships to introduce dependencies in wheel-running activity curves; how to account for this would be a major focus when developing any subsequent genetics analysis and it will be scope for future work.

Acknowledgments

A. Cabassi was supported by the MRC (project reference `MC_UP_0801/1`) and by a "Tesi all'estero" scholarship from Politecnico di Milano, Italy. The authors also wish to thank The Washington State University College of Arts and Sciences, Office of International Programs and Office of Research for travel grants to P.A. Carter.

References

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

Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*, 49(267):559–574. MR0065101

Boente, G., Rodriguez, D., and Sued, M. (2014). A test for the equality of covariance operators. *arXiv preprint arXiv:1404.7080*. MR2815560

Bosq, D. (2012). *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media. MR1783138

Bronikowski, A., Morgan, T., Garland, T., and Carter, P. (2006). The evolution of aging and age-related physical decline in mice selectively bred for high voluntary exercise. *Evolution*, 60(7):1494–1508.

Cabassi, A. and Kashlak, A. B. (2016). *fdcov: Analysis of Covariance Operators*. R package version 1.0.0.

Careau, V., Wolak, M. E., Carter, P. A., and Garland, T. (2015). Evolution of the additive genetic variance–covariance matrix under continuous directional selection on a complex behavioural phenotype. In *Proc. R. Soc. B*, volume 282, page 20151119. The Royal Society.

Colosimo, B. M. and Pacella, M. (2010). A comparison study of control charts for statistical monitoring of functional data. *International Journal of Production Research*, 48(6):1575–1601.

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*, pages 1102–1123. MR2750388

Falconer, D. S. and Mackay, T. F. (1996). *Introduction to quantitative genetics*. Longman, Essex United Kingdom.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media. MR2229687

Fremdt, S., Steinebach, J. G., Horvath, L., and Kokoszka, P. (2013). Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1):138–152. MR3024036

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media. MR2920735

Illian, J. B., Prosser, J. I., Baker, K. L., and Rangel-Castro, J. I. (2009). Functional principal component data analysis: A new method for analysing microbial community fingerprints. *Journal of microbiological methods*, 79(1):89–95.

Irwin, K. and Carter, P. (2013). Constraints on the evolution of function-valued traits: a study of growth in tribolium castaneum. *Journal of evolutionary biology*, 26(12):2633–2643.

Irwin, K. and Carter, P. (2014). Artificial selection on larval growth curves in tribolium: correlated responses and constraints. *Journal of evolutionary biology*, 27(10):2069–2079.

Ji, X. and Ruymgaart, F. H. (2008). Fréchet-differentiation of functions of operators with application to testing the equality of two covariance operators. In *Journal of Physics: Conference Series*, volume 124, page 012028. IOP Publishing.

Jiang, C.-R., Aston, J. A., and Wang, J.-L. (2009). Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage*, 47(1):184–193.

Kane, S. L., Garland Jr, T., and Carter, P. A. (2008). Basal metabolic rate of aged mice is affected by random genetic drift but not by selective breeding

for high early-age locomotor activity or chronic wheel access. *Physiological and Biochemical Zoology*, 81(3):288–300.

Kashlak, A. B., Aston, J. A., and Nickl, R. (2016). Inference on covariance operators via concentration inequalities: k-sample tests, classification, and clustering via rademacher complexities. *arXiv preprint arXiv:1604.06310*.

Kingsolver, J. G., Gomulkiewicz, R., and Carter, P. A. (2001). Variation, selection and evolution of function-valued traits. In *Microevolution Rate, Pattern, Process*, pages 87–104. Springer.

Koteja, P., Garland, T., Sax, J. K., Swallow, J. G., and Carter, P. A. (1999). Behaviour of house mice artificially selected for high levels of voluntary wheel running. *Animal behaviour*, 58(6):1307–1318.

Laukaitis, A. (2008). Functional data analysis for cash flow and transactions intensity continuous-time prediction using hilbert-valued autoregressive processes. *European Journal of Operational Research*, 185(3):1607–1614.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media. MR2135927

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660. MR0468056

Morgan, T. J., Garland, T., and Carter, P. A. (2003). Ontogenies in mice selected for high voluntary wheel-running activity. i. mean ontogenies. *Evolution*, 57(3):646–657.

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

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

Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons. MR1855501

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

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer. MR2168993

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2014). *fda: Functional Data Analysis*. R package version 2.4.4.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592. MR0455196

Solari, A., Salmaso, L., Pesarin, F., and Basso, D. (2009). *Permutation Tests for Stochastic Ordering and ANOVA*. Springer New York.

Stinchcombe, J. R., Kirkpatrick, M., and Function-valued Traits Working Group (2012). Genetics and evolution of function-valued traits: understanding environmentally responsive phenotypes. *Trends in Ecology & Evolution*, 27(11):637–647.

Swallow, J. G., Carter, P. A., and Garland Jr, T. (1998). Artificial selection for increased wheel-running behavior in house mice. *Behavior genetics*, 28(3):227–237.

Torres, J. M., Nieto, P. G., Alejano, L., and Reyes, A. (2011). Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of hazardous materials*, 186(1):144–149.

Tucker, J. D. (2016). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.6.1.

Tucker, J. D., Wu, W., and Srivastava, A. (2013). Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66. MR3063000

Viviani, R., Grön, G., and Spitzer, M. (2005). Functional principal component analysis of fmri data. *Human brain mapping*, 24(2):109–129.

West, R. M., Harris, K., Gilthorpe, M. S., Tolman, C., and Will, E. J. (2007). Functional data analysis applied to a randomized controlled clinical trial in hemodialysis patients describes the variability of patient responses in the control of renal anemia. *Journal of the American Society of Nephrology*, 18(8):2371–2376.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Wu, P.-S. and Müller, H.-G. (2010). Functional embedding for the classification of gene expression profiles. *Bioinformatics*, 26(4):509–517.