MOX–Report No. 23/2014

Covariance Based Unsupervised Classification in Functional Data Analysis

Ieva, F., Paganoni, A.M., Tarabelloni, N.

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

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

COVARIANCE BASED UNSUPERVISED CLASSIFICATION IN FUNCTIONAL DATA ANALYSIS

Francesca Ieva*, Anna Maria Paganoni♯ and Nicholas Tarabelloni♯

* Department of Mathematics "F. Enriques"
Università degli Studi di Milano
Via Cesare Saldini 50, 20133 Milano, Italy

francesca.ieva@unimi.it

♯ MOX– Modeling and Scientific Computing
Department of Mathematics
Politecnico di Milano
Via Bonardi 9, 20133 Milano, Italy

anna.paganoni@polimi.it,
nicholas.tarabelloni@.polimi.it

Keywords: Unsupervised classification, covariance operator, operator distance, shrinkage estimation, functional data analysis.

Abstract

In this paper we propose a new algorithm to perform unsupervised classification of multivariate and functional data when the difference between the two populations lies in their covariances, rather than in their means. The algorithm relies on a proper quantification of distance between the estimated covariance operators of the populations, and identifies as clusters those groups maximising the distance between their induced covariances. The naive implementation of such an algorithm is computationally forbidding, so we propose an heuristic formulation with a much lighter complexity and we study its convergence properties, along with its computational cost. We also propose to use an enhanced estimator for the estimation of discrete covariances of functional data, namely a linear shrinkage estimator, in order to improve the precision of the classification. We establish the effectiveness of our algorithm through applications to both synthetic data and a real dataset coming from a biomedical context, showing also how the use of shrinkage estimation may lead to substantially better results.

1 Introduction

The goal of performing unsupervised classification of data, in order to outline groups of observations based on some notion of similarity, has been of primary interest in applied statistics since ages. Literature is plenty of methods focusing their attention on the aggregation and separation of a sample into groups depending on similarities in locations of data (e.g. hierarchical clustering, k-means, PAM; see for instance [Har75]). To the best of our knowledge, almost nothing can be found on methods attaining the classification entirely on the basis of differences in the covariance structures of random models generating data. This target is not trivial, and less easy to translate into practice, since it calls for a proper quantification of distances between covariances of data. Nevertheless, it might happen to analyse data that are scarcely distinguishable in terms of locations, while show differences in their variability. Examples can be found in the biostatistics field where, for instance, the dichotomy between physiological and pathological features often shows an interesting pattern in change of variability.
In this paper we introduce a new method developed in order to reach this goal. We focus on

the specific case of a set of observations from two populations whose probability distributions have the same mean but differ in terms of covariances. Our aim is to perform classification in an unsupervised framework. The method we propose is general, therefore can be applied both to the traditional case of random vectors, and to the recently developed setting of functional data, arising as outcomes of infinite-dimensional stochastic processes (see, for instance, monographs [HK12], [RS05]). We will introduce the method according to the latter case.
In particular, we first introduce a suitable notion of distance between covariance operators, i.e. the functional generalisation of covariance matrices, which is the instrument we use to measure dissimilarities. Then we make use of such distance to search, among partitions of data, the one maximising the distance between the related covariances, under the assumption that, if the two populations can be distinguished from their covariances, this would be the most likely subdivision detecting the true groups.
A naive implementation of this algorithm, involving an exhaustive sampling strategy inside the set of subsets of data, would face a combinatorial complexity with respect to the number of observations, thus forbidding the analysis of datasets with common sizes. We therefore translate the method into a heuristic, greedy algorithm, with greatly reduced complexity, which can be efficiently implemented and effectively applied.
Due to its construction, our algorithm benefits from the accuracy of the estimation of covariances. Owing to the typically large dimensionality (compared to the number of data available) of discrete approximations of functional observations, covariance estimation through classical sample estimators may be non-optimal. To remedy this shortcoming, we propose to replace standard unbiased covariance estimator with a shrinkage estimator with enhanced accuracy properties (see, for instance, [LW03; LW04] and [SS05]). We show through experiments that this choice leads to a substantially improved classification.

The paper is organised as follows: in Section 2 we briefly recall some properties of covariance operators for functional data. In Section 3 we introduce the new classification method for data which differ in variance-covariance structures, we derive its heuristic formulation and describe the shrinkage strategy we used to enhance the estimation performances. In Section 4 we assess the classification performances through the application to both synthetic and real datasets. Discussion and conclusions are presented in Section 5.

2 Covariance operators for functional data

Whenever our data can be interpreted as finite-dimensional samplings of quantities that, instead, are intrinsically dependent on some continuous variable, such as time, we might be resorting to the model of functional data (see, for instance [HK12] and [RS05]). At its core is the assumption that data are sample measurements of trajectories of stochastic processes valued in suitable function spaces.
Hereby we recall the definition of covariance operator for functional data, along with its most important properties (for more details see, e.g., [Bos00]). Let $\mathcal{X}$ be a stochastic process taking values in $L^2(I;\mathbb{R})$, with $I \subset \mathbb{R}$ compact interval, having mean function $\mathbb{E}[\mathcal{X}] = \mu$ and such that $\mathbb{E}\|\mathcal{X}\|_0^2 < \infty$, where we denote by $\|\cdot\|_0$ the $L^2(I)$ norm induced by the scalar product $\langle\cdot,\cdot\rangle_0$. Without loss of generality we can assume $\mu = 0$ and define the following covariance operator:

$$\mathcal{C} \in \mathcal{L}\left(L^2(I);L^2(I)\right) \quad : \quad \langle y, \mathcal{C}(x)\rangle_0 = \mathbb{E}\left[\langle x,\mathcal{X}\rangle_0\langle y,\mathcal{X}\rangle_0\right], \qquad \forall x,y \in L^2(I), \tag{2.1}$$

$\mathcal{C}$ is a compact, self-adjoint, positive operator between $L^2(I)$ and $L^2(I)$. Therefore it can be decomposed into:

$$\mathcal{C} = \sum_{k=1}^{\infty} \lambda_k \; e_k \otimes e_k, \tag{2.2}$$

where $\{e_k\}_{k=1}^{\infty}$ is the sequence of orthonormal eigenfunctions, forming a basis of $L^2(I)$, and $\{\lambda_k\}_{k=1}^{\infty}$ is the sequence of eigenvalues. We assume eigenvalues to be sorted in decreasing order, so that:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq 0.$$

By expressing $\mathcal{X}$ with respect to the eigenfunctions basis, $\mathcal{X} = \sum_{k=1}^{\infty} \xi_k e_k$, it holds

$$\lambda_k = \langle e_k, \mathcal{C}(e_k) \rangle_0 = \mathbb{E}\left[\xi_k^2\right],$$

thus, the covariance operator is nuclear, meaning that

$$\mathbb{E}\|\mathcal{X}\|^2 = \sum_{k=1}^{\infty} \lambda_k = \sum_{k=1}^{\infty} |\lambda_k| < \infty.$$

$\mathcal{C}$ is also a Hilbert-Schmidt operator (see, for instance, [Bos00]), since it holds:

$$\sum_{k=1}^{\infty} \lambda_k^2 < \infty. \tag{2.3}$$

We equip the space of Hilbert-Schmidt operators with the Hilbert-Schmidt norm, defined as $\|\mathcal{U}\|_S^2 = \sum_{k=1}^{\infty} \lambda_k^2$, where $\{\lambda_k\}_{k=1}^{\infty}$ are the eigenvalues of $\mathcal{U}$. This is induced by the following scalar product:

$$\langle \mathcal{U}, \mathcal{V} \rangle_S = \sqrt{\mathrm{Tr}\left(\mathcal{U} - \mathcal{V}\right)\left(\mathcal{U} - \mathcal{V}\right)^*}, \tag{2.4}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace operator, and $\mathcal{U}^*$ is the Hilbertian adjoint of $\mathcal{U}$, i.e.

$$\langle \mathcal{U}(x),\ y \rangle_0 = \langle x,\ \mathcal{U}^*(y) \rangle_0 \quad \forall x, y \in L^2(I).$$

The space of Hilbert-Schmidt operators on $L^2(I)$, endowed with the scalar product (2.4) and the associated norm, becomes a separable Hilbert space itself.
Within this theoretic framework, a natural definition of dissimilarity between Hilbert-Schmidt operators (among which are covariance operators) may be the Hilbert-Schmidt distance:

$$d\left(\mathcal{U}, \mathcal{V}\right) = \|\mathcal{U} - \mathcal{V}\|_S = \sum_{k=1}^{\infty} \eta_k^2. \tag{2.5}$$

where $\{\eta_k\}_{k=1}^{\infty}$ is the sequence of eigenvalues of $\mathcal{U} - \mathcal{V}$.

3 Covariance based classification

We face now the problem of classifying observations belonging to two different functional populations. Let $\mathcal{X}$ be a stochastic process on $L^2(I)$ generated by the law $P_{\mathcal{X}}$, and $\mathcal{Y}$ a likewise process generated according to $P_{\mathcal{Y}}$. We imagine to have a set of N data in some dataset D generated according to these two populations in a balanced proportion, i.e:

$$\begin{aligned} [D]_i &= X_i \sim P_{\mathcal{X}}, \qquad \forall i = 1, \ldots, K \\ [D]_j &= Y_j \sim P_{\mathcal{Y}}, \qquad \forall j = K+1, \ldots, N \end{aligned}$$

with $K = N/2$.
We introduce the following quantities:

$$\mu_1 := \mathbb{E}[X_i], \quad \mathcal{C}_1 := \mathbb{E}[X_i \otimes X_i], \quad \forall\, i = 1, \dots, K,$$
$$\mu_2 := \mathbb{E}[Y_j], \quad \mathcal{C}_2 := \mathbb{E}[Y_j \otimes Y_j], \quad \forall\, j = K+1, \dots, N.$$

Let us consider the vector of indexes of units constituting the two populations in D:

$$I^{(0)} = \Big(\overbrace{1, 2, \dots, K}^{I_1^{(0)}}, \overbrace{K+1, \dots, N}^{I_2^{(0)}} \Big). \tag{3.1}$$

which is unique, provided we don't distinguish among permutations of sub-intervals $I_1^{(0)}$ and $I_2^{(0)}$.
In the following we shall consider recombinations of these indexes:

$$I^{(i)} = \mathbf{R}_i I^{(0)} = \left[I_1^{(i)};\ I_2^{(i)} \right], \quad i \in \{1, \dots, N_C\}, \tag{3.2}$$

where $I^{(i)}$ denotes the i-th combination out of $N_C = \binom{N}{K}$, however enumerated, and $\mathbf{R}_i$ is the recombination matrix realising it.
The sample estimators of means and covariance operators induced by this subdivision are indicated, respectively, with $\widehat{\mu}_1^{(i)}$, $\widehat{\mu}_2^{(i)}$ and $\widehat{\mathcal{C}}_1^{(i)}$, $\widehat{\mathcal{C}}_2^{(i)}$. We point out that, when $i = 0$, we recover the estimators of μ_1, μ_2 and $\mathcal{C}_1$ and $\mathcal{C}_2$. For this reason we rename the latter quantities as $\mu_1^{(0)}$, $\mu_2^{(0)}$, and $\mathcal{C}_1^{(0)}$, $\mathcal{C}_2^{(0)}$.

Our classification method is based on the following, crucial assumption:

Hypothesis (H). *We assume that observations drawn from families $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ constituting the dataset D may be distinguished on the basis of their covariances, but not from their means, i.e.:*

$$\begin{aligned} \mathcal{C}_1^{(0)} \neq \mathcal{C}_2^{(0)}, \quad &\Longrightarrow \quad d(\mathcal{C}_1^{(0)}, \mathcal{C}_2^{(0)}) \gg 0, \\ \mu_1^{(0)} = \mu_2^{(0)}, \quad &\Longrightarrow \quad \|\mu_1^{(0)} - \mu_2^{(0)}\|_0 = 0. \end{aligned} \tag{H}$$

As a consequence of this hypothesis we conveniently center data and assume they have zero means.

Let us pretend the true covariance operators $\mathcal{C}_1^{(0)}$ and $\mathcal{C}_2^{(0)}$ are known. Then, for the generic i-th recombination of data, we define the quantity:

$$\rho_i = \frac{\mathrm{Card}\left(I_1^{(i)} \cap I_1^{(0)} \right)}{K} \in [0, 1], \quad i \in \{1, \dots, N_C\}, \tag{3.3}$$

that measures the proportion of indexes of data from the first population that are fixed under the i-th recombination.
Then, if we consider the sample estimators of covariance operators for the generic i-th recombination:

$$\widehat{\mathcal{C}}_1^{(i)} = \frac{1}{K} \left(\sum_{l \in L_{(1,1)}} X_l \otimes X_l + \sum_{l \in L_{(1,2)}} Y_l \otimes Y_l \right), \quad L_{(1,1)} = I_1^{(i)} \cap I_1^{(0)},\ L_{(1,2)} = I_1^{(i)} \cap I_2^{(0)}$$
$$\widehat{\mathcal{C}}_2^{(i)} = \frac{1}{K} \left(\sum_{l \in L_{(2,1)}} X_l \otimes X_l + \sum_{l \in L_{(2,2)}} Y_l \otimes Y_l \right), \quad L_{(2,1)} = I_2^{(i)} \cap I_1^{(0)},\ L_{(2,2)} = I_2^{(i)} \cap I_2^{(0)}$$

by exploiting (3.3), it holds:

$$\mathcal{C}_1^{(i)} := \mathbb{E}\left[\widehat{\mathcal{C}}_1^{(i)}\right] = \rho_i\, \mathcal{C}_1^{(0)} + (1-\rho_i)\, \mathcal{C}_2^{(0)},$$
$$\mathcal{C}_2^{(i)} := \mathbb{E}\left[\widehat{\mathcal{C}}_2^{(i)}\right] = (1-\rho_i)\, \mathcal{C}_1^{(0)} + \rho_i\, \mathcal{C}_2^{(0)},$$

then:

$$d\left(\mathcal{C}_1^{(i)}, \mathcal{C}_2^{(i)}\right) = (2\rho_i - 1)^2\; d\left(\mathcal{C}_1^{(0)}, \mathcal{C}_2^{(0)}\right). \tag{3.4}$$

The last identity confirms the natural idea that, if hypothesis (**H**) is true, the distance between the induced covariance operators reaches its maximum at the labelling expressing the true subdivision of data.
Nevertheless, in real applications the *true* indexing $I^{(0)} = [I_1^{(0)}; I_2^{(0)}]$ is typically unknown and an estimate of it is required. If (**H**) holds true, by looking at (3.4), a natural way to recover the true indexing may be to find the recombination of data in two groups maximising the distance between the induced covariance operators, i.e. to solve the optimization problem:

$$[I_1^*; I_2^*] = \arg\max_{i \in J} \left\{ d\left(\mathcal{C}_1^{(i)}, \mathcal{C}_2^{(i)}\right) \right\}, \quad J = \{1, 2, \ldots, N_C\}. \tag{$\mathbf{P}$}$$

Identity (3.4) assures that either $I_1^* = I_1^{(0)}$ and $I_2^* = I_2^{(0)}$, or $I_1^* = I_2^{(0)}$ and $I_2^* = I_1^{(0)}$. The double solution is due to the symmetry of (3.4), yet the groups represent the same partition of data, therefore in the following we will not distinguish between them.
Practically, only approximate estimates of $\mathcal{C}_1^{(i)}$ and $\mathcal{C}_2^{(i)}$ are available, thus we must recast problem (**P**) into:

$$\left[\widehat{I}_1^*; \widehat{I}_2^*\right] = \arg\max_{i \in J} \left\{ d\left(\widehat{\mathcal{C}}_1^{(i)}, \widehat{\mathcal{C}}_2^{(i)}\right) \right\}, \quad J = \{1, 2, \ldots, N_C\}, \tag{$\widehat{\mathbf{P}}$}$$

The method we propose coincides with finding a solution to problem ($\widehat{\mathbf{P}}$).
In general $\widehat{I}_1^*$ and $\widehat{I}_2^*$ may differ from $I_1^{(0)}$ and $I_2^{(0)}$, since they are determined based on estimates of covariance operators. Indeed, provided that the chosen distance is capable of emphasizing the actual differences between covariances of the two populations, results could be improved by enhancing the accuracy of estimators. In Subsection 3.2 we will address the former issue.

3.1 Greedy formulation

In order to solve problem ($\widehat{\mathbf{P}}$), it would be required to test each of the N_C recombinations of indexes in order to find the desired pair $\widehat{I}_1^*$ and $\widehat{I}_2^*$. Of course, the number of tests to be performed, $N_C = \binom{N}{K}$, with $K = N/2$, undergoes a combinatorially-fast growth, as N increases. Thus, unless we have only a small number of observations in our dataset, the naive approach of performing an exhaustive search in the set of recombinations is not feasible. This calls for a proper complexity-reduction strategy, aimed at restraining the complexity and enabling the application of our classification method also to datasets with a common size.

Max-Swap algorithm: We propose to rephrase problem ($\widehat{\mathbf{P}}$) into a greedy algorithm, with a greatly reduced complexity. The driving idea is to interpret $d(\widehat{\mathcal{C}}_1^{(i)}, \widehat{\mathcal{C}}_2^{(i)})$ as an objective function of i, and, starting from an initial guess (I_1^0, I_2^0), to iteratively increase it by allowing exchanges of units between the two groups. The exchange of data must preserve the total number of units

inside each group, so each group discards and receives an equal number of units, say up to J per group.
We propose to choose the swapping units in such a way that the distance between the estimated covariance operators at the next step be strictly higher than the previous one and, heuristically, the highest possible. Convergence is reached when no further swap can increase that distance.

When searching the best swap of size up to J, we must explore a number of permutations of the current groups equal to $\sum_{i=1}^{J} \binom{K}{i}^2$. Therefore it is evident that J affects both the computational effort and the robustness of our algorithm: the lower is J, the less permutations we have to search among to find the optimal swap; the greater is J, the more likely we are to detect and exchange at once a block of truly extraneous units. We point out that, for $J = K$ we recover the original complexity of solving problem ($\widehat{\mathbf{P}}$) in just one step, since it holds:

$$\binom{N}{K} = \sum_{i=1}^{K} \binom{K}{i}^2.$$

We propose to set $J = 1$, in order to save computations, and to choose the units to be exchanged by exploring the K^2 swaps of one unit from the first group with another unit of the second group. Then we select the one yielding the maximum increment in the distance. The complete formulation of our Max-Swap algorithm is summarised in Algorithm 1. In the following we will denote the estimated set of indexes at step k of algorithm with the apex k without brackets, (I_1^k, I_2^k).

Convergence We turn now to the study of the convergence of our proposed algorithm. With reference to the notation of Algorithm 1, it is easy to prove that

Proposition 1. *The monotonicity constraint:*

$$(\Delta d)^k > 0 \quad \forall k \geq 1, \tag{3.5}$$

ensures that convergence always happens, at least to a local maximum of $d(\widehat{\mathcal{C}}_1^{(i)}, \widehat{\mathcal{C}}_2^{(i)})$.

Proof. As a simple consequence of (3.5), the list of intermediate indexings:

$$\left(I_1^0, I_2^0\right), \left(I_1^1, I_2^1\right), \ldots, \left(I_1^k, I_2^k\right), \ldots$$

does not have cycles (a contiguous sub-sequence with equal extrema). In fact, let there be a cycle of minimal period L starting at iteration k_0, then it should hold:

$$0 = d^{k_0+L} - d^{k_0} = \sum_{j=1}^{L} (\Delta d)^{j+k_0} > 0,$$

which is a contradiction. Thus each element in the list is unique and contained in the set of all the possible recombinations of data:

$$(I_1^{(0)}, I_2^{(0)}), (I_1^{(1)}, I_2^{(1)}), \ldots, (I_1^{(N_C)}, I_2^{(N_C)})$$

which has a finite number of elements. Therefore the algorithm, however initialised, converges. □

Now that convergence has been established, we can formulate the following proposition:

Algorithm 1: Max-Swap algorithm

Input: Initial guess: $\left(I_1^0, I_2^0\right)$

Output: Estimated indexing $\left(\widehat{I}_1^{**}, \widehat{I}_2^{**}\right)$

Compute $\left(\widehat{\mathcal{C}}_1^0, \widehat{\mathcal{C}}_2^0\right)$ induced by $\left(I_1^0, I_2^0\right)$;

$d^0 = d\left(\widehat{\mathcal{C}}_1^0, \widehat{\mathcal{C}}_2^0\right)$;

$k = 1$;

$(\Delta d)^k = 1$;

while $(\Delta d)^k > 0$ **do**

  for $s \in 1, \dots, K$ **do**

    for $t \in 1, \dots, K$ **do**

      Swap in first group: $\tilde{I}_1 = \bigcup_{p \neq s} I_1^{k-1}(p) \cup I_2^{k-1}(t)$;

      Swap in second group: $\tilde{I}_2 = \bigcup_{q \neq t} I_2^{k-1}(q) \cup I_1^{k-1}(s)$;

      Compute $\left(\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2\right)$ induced by $(\tilde{I}_1, \tilde{I}_2)$;

      $D_{s,t} = d\left(\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2\right)$;

  $(s^*, t^*) = \arg\max_{s,t} D_{s,t}$;

  $d^k = D_{s^*,t^*}$;

  $(\Delta d)^k = d^k - d^{k-1}$;

  $I_1^k = \bigcup_{p \neq s^*} I_1^{k-1}(p) \cup I_2^{k-1}(t^*)$;

  $I_2^k = \bigcup_{q \neq t^*} I_2^{k-1}(q) \cup I_1^{k-1}(s^*)$;

  $k = k + 1$;

$d^{**} = d^{k-1}$;

$I_1^{**} = I_1^{k-1}$;

$I_2^{**} = I_2^{k-1}$;

Figure 1: Pseudo-code of the Max-Swap algorithm.

Proposition 2. *When estimates of covariance operators, $\widehat{\mathcal{C}}_1$ and $\widehat{\mathcal{C}}_2$, are exact, i.e., when $\widehat{\mathcal{C}}_1 = \mathcal{C}_1$ and $\widehat{\mathcal{C}}_2 = \mathcal{C}_2$, the greedy algorithm converges to the exact solution (I_1^*, I_2^*) of problem (**P**) in at most $K/2$ steps.*

Proof. This is a consequence of Proposition 1 and the convexity of the objective function $d(\mathcal{C}_1^{(i)}, \mathcal{C}_2^{(i)})$ showed in (3.4), making it impossible that local maxima exist. □

A direct consequence of Proposition 1 is that in general $(I_1^{**}, I_2^{**}) \neq (\widehat{I}_1^*, \widehat{I}_2^*)$, since the algorithm may converge only to a local maximizer of the covariances' distance. This is a well-known drawback affecting greedy methods for optimization problems based on local-search patterns and the development of possible remedies is a very active research field in algorithmics and optimization disciplines.
However, Proposition 2 assures that the algorithm converges to the exact solution, provided that we have a thorough knowledge of covariance operators. Therefore, the possible non-convexity of the objective function $d(\mathcal{C}_1^{(i)}, \mathcal{C}_2^{(i)})$, being the reason why local maxima exist, would be a direct consequence of the finite precision in estimating covariances.
Under this light, we can rephrase the problem of enhancing our method from finding an

algorithmics-like remedy to the aforementioned drawback, to the study of accuracy properties of covariance estimators. This will be the focus of Subsection 3.2. Another possibility to reduce the risk of selecting only local maximisers, which we don't investigate in the following, would be to assess the stability of the maximiser found when running a standard Max-Swap algorithm by running a general version of Max-Swap with $J > 1$.

Rate of convergence and complexity Max-Swap algorithm increases the objective function starting from an initial guess, (I_1^0, I_2^0). In general, we have no prior knowledge on the distribution of true groups among the dataset, then we should choose the initial guess at random, and in particular by drawing from the set of indexes without replacement, assigning equal probability to each outcome. This strategy causes the quantity ρ in (3.3) to follow a scaled hypergeometric distribution:

$$\rho K \sim \text{Hyper}(N, K, K),$$

i.e:

$$\mathbb{P}(\rho = h) = \frac{\binom{K}{Kh}\binom{K}{K-Kh}}{\binom{N}{K}}, \quad \mathbb{E}[\rho K] = \frac{K}{2},$$

with typically large values of N and $K = N/2$. Owing to the fast decay of hypergeometric mass function away from its mean, we can presume that the random initial draw will cause ρ to be most likely in some neighbourhood of $K/2$ (we imagine K to be even).
Let us assume that hypotheses of Proposition 2 are true, then if we consider a mixing proportion ρ resulting from the initial guess, the number of iterations to convergence is $N_{\text{it}} = \rho K$, corresponding to ρK consecutive and correct swaps.
Then, by initialising at random the algorithm:

$$\mathbb{E}[N_{\text{it}}] = \frac{K}{2}.$$

If we summarise the complexity of solving problem ($\widehat{\mathbf{P}}$) with the number of combinations N_{comb} processed to recover the estimated groups (I_1^*, I_2^*), and we compare our proposed method (MS) with the naive, exhaustive strategy (N), we have:

$$\mathbb{E}\left[N_{\text{comb}}^{MS}\right] = K^2\,\mathbb{E}[N_{\text{it}}] = \frac{K^3}{2}, \qquad N_{\text{comb}}^{\text{N}} = \binom{N}{K},$$

where the multiplicative factor K^2 in $\mathbb{E}\left[N_{\text{comb}}^{MS}\right]$ accounts for the number of combinations searched for to find the best swap at each step of Max-Swap algorithm.
This shows how Max-Swap algorithm entails a far lower complexity than the brute force approach.

3.2 Shrinkage estimation of covariance

In this subsection we consider the problem of improving the estimation of covariance operators, so that classification be more accurate.

Let us consider a generic family of functional data $\mathcal{X} \sim P_{\mathcal{X}}$, such that $\mathbb{E}[\mathcal{X}] = 0$, $\mathbb{E}\|\mathcal{X}\|_0^2 < \infty$. We denote its covariance with $\mathcal{C}$ and we imagine to estimate it with $\widehat{\mathcal{C}}$. Our purpose is to find the best possible approximation, or saying it otherwise, being:

$$\text{MSE}_S(\widehat{\mathcal{C}}) := \mathbb{E}\,\|\widehat{\mathcal{C}} - \mathcal{C}\|_S^2$$

our measure of the estimation error of $\widehat{\mathcal{C}}$, to solve the following estimation problem (**E**):

$$\widehat{\mathcal{C}}^* = \arg\min_{\widehat{\mathcal{C}}} \mathrm{MSE}_S(\widehat{\mathcal{C}}) = \arg\min_{\widehat{\mathcal{C}}} \mathbb{E}\|\widehat{\mathcal{C}} - \mathcal{C}\|_S^2, \tag{**E**}$$

where the minimum is sought among all possible estimators $\widehat{\mathcal{C}}$ of $\mathcal{C}$. We point out that in MSE_S we use our selected distance to measure the discrepancy of estimation.
Of course, from a practical viewpoint, only a finite dimensional estimation of $\mathcal{C}$ can be attained, given data. In addition, functional data are often available from sources as discrete measurements of a signal over some one dimensional grid. Let us indicate by X_i the i-th (out of N) sample realisation of process $\mathcal{X}$, i.e:

$$X_i = \{X_i(t_j)\}_{j=1}^P, \qquad I^h = [t_1, \dots, t_P], \quad t_{i+1} - t_i = h > 0, \quad i = 1, \dots, N. \tag{3.6}$$

where, for the sake of simplicity, we have imagined the grid I^h to be regularly spaced (although this is not mandatory). A crucial point when analysing functional data is to reconstruct functions from scattered measurements X_i, which requires the use of some proper smoothing technique. Furthermore, the so called *phase variability* of reconstructed signals, involving the dispersion of features along the grid axis, can be separated from *amplitude variability*, appearing as the dispersion of magnitudes of values of X_i. This process is known as registration (see, for instance, [RS05]). Once data have been smoothed and registered, they can be re-evalued onto another one dimensional grid. To save notation we will assume that discrete representations in (3.6) have already been preprocessed.

It is clear that, within this habit, covariance estimators of $\mathcal{C}$ are discrete, matrix-type approximations constructed starting from pointwise observations X_i. For instance, standard sample covariance estimator for zero-mean data is:

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} X_i \, X_i^T. \tag{3.7}$$

If we denote the true, discrete covariance structure related to each X_i by $\mathbf{C}$ the discrete version of problem (**E**) is:

$$\mathbf{C}^* = \arg\min_{\widehat{\mathbf{C}}} \mathbb{E}\|\widehat{\mathbf{C}} - \mathbf{C}\|_F^2, \tag{$\widehat{\mathbf{E}}$}$$

where the minimum is sought inside the set of symmetric and positively defined matrix-type estimators of dimension P. We point out that the subscript F in ($\widehat{\mathbf{E}}$) indicates the Frobenius norm, that is the finite-dimensional counterpart of the Hilbert-Schmidt norm for operators.

When the sample size N is low compared to the number of features P, sample covariance may loose in accuracy, meaning that the actual estimate might be quite distant from the true covariance $\mathbf{C}$ (this can be seen as a consequence of the so-called *Stein's phenomenon*, [Ste56]).
A typical remedy to the poor performances of sample covariance, often used in the setting of *Large P - Small N* problems, is to replace it with a biased, shrinkage estimator. Other solutions might be jacknife or bootstrap, but their computational cost renders them practically useless in our classification algorithm, which requires to repeatedly estimate covariance matrices. Shrinkage estimation has been explicitly applied to the context of large covariance matrices in [LW03; LW04] and [SS05], turning out in a sufficiently lightweight procedure. In those works, authors start from problem ($\widehat{\mathbf{E}}$) and build an estimator that is asymptotically more accurate and better conditioned

than sample covariance. We follow [LW04] and consider the class of linear shrinkage estimators of the form:

$$\widehat{\mathbf{C}} = \mu\gamma\,\mathbf{I} + (1-\gamma)\mathbf{S}, \tag{3.8}$$

where $\mathbf{I}$ is the $P \times P$ identity matrix and $\gamma \in [0,1]$, $\mu \in \mathbb{R}^+$ and $\mathbf{S}$ is the sample covariance estimator. Obviously, the class contains the sample covariance estimator itself. Then ($\widehat{\mathbf{E}}$) is solved with respect to the optimal values of μ and γ:

$$(\mu^*, \gamma^*) = \arg\min_{\mu,\gamma} \; \mathbb{E}\frac{\|\mathbf{C} - \mu\gamma\mathbf{I} - (1-\gamma)\mathbf{S}\|_F^2}{P}. \tag{3.9}$$

If we introduce the quantities:

$$\alpha^2 = \frac{\|\mathbf{C} - \mu\mathbf{I}\|_F^2}{P}, \qquad \beta^2 = \frac{\mathbb{E}\,\|\mathbf{S} - \mathbf{C}\|_F^2}{P}, \qquad \delta^2 = \frac{\mathbb{E}\,\|\mathbf{S} - \mu\mathbf{I}\|_F^2}{P}, \tag{3.10}$$

and note that these are subjected to $\alpha^2 + \beta^2 = \delta^2$, we can perform the explicit minimization in equation (3.9). The expressions of μ^* and γ^* are:

$$\mu^* = \frac{\langle \mathbf{C}, \mathbf{I} \rangle_F}{P} = \frac{\mathrm{Tr}\,(\mathbf{C})}{P}, \qquad \gamma^* = \frac{\beta^2}{\delta^2}, \tag{3.11}$$

where we have used $\mu = \mu^*$ in the computation of β and δ. The desired shrinkage estimator becomes:

$$\mathbf{S}^* = \mu^*\frac{\beta^2}{\delta^2}\mathbf{I} + \frac{\alpha^2}{\delta^2}\mathbf{S}. \tag{3.12}$$

Of course, estimator (3.12) depends on the unknown exact covariance matrix $\mathbf{C}$, even though only through four scalar functions. In [LW04] authors solve this problem by proposing the following estimators for α, β and δ:

$$\widehat{\mu^*} = \frac{\mathrm{Tr}\,(\mathbf{S})}{P}, \qquad \widehat{\delta^2} = \frac{\|\mathbf{S} - \widehat{\mu^*}\mathbf{I}\|_F^2}{P}, \qquad \widehat{\beta^2} = \min\left(\widehat{\delta^2}; \frac{1}{N^2}\sum_{k=1}^{N}\frac{\|X_k\, X_k^T - \mathbf{S}\|_F^2}{P}\right) \tag{3.13}$$

and $\widehat{\alpha^2} = \widehat{\delta^2} - \widehat{\beta^2}$.
Then, the actual shrinkage estimator is:

$$\widehat{\mathbf{S}}^* = \widehat{\mu^*}\frac{\widehat{\beta^2}}{\widehat{\delta^2}}\mathbf{I} + \frac{\widehat{\alpha^2}}{\widehat{\delta^2}}\mathbf{S} \tag{3.14}$$

In [LW04] they show how estimates (3.13) are consistent, in the sense that they converge to the exact values in quadratic mean, under the general asymptotic limits of P and N, i.e. when both P and N are allowed to go to infinity but there exists a $c \in \mathbb{R}$ independent on N such that $P/N < c$ (see [LW04] and references therein for theoretical details on general asymptotics). Moreover, estimator $\widehat{\mathbf{S}}^*$ is an asymptotically optimal linear shrinkage estimator for covariance matrix $\mathbf{C}$ with respect to quadratic loss.
Besides its asymptotic properties, extensive use in applications shows that the accuracy gain resulting from $\widehat{\mathbf{S}}^*$ in terms of decrease in MSE is substantial also in many finite sample cases, and that standard covariance is almost always matched and often outperformed by $\widehat{\mathbf{S}}^*$.

4 Case studies

In this section we provide three simulations involving our proposed classification method. In Subsection 4.1 we show a first example, regarding standard bivariate data, in order to give a clear geometric idea of classification based on covariance structures. In Subsection 4.2 we show an application to synthetic functional data. In these former two examples the true subdivision of samples is known, so the goodness of the classification arising from Max-Swap algorithm is assessed against the true identities of data. In Subsection 4.3, instead, we apply the classification algorithm in a truly unsupervised way on real functional data expressing the concentration of deoxygenated hemoglobin measured in human subjects' brains.

4.1 Multivariate data

This test is meant to provide a first, visual example of the features of the classification arising from using Max-Swap algorithm. In order to ease the geometrical interpretation, we chose to focus on two bivariate datasets, composed of simulated data with a-priori designed covariances. Indeed, by representing bi-dimensional data we are able to support our considerations with a clear graphical counterpart.
We exploit two reference datasets having the same means but different variance-covariance structures. In particular, a generic classification based on locations run on these data is meant to fail. The first set of data, hereafter *hourglass* data, has covariances whose difference lies in the directions along which variability expresses. We generated it according to the following laws:

$$\begin{aligned} X &= \big(\, \rho_x \cos\theta_x,\ \rho_x \sin\theta_x \big), \quad \rho_x \sim \mathcal{U}\left[-1, 1\right], \quad \theta_x \sim \mathcal{U}\left[\tfrac{\pi}{12}, \tfrac{5\pi}{12}\right], \\ Y &= \big(\, \rho_y \cos\theta_y,\ \rho_y \sin\theta_y \big), \quad \rho_y \sim \mathcal{U}\left[-1, 1\right], \quad \theta_y \sim \mathcal{U}\left[\tfrac{7\pi}{12}, \tfrac{11\pi}{12}\right], \end{aligned}$$

where the four random variables, $\rho_x, \rho_y, \theta_x, \theta_y$ are independent. Simple calculations reveal that $\mathbb{E}\left[X\right] = 0$ and $\mathbb{E}\left[Y\right] = 0$, while covariances are:

$$\mathbf{C}_x = \begin{pmatrix} 1/6 & \sqrt{3}/4\pi \\ \sqrt{3}/4\pi & 1/6 \end{pmatrix}, \qquad \mathbf{C}_y = \begin{pmatrix} 1/6 & -\sqrt{3}/4\pi \\ -\sqrt{3}/4\pi & 1/6 \end{pmatrix}.$$

Note that X and Y differ only in their covariances. Moreover, since only off-diagonal terms of $\mathbf{C}_1$ and $\mathbf{C}_2$ are different (and indeed opposed), the two families have the same kind of variability, only expressed along orthogonal directions in the plane. We generate a dataset D of $N = 400$ data, according to the previous laws, made up of $K = 200$ samples from X and $K = 200$ samples from Y, which are displayed in Fig. 2.
We considered also another dataset, referred to as *bull's eye*, whose features are somehow complementary to the ones of *hourglass*, since variabilities of *bull's eye* sub-populations express along the same directions, though with different magnitudes. In particular, we considered the following laws:

$$\begin{aligned} X &= \big(\, \rho_x \cos\theta_x,\ \rho_x \sin\theta_x \big), \quad \rho_x \sim \mathcal{U}\left[0, \tfrac{1}{2}\right], \quad \theta_x \sim \mathcal{U}\left[0, 2\pi\right], \\ Y &= \big(\, \rho_y \cos\theta_y,\ \rho_y \sin\theta_y \big), \quad \rho_y \sim \mathcal{U}\left[2, \tfrac{5}{2}\right], \quad \theta_y \sim \mathcal{U}\left[0, 2\pi\right], \end{aligned}$$

where, still, the four random variables $\rho_x, \rho_y, \theta_x, \theta_y$ are independent. This leads to covariances:

$$\mathbf{C}_x = \frac{1}{24}\mathbf{I}, \qquad \mathbf{C}_y = \frac{61}{24}\mathbf{I},$$

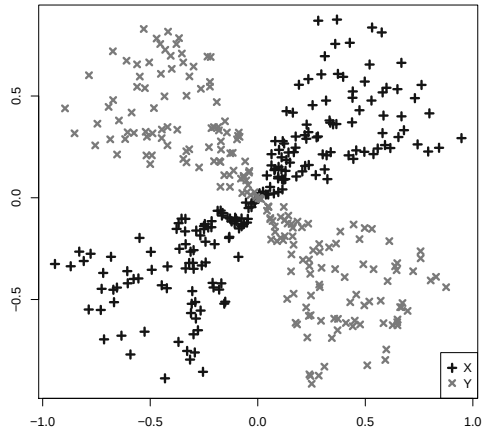

Figure 2: Hourglass dataset used in the first multivariate experiment, collecting $N = 400$ points subdivided into family X ($K = 200$ points, marked by $+$) and family Y ($K = 200$ points, marked by $\times$).

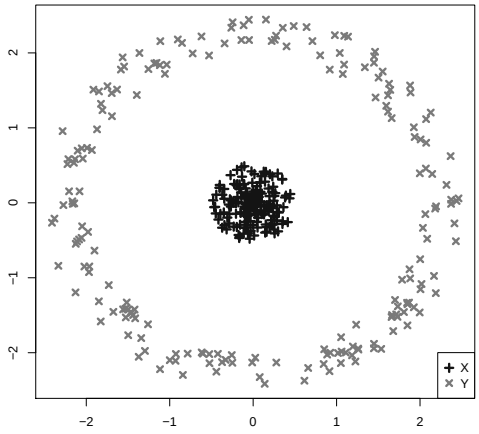

Figure 3: Bull's eye dataset used in the second multivariate experiment, collecting $N = 400$ points subdivided into family X ($K = 200$ points, marked by $+$) and family Y ($K = 200$ points, marked by $\times$).

	Hourglass		**Bull's eye**	
	X (true)	**Y (true)**	**X (true)**	**Y (true)**
X (estimated)	197	3	200	0
Y (estimated)	3	197	0	200
Misclassification	1.5%		0%	

Table 1: Misclassification table for the experiment on *hourglass* and *bull's eye* datasets.

that clearly differ only in terms of their variability's magnitude. *Bull's eye* dataset is generated according to these laws for an overall cardinality of $N = 400$ data subdivided in two groups of size $K = 200$ each, and it is shown in Fig. 3.
We point out that, in order to improve the robustness of results with respect to the chance of selecting only a local maximiser of the distance between variance-covariance structures, Max-Swap algorithm has been run for 10 times, keeping the result for which the objective function was highest. We summarize the results of both the experiments in Tab. 1. Since the number of data in each sub-population, K, is high with respect to their dimensionality, $P = 2$, we used Max-Swap algorithm in combination with the standard sample estimator of covariance, $\mathbf{S}$.

4.2 Synthetic functional data

In this subsection we apply our classification algorithm to functional data. We use a dataset composed of two populations of functions, $\mathcal{X}$ and $\mathcal{Y}$, with null means and covariance operators:

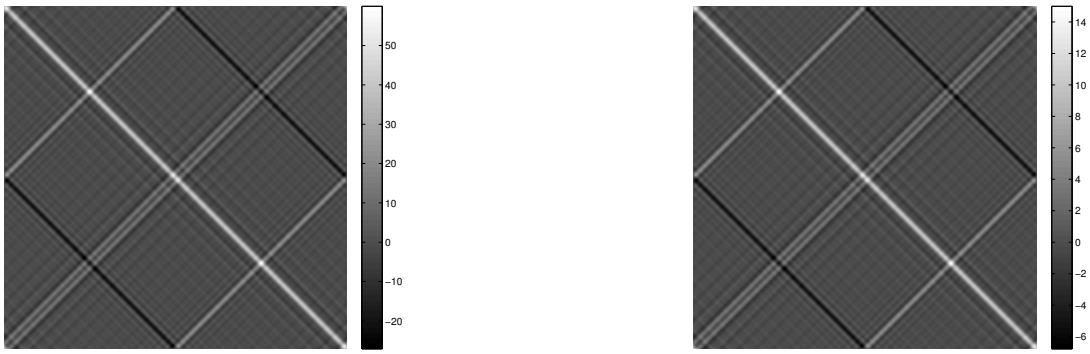

Figure 4: Contour plot of covariances $\mathcal{C}_x$ and $\mathcal{C}_y$ for the experiment with synthetic functional data. The different scales of contour plots show that the difference between the variance-covariance structures is only in magnitude.

$$\mathcal{C}_x = \sum_{i=1}^{L} \sigma_i \; e_i \otimes e_i, \qquad \mathcal{C}_y = \sum_{i=1}^{L} \eta_i \; e_i \otimes e_i, \tag{4.1}$$

where $\{e_i\}_{i=1}^{L}$ are the first L elements of the orthonormal Fourier basis on the interval I, save for the constant, i.e.:

$$e_{2k-1} = \sqrt{\frac{I}{2}} \sin\left(\frac{2k\pi x}{I}\right), \quad e_{2k} = \sqrt{\frac{I}{2}} \cos\left(\frac{2k\pi x}{I}\right), \qquad k = 1, \ldots, L/2$$

and the eigenvalues are:

$$\sigma_i = 1, \qquad \eta_i = \frac{\sigma_i}{4}, \quad \forall \; i = 1, \ldots, L, \tag{4.2}$$

In what follows we considered $I = [0, 1]$ and $L = 40$. A visual representation of the related covariance functions is in Fig. 4. It is clear from (4.1) and from Fig. 4 that covariances $\mathcal{C}_x$ and $\mathcal{C}_y$ are different only in terms of variability's magnitudes, while their eigenfunctions are the same.
We generated several sets of the following synthetic families of gaussian functional data having covariances like in (4.1):

$$X_i = \sum_{j=0}^{L} \xi_{ij} \sqrt{\sigma_j} \; e_j, \qquad Y_i = \sum_{j=0}^{L} \zeta_{ij} \sqrt{\eta_j} \; e_j, \qquad i = 1, \ldots, K, \tag{4.3}$$

where $\xi_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$, and are independent from $\zeta_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Each synthetic functional unit has been evaluated on a grid of $P = 100$ points, evenly spaced on I. The different sets have been generated choosing $K \in \{20, 25, 40, 35, 40, 45, 50\}$, corresponding to total cardinalities of $N \in \{40, 50, 60, 70, 80, 90, 100\}$.
We applied our classification algorithm to each synthetic dataset. The different values of K (i.e. of N) allow to study the performances of classification as the sample size increases. This is of interest since our method relies on the estimation of covariance matrices, thus we expect that when the number of data increases the classification tends to improve.
We used Max-Swap algorithm both with the standard sample covariance estimator, $\mathbf{S}$, and with

		Sample covariance (S)		Shrinkage estimator ($\widehat{\mathbf{S}}^*$)	
N	K	**Miscl.** $(\mathbf{1},\mathbf{2})$	**Err.**	**Miscl.** $(\mathbf{1},\mathbf{2})$	**Err.**
40	20	6 $(3,3)$	15%	0 $(0,0)$	0%
50	25	6 $(3,3)$	12%	0 $(0,0)$	0%
60	30	4 $(2,2)$	6.67%	0 $(0,0)$	0%
70	35	4 $(2,2)$	5.71%	0 $(0,0)$	0%
80	40	4 $(2,2)$	5%	0 $(0,0)$	0%
90	45	4 $(2,2)$	4.4%	0 $(0,0)$	0%
100	50	0 $(0,0)$	0%	0 $(0,0)$	0%

Table 2: Classification results for the application with synthetic functional data.

the shrinkage covariance estimator $\widehat{\mathbf{S}}^*$.

We report the results of the classification procedure in Tab. 2. Similarly to the case of multivariate synthetic data of Subsection 4.1, each of them is related to the one trial in a pool of 10 for which the distance between covariances was highest; this was done in order to take account of the heuristic component of the algorithm.
Results undoubtedly highlight that covariance-based classification is utterly effective, yielding misclassification error rates always lower than 15% also for challenging cases of scarce data and, for reasonable sample sizes ($K \geq 30$), always lower than 6.67%. In addition, we notice that misclassification rate decreases as the number of data increases. The results are even more satisfactory if related to the dimension of the covariance matrices of these data, i.e. $P = 100$, since a successful classification may be carried out with only 25-30 units per family.
If we compare the performances gained when using $\mathbf{S}$ with those attained by using $\widehat{\mathbf{S}}^*$, we see that a substantial improvement in accuracy has been achieved, being the classification in the second case, in practice, always perfect. Moreover, in this experiment the performances of Max-Swap combined with $\widehat{\mathbf{S}}^*$ were more stable across the trials, and almost always close to the best one for all trials. This may be an advantage with respect to $\mathbf{S}$, which in turn gave results more variable from trial to trial. On the contrary, from a computational point of view, resorting to $\mathbf{S}$ leads to faster simulations, while using $\widehat{\mathbf{S}}^*$ requires higher effort, especially when K is large.

4.3 Deoxygenated hemoglobin data

In this Subsection we apply our classification method to a real dataset belonging to a biomedical context. In particular, we deal with measurements along time of the concentration of deoxygenated hemoglobin in the brain of a group of six subjects, while they are carrying out a certain task. The measures of each subject were made on eight different points of the brain, four located in the central part of the left hemisphere and another four located in the central part of the right hemisphere. The measurement and preprocessing techniques as well as the experimental instruments used to collect data are described in [Tor+14; Zuc+13; Re+13].

Each statistical unit of the dataset consists of a sampling along 40 seconds of deoxygenated hemoglobin's concentration at the related location on the brain. Our classification purpose is to recognize the signals of patients whose trends in hemoglobin's concentration show wide fluctuations across their mean profiles from signals where the concentration varies little. This difference in dispersion has a clear counterpart in terms of difference in covariance operators, thus

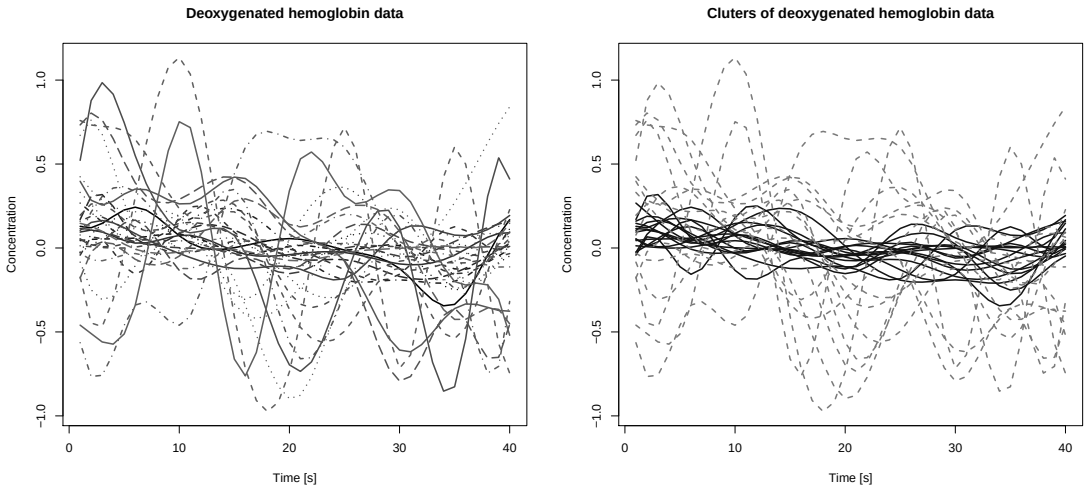

Figure 5: Deoxygenated hemoglobin's concentration data. In the left panel are represented the preprocessed data on which the classification is carried out. In the right panel is shown the output of classification.

we wish to apply our classification algorithm in order to detect the two clusters.
In a pre-processing stage, signals affected by artefacts due to the measurement procedure were removed, while the other were detrended and smoothed thanks to a B-Spline smoothing basis.
At the end of the pre-processing stage, a set of $N = 30$ signals, subdivided into two groups of $K = 15$ are available with a sampling rate of $1s$, so that $P = 40$. These data are depicted in Fig. 5.

We run the Max-Swap classification algorithm on these data to perform an unsupervised classification, both using $\mathbf{S}$ and $\widehat{\mathbf{S}}^*$ estimators, finding equal partitions of initial data. The results are shown in Fig. 5, and highlight how the algorithm is able to answer to our request, i.e. to detect two clusters of functions that are well distinguishable in terms of their different variability. We interpret these as different intensities in the reaction to the experiment.

5 Conclusions

In this paper we have studied the problem of performing an unsupervised classification on data whose difference lies in their variance-covariance structures rather than in their means. We have formulated it according to the general statistical framework of functional data, yet it can be of interest also in other contexts, such as for multivariate data. We have shown how the naive classification strategy is computationally intractable and we have proposed a new heuristic algorithm to override such issue. The algorithm is based on a proper quantification of the distance between estimates of covariance operators, which we assumed to be the natural Hilbert-Schmidt norm, and seeks for the partition of data producing the highest possible distance among estimated covariances. The partition is sought by modifying two initial guesses of the true groups with subsequent exchanges of units, in order to maximise the distance between estimated covariances. We have given its pseudo-code formulation and studied its convergence properties and complexity. A crucial point of the algorithm is the estimation of covariance operators, which can be done by standard sample covariance, but we have proposed a variant involving a linear shrinkage estimator,

which promises to be at least as accurate as sample covariance, and often better in terms of mean square error. By means of some examples we have collected empirical evidence to prove that the algorithm is able to solve suitably the classification problem, both when the variabilities are different in their magnitudes or in their directions. We compared the performances gained on functional data under the use of the sample estimator and of the linear shrinkage one, and found that both of them give definitely satisfactory results and that the use of linear shrinkage may provide a substantial improvement in terms of misclassification error.

References

[Bos00] D. Bosq. *Linear processes in function spaces: theory and applications*. Vol. 149. Springer, 2000.

[Efr82] B. Efron. "Maximum likelihood and decision theory". In: *The Annals of Statistics* (1982), pp. 340–356.

[Har75] J. A. Hartigan. *Clustering Algorithms*. 99th. New York, NY, USA: John Wiley & Sons, Inc., 1975.

[HK12] L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Vol. 200. Springer, 2012.

[LW03] O. Ledoit and M. Wolf. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection". In: *Journal of empirical finance* 10.5 (2003), pp. 603–621.

[LW04] O. Ledoit and M. Wolf. "A well-conditioned estimator for large-dimensional covariance matrices". In: *Journal of multivariate analysis* 88.2 (2004), pp. 365–411.

[Re+13] R. Re et al. "Multi-channel medical device for time domain functional near infrared spectroscopy based on wavelength space multiplexing". In: *Biomedical optics express* 4.10 (2013), pp. 2231–2246.

[RS05] J.O. Ramsay and B.W. Silverman. *Functional data analysis*. Springer, New York, 2005.

[SS05] J. Schafer and K. Strimmer. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics". In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005), pp. 1175–1189.

[Ste56] C. Stein. "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution". In: *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*. Vol. 1. 399. 1956, pp. 197–206.

[Tor+14] A. Torricelli et al. "Time domain functional NIRS imaging for human brain mapping". In: *Neuroimage* 85 (2014), pp. 28–50.

[Zuc+13] L. Zucchelli et al. "Method for the discrimination of superficial and deep absorption variations by time domain fNIRS". In: *Biomedical optics express* 4.12 (2013), pp. 2893–2910.

MOX Technical Reports, last issues

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

23/2014 Ieva, F., Paganoni, A.M., Tarabelloni, N.
Covariance Based Unsupervised Classification in Functional Data Analysis

22/2014 Arioli, G.
Insegnare Matematica con Mathematica

21/2014 Artina, M.; Fornasier, M.; Micheletti, S.; Perotto, S.
The benefits of anisotropic mesh adaptation for brittle fractures under plane-strain conditions

20/2014 Artina, M.; Fornasier, M.; Micheletti, S.; Perotto, S.
Anisotropic mesh adaptation for crack detection in brittle materials

19/2014 L.Bonaventura; R. Ferretti
Semi-Lagrangian methods for parabolic problems in divergence form

18/2014 Tumolo, G.; Bonaventura, L.
An accurate and efficient numerical framework for adaptive numerical weather prediction

17/2014 Discacciati, M.; Gervasio, P.; Quarteroni, A.
Interface Control Domain Decomposition (ICDD) Method for Stokes-Darcy coupling

16/2014 Dede, L.; Jaggli, C.; Quarteroni, A.
Isogeometric numerical dispersion analysis for elastic wave propagation

15/2014 Esfandiar, B.; Porta, G.; Perotto, S.; Guadagnini, A;
Anisotropic mesh and time step adaptivity for solute transport modeling in porous media

14/2014 Dassi, F.; Formaggia, L.; Zonca, S.
Degenerate Tetrahedra Recovering