Journal of Machine Learning Research 17 (2016) 1-21 Submitted 11/15; Revised 7/16; Published 9/16

# Covariance-based Clustering in Multivariate and Functional Data Analysis

**Francesca Ieva** FRANCESCA.IEVA@UNIMI.IT
*Department of Mathematics "F. Enriques"*
*Università degli Studi di Milano*
*Via Cesare Saldini 50, 20133 Milano, Italy*

**Anna Maria Paganoni** ANNA.PAGANONI@POLIMI.IT
**Nicholas Tarabelloni** NICHOLAS.TARABELLONI@POLIMI.IT
*MOX – Modeling and Scientific Computing*
*Department of Mathematics*
*Politecnico di Milano*
*Via Bonardi 9, 20133 Milano, Italy*

**Editor:** Sara van de Geer

## Abstract

In this paper we propose a new algorithm to perform clustering of multivariate and functional data. We study the case of two populations different 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 subdivides data in two groups maximising the distance between their induced covariances. The naive implementation of such an algorithm is computationally forbidding, so we propose a 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 clustering. We establish the effectiveness of our algorithm through applications to both synthetic data and a real data set coming from a biomedical context, showing also how the use of shrinkage estimation may lead to substantially better results.

**Keywords:** Clustering, covariance operator, operator distance, shrinkage estimation, functional data analysis

## 1. Introduction

The goal of performing clustering of data, in order to point out 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 Hartigan, 1975). Considerably less work can be found on methods attaining the clustering 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 differential correlation or distances between covariances of data. Nevertheless, it might happen to analyse groups of data that are scarcely distinguishable in terms of locations, while showing differences in their variability.

Examples can be found in biostatistics where, for instance, the dichotomy between physiological and pathological features often shows an interesting change in pattern of variability. Also, this is of great interest in genomics, where instead of focusing on gene expression levels, one could be interested in finding different correlation structures among subsets of data. This is the core task in the analysis of the differential co-expression of genes, namely the differential correlation structure among expression levels in different subsets of experimental conditions. In (Watson, 2006), for instance, the author proposes a method to identify groups of genes that are correlated in a first group of microarrays, and are not in a second one. In (Mitra et al., 2016), the authors provide a more complex modeling strategy that is able to specify the differential dependence structure by using Bayesian graphical models, and in (Cai and Zhang, 2016) authors provide, within a supervised framework, a way to estimate the differential correlation matrices of data belonging to two differentially expressed groups.

In this paper we tackle the problem from a different point of view, and focus on differences between global covariance structures of data belonging to two unknown groups, which will also be identified. 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. The method we propose 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, the monographs Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012). 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 two-class partitions of data, the one maximising the distance between the class-specific 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 data sets with common sizes. We therefore transform 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, Ledoit and Wolf, 2003, 2004 and Schafer and Strimmer, 2005). We show through experiments that this choice leads to a substantially improved clustering.

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 clustering method for two groups of 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 clustering performances through the application to both synthetic and real data sets. Discussion and conclusions are presented in Section 5.

## 2. Covariance Operators for Functional Data

Whenever our data can be interpreted as finite-dimensional samples of quantities that are intrinsically dependent on some continuous variable, such as time, we may resort to the model of functional data (see, for instance, Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012). At its core is the assumption that data are sample measurements of trajectories of stochastic processes valued in suitable function spaces.

In the following we recall the definition of covariance operator for functional data, along with its most important properties (for more details see, e.g., Bosq, 2000). Let $\mathcal{X}$ be a stochastic process taking values in $L^2(I)$, with $I \subset \mathbb{R}$ a compact interval, having mean function $\mathbb{E}\left[\mathcal{X}\right] = \mu$ and such that $\mathbb{E}\|\mathcal{X}\|^2 < \infty$, where we denote by $\|\cdot\|$ the $L^2(I)$ norm induced by the scalar product $\langle\cdot,\cdot\rangle$. Without loss of generality we can assume $\mu = 0$ and define the following covariance operator $\mathcal{C} \in \mathcal{L}\left(L^2(I); L^2(I)\right)$:

$$\langle y, \mathcal{C}x\rangle = \mathbb{E}\left[\langle x, \mathcal{X}\rangle\langle y, \mathcal{X}\rangle\right], \quad \forall x, y \in L^2(I). \tag{1}$$

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

$$\mathcal{C} = \sum_{k=1}^{\infty} \lambda_k \; e_k \otimes e_k, \tag{2}$$

where $\otimes$ indicates an outer product in $L^2(I)$, $\{e_k\}_{k=1}^{\infty}$ 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 that eigenvalues are 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

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

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, Bosq, 2000), since it holds:

$$\sum_{k=1}^{\infty} \lambda_k^2 < \infty. \tag{3}$$

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:

$$\langle \mathcal{U}, \mathcal{V} \rangle_S = \sqrt{\mathrm{Tr}\,(\mathcal{U} - \mathcal{V})\,(\mathcal{U} - \mathcal{V})^*}, \tag{4}$$

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

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

The space of Hilbert-Schmidt operators on $L^2(I)$, endowed with the scalar product (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:

$$d\,(\mathcal{U}, \mathcal{V}) = \|\mathcal{U} - \mathcal{V}\|_S^2 = \sum_{k=1}^{\infty} \eta_k^2, \tag{5}$$

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

## 3. Covariance-based Clustering

We face now the problem of classifying observations belonging to two different functional populations. Let $\mathcal{X}$ and $\mathcal{Y}$ be stochastic processes on $L^2(I)$ generated by the laws $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$. We imagine to have a set of $N$ data in some data set $D$ (i.e., a collection of observations) composed in the following way by an equal number of observations from two families: $D = \{X_1, \ldots, X_K, Y_1, \ldots, Y_K\}$, with $K = N/2$ and where $\{\mathcal{X}_i\}_{i=1}^K, \{\mathcal{Y}_j\}_{j=1}^K$ are i.i.d and follow respectively $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$.
We introduce the following quantities:

$$\begin{aligned} \mu_1 &= \mathbb{E}\left[X_i\right], \quad \mathcal{C}_1 = \mathbb{E}\left[X_i \otimes X_i\right], \quad \forall\ i = 1, \ldots, K, \\ \mu_2 &= \mathbb{E}\left[Y_j\right], \quad \mathcal{C}_2 = \mathbb{E}\left[Y_j \otimes Y_j\right], \quad \forall\ j = 1, \ldots, K. \end{aligned}$$

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

$$I^{(0)} = \Big( \overbrace{1, 2, \ldots, K}^{I_1^{(0)}}, \overbrace{K+1, \ldots, N}^{I_2^{(0)}} \Big), \tag{6}$$

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 into two subsets:

$$I^{(i)} = \left( I_1^{(i)};\ I_2^{(i)} \right), \quad i \in \{1, \ldots, N_C\}, \tag{7}$$

where $I^{(i)}$ denotes the $i$-th combination out of $N_C = \binom{N}{K}$, however enumerated.
The sample estimators of means and covariance operators induced by this subdivision are denoted, 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 $\mu_1$, $\mu_2$ 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 clustering method is based on the following, crucial assumption:

**Assumption 1** *We assume that observations drawn from families $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ constituting the data set $D$ may be distinguished on the basis of their covariances, but not of their means, i.e. $\mu_1^{(0)} = \mu_2^{(0)}$ and $\mathcal{C}_1^{(0)} \neq \mathcal{C}_2^{(0)}$, and therefore $\|\mu_1^{(0)} - \mu_2^{(0)}\| = 0$ and $d(\mathcal{C}_1^{(0)}, \mathcal{C}_2^{(0)}) \gg 0$.*

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

In order to illustrate the clustering method we propose, let us consider a situation where the original data set has been split according to a vector of indices $I^{(i)} = [I_1^{(i)}; I_2^{(i)}]$. For the sake of simplicity, let us encode this through the binary variables $w_{j,g} = \mathbb{1}(X_j \in I_g^{(i)})$, $j = 1, \dots, K$, $g = 1, 2$, and $v_{j,g} = \mathbb{1}(Y_j \in I_g^{(i)})$, $j = 1, \dots, K$, $g = 1, 2$. In other words, such variables express the fact that observation $j$ from the original population $\mathcal{X}$ or $\mathcal{Y}$ belongs to the first ($I_1^{(i)}$) or second ($I_2^{(i)}$) group into which data are split. According to the setting previously introduced, it is:

$$K = \sum_{g=1}^{2} \sum_{j=1}^{K} w_{j,g}, \qquad K = \sum_{g=1}^{2} \sum_{j=1}^{K} v_{j,g},$$
$$K = \sum_{j=1}^{K} w_{j,1} + \sum_{j=1}^{K} v_{j,1}, \qquad K = \sum_{j=1}^{K} w_{j,2} + \sum_{j=1}^{K} v_{j,2}.$$

Then, we can re-write the sample covariances $\widehat{\mathcal{C}}_1^{(i)}$ and $\widehat{\mathcal{C}}_2^{(i)}$ as:

$$\widehat{\mathcal{C}}_1^{(i)} = \frac{\sum_{j=1}^{K} w_{j,1} X_j \otimes X_j + \sum_{k=1}^{K} v_{k,1} Y_k \otimes Y_k}{K},$$
$$\widehat{\mathcal{C}}_2^{(i)} = \frac{\sum_{j=1}^{K} w_{j,2} X_j \otimes X_j + \sum_{k=1}^{K} v_{k,2} Y_k \otimes Y_k}{K}.$$

If we compute the difference $\widehat{\mathcal{C}}_1^{(i)} - \widehat{\mathcal{C}}_2^{(i)}$, we obtain:

$$\widehat{\mathcal{C}}_1^{(i)} - \widehat{\mathcal{C}}_2^{(i)} = \frac{1}{K} \sum_{j=1}^{K} (w_{j,1} - w_{j,2})\, X_j \otimes X_j + \frac{1}{K} \sum_{k=1}^{K} (v_{k,1} - v_{k,2})\, Y_k \otimes Y_k,$$

hence, exploiting the distance between covariance operators:

$$
\begin{aligned}
K^2 \| \widehat{\mathcal{C}}_1^{(i)} - \widehat{\mathcal{C}}_2^{(i)} \|_S^2 = & \left\| \sum_{j=1}^{K} \left( w_{j,1} - w_{j,2} \right) X_j \otimes X_j \right\|_S^2 + \left\| \sum_{j=1}^{K} \left( v_{j,1} - v_{j,2} \right) Y_j \otimes Y_j \right\|_S^2 + \\
& + 2 \sum_{j=1}^{K} \sum_{k=1}^{K} \left( w_{j,1} - w_{j,2} \right) \left( v_{k,1} - v_{k,2} \right) \langle X_j , Y_k \rangle^2 \\
& = \sum_{j=1}^{K} \| X_j \otimes X_j \|_S^2 + 2 \sum_{j<k} \left( w_{j,1} - w_{j,2} \right) \left( w_{k,1} - w_{k,2} \right) \langle X_j , X_k \rangle^2 + \\
& + \sum_{j=1}^{K} \| Y_j \otimes Y_j \|_S^2 + 2 \sum_{j<k} \left( v_{j,1} - v_{j,2} \right) \left( v_{k,1} - v_{k,2} \right) \langle Y_j , Y_k \rangle^2 + \\
& + 2 \sum_{j=1}^{K} \sum_{k=1}^{K} \left( w_{j,1} - w_{j,2} \right) \left( v_{k,1} - v_{k,2} \right) \langle X_j , Y_k \rangle^2 .
\end{aligned}
\tag{8}
$$

Let us now call $\delta^X_{j,k} = \left( w_{j,1} - w_{j,2} \right) \left( w_{k,1} - w_{k,2} \right)$, $\delta^Y_{j,k} = \left( v_{j,1} - v_{j,2} \right) \left( v_{k,1} - v_{k,2} \right)$ and $\delta^{XY}_{j,k} = \left( w_{j,1} - w_{j,2} \right) \left( v_{k,1} - v_{k,2} \right)$. Now, it is: $\delta^X_{j,k} = +1$ if observations $X_j$ and $X_k$ are assigned to the same group, while on the contrary it is $\delta^X_{j,k} = -1$. The same applies for $\delta^Y_{j,k}$ with $Y_j$ and $Y_k$. Finally, $\delta^{XY}_{j,k} = +1$ if $X_j$ and $Y_k$ are assigned to different groups, and $\delta^{XY}_{j,k} = -1$ on the contrary. It is now clear that, the distance between covariance operators is increased when two observations of populations $\mathcal{X}$ or $\mathcal{Y}$ are both assigned to the same group, or when two observations of the opposite populations $\mathcal{X}$ and $\mathcal{Y}$ are assigned to different groups. These remarks suggest the idea that, under the previous assumption, by recovering the original labelling of the two populations $\mathcal{X}$ and $\mathcal{Y}$ the distance between the induced covariance operators is increased.

If we replace the estimators of covariance operators in (8) with their expected values,

$$
\begin{aligned}
\mathbb{E} \left[ \widehat{\mathcal{C}}_1^{(i)} \right] = \mathcal{C}_1^{(i)} = \frac{\sum_{j=1}^{N} w_{j,1}}{K} \, \mathcal{C}_1 + \frac{\sum_{j=1}^{K} v_{j,1}}{K} \, \mathcal{C}_2 , \\
\mathbb{E} \left[ \widehat{\mathcal{C}}_2^{(i)} \right] = \mathcal{C}_2^{(i)} = \frac{\sum_{j=1}^{N} w_{j,2}}{K} \, \mathcal{C}_1 + \frac{\sum_{j=1}^{K} v_{j,2}}{K} \, \mathcal{C}_2 ,
\end{aligned}
$$

then, by denoting $N_{1,2} = \sum_{j=1}^{K} w_{j,2}$ and also considering the relations among the variables $w_{j,g}$ and $v_{j,g}$ we get:

$$
d(\mathcal{C}_1^{(i)}, \mathcal{C}_2^{(i)}) = \left\| \mathbb{E} \left[ \widehat{\mathcal{C}}_1^{(i)} \right] - \mathbb{E} \left[ \widehat{\mathcal{C}}_2^{(i)} \right] \right\|_S^2 = \left( 1 - 2 \frac{N_{1,2}}{K} \right)^2 \| \mathcal{C}_1 - \mathcal{C}_2 \|_S^2 , \tag{9}
$$

specifying that the maximum distance between (exact) covariances is attained when the groupings coincide with the original but unknown indexing of the data set.

If assumption (1) is true and in view of (9), a natural way to recover the true indexing can 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:

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

Identity (9) ensures 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 (9), yet the groups represent the same partition of data, for this reason 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 ($\mathbf{P}$) into:

$$\left[\widehat{I}_1^*; \widehat{I}_2^*\right] = \underset{i \in R_C}{\arg\max} \left\{ d\left(\widehat{\mathcal{C}}_1^{(i)}, \widehat{\mathcal{C}}_2^{(i)}\right) \right\}, \quad R_C = \{1, \dots, N_C\}, \tag{$\hat{\mathbf{P}}$}$$

The method we propose coincides with finding a solution to problem ($\hat{\mathbf{P}}$).
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 ($\hat{\mathbf{P}}$), 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 data set, 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 method also to data sets with a common size.

#### 3.1.1 Max-Swap algorithm

We propose to rephrase problem ($\hat{\mathbf{P}}$) 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.

This strategy can be also motivated by (8), since it performs swaps between observations in order to increase $\|\widehat{\mathcal{C}}_1^{(i)} - \widehat{\mathcal{C}}_2^{(i)}\|$, thus trying to assign the correct values to $w_{j,g}$ and $v_{j,g}$.

**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 $s \in 1, \ldots, K$ do
        for $t \in 1, \ldots, K$ 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 $(\tilde{I}_1, \tilde{I}_2)$;
            $D_{s,t} = d\left(\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2\right)$;
    $(s^*, t^*) = \arg\max_{s,t} D_{s,t}$;
    $d^k = D_{s^*,t^*}$;
    $(\Delta d)^k = d^k - d^{k-1}$;
    $I_1^k = \bigcup_{p \neq s^*} I_1^{k-1}(p) \cup I_2^{k-1}(t^*)$;
    $I_2^k = \bigcup_{q \neq t^*} I_2^{k-1}(q) \cup I_1^{k-1}(s^*)$;
    $k = k + 1$;
$d^{**} = d^{k-1}$;
$I_1^{**} = I_1^{k-1}$;
$I_2^{**} = I_2^{k-1}$;
```

When searching the best swap of size up to $J$, we must explore a number of combinations 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 $(\hat{\mathbf{P}})$ in just one step, since it holds:

$$\binom{N}{K} = \sum_{i=0}^{K} \binom{K}{i}^2 = \sum_{i=1}^{K} \binom{K}{i}^2 + 1.$$

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, where we specify for the sake of clarity that the symbol $I_1^k(p)$, for instance, indicates the $p$-th element of the set of indexes $I_1^k$. In the following we will denote the estimated set of indexes at step $k$ of algorithm with superscript $k$ without brackets, $(I_1^k,\, I_2^k)$.

### 3.1.2 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:*

$$(\Delta d)^k > 0 \quad \forall k \geq 1, \tag{10}$$

*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 (10), 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:

**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* $(\mathbf{P})$ *in at most* $K/2$ *steps.*

**Proof** This is a consequence of Proposition 1 and the convexity of the objective function $d(\mathcal{C}_1^{(i)}, \mathcal{C}_2^{(i)})$ w.r.t $N_{1,2}$ showed in (9), making it impossible that local maxima exist. ■

A 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. A simple way to correct for the possibility to select only a local optimum is to implement a *restart*-like strategy, namely run multiple instances of the algorithm with different initialisations and to select the best results in terms of optimised objective function. This is a simple and classic solution to the drawbacks of general local-optimisation algorithms.

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 small 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 $J > 1$.

### 3.1.3 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 data set, 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 $N_{1,2}$ in (9) to follow a hypergeometric distribution:

$$N_{1,2} \sim \mathrm{Hyper}(N, K, K),$$

i.e.,

$$\mathbb{P}\left(N_{1,2} = h\right) = \frac{\binom{K}{h}\binom{K}{K-h}}{\binom{N}{K}}, \quad \mathbb{E}\left[N_{1,2}\right] = \frac{K}{2},$$

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 $N_{1,2}$ 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 value for $N_{1,2}$ resulting from the initial guess, the number of iterations to convergence is $N_{\text{it}} = \min(N_{1,2}, K - N_{1,2})$, corresponding to $N_{\text{it}}$ consecutive and correct swaps.

Then, by initialising at random the algorithm:

$$N_{\text{it}} \leq \frac{K}{2}.$$

If we summarise the complexity of solving problem $(\hat{\mathbf{P}})$ with the number of combinations $N_{\text{comb}}$ 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:

$$N_{\text{comb}}^{MS} = K^2 \; N_{\text{it}} \leq \frac{K^3}{2}, \qquad N_{\text{comb}}^{\text{N}} = \binom{N}{K},$$

where the multiplicative factor $K^2$ in $N_{\text{comb}}^{MS}$ 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.

### 3.1.4 Classification of new observations

Once the algorithm has been run and the two groups constituting the mixed data set have been found, a simple rule can be applied to use the clusters in order to classify new observations. We suggest to consider each new unit separately, and compute the covariances obtained by including the unit either in the first or second group. For the sake of simplicity, we denote them, respectively, by $\widetilde{\mathcal{C}}_1$ and $\widetilde{\mathcal{C}}_2$. Then we compute the distances $\tilde{d}_1 = d(\widetilde{\mathcal{C}}_1, \widehat{\mathcal{C}}_1)$ and $\tilde{d}_2 = d(\widetilde{\mathcal{C}}_2, \widehat{\mathcal{C}}_2)$, where here we indicate by $\widehat{\mathcal{C}}_1$ and $\widehat{\mathcal{C}}_2$ the covariances of the two groups determined at the end of the clustering stage. Then we attribute the new unit to the group for which the distance $\tilde{d}$ is minimum.

## 3.2 Shrinkage estimation of covariance

In this subsection we consider the problem of improving the estimation of covariance operators, so that clustering is more accurate. In particular, we will describe an alternative estimator of data covariance than the sample covariance, which is better conditioned and in some circumstances achieves lower MSE. We will make use of it in our clustering algorithm as an alternative to sample covariance.

Let us consider a generic family of functional data $\mathcal{X} \sim P_{\mathcal{X}}$, such that $\mathbb{E}\left[\mathcal{X}\right] = 0$, $\mathbb{E}\|\mathcal{X}\|^2 < \infty$. We denote its covariance with $\mathcal{C}$ and we imagine to estimate it with $\widehat{\mathcal{C}}$. Our purpose is to find its 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**):

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

where the minimum is sought among all possible estimators $\widehat{\mathcal{C}}$ of $\mathcal{C}$. We point out that in $\text{MSE}_S$ 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,:

$$X_i = \left(X_i(t_j)\right)_{j=1}^P\,, \quad I^h = [t_1, \dots, t_P]\,, \tag{11}$$

where, for the sake of simplicity, we have imagined the grid $I^h$ to be regularly spaced (although this is not mandatory), i.e., $t_{j+1} - t_j = h > 0$ for $j = 1, \dots, P-1$. 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, Ramsay and Silverman, 2005). 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 (11) have already been preprocessed.

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

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} X_i \; X_i^T. \tag{12}$$

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

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

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 ($\hat{\mathbf{E}}$) 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*, Stein, 1956).
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 jackknife or bootstrap, but their computational cost renders them practically useless in our clustering algorithm, which requires to repeatedly estimate covariance matrices. Shrinkage estimation has been explicitly applied to the context of large covariance matrices in (Ledoit and Wolf, 2003, 2004) and (Schafer and Strimmer, 2005), turning out in a sufficiently lightweight procedure. In those works, authors start from problem ($\hat{\mathbf{E}}$) and build an estimator that is asymptotically more accurate and better conditioned than sample covariance. We follow the approach described in (Ledoit and Wolf, 2004) and consider the class of linear shrinkage estimators of the form:

$$\widehat{\mathbf{C}} = \mu\gamma \, \mathbf{I} + (1-\gamma)\mathbf{S}, \tag{13}$$

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 ($\hat{\mathbf{E}}$) is solved with respect to the optimal values of $\mu$ and $\gamma$:

$$(\mu^*, \gamma^*) = \arg\min_{\mu,\gamma} \; \mathbb{E} \frac{\|\mathbf{C} - \mu\gamma\mathbf{I} - (1-\gamma)\mathbf{S}\|_F^2}{P}. \tag{14}$$

If we introduce the quantities:

$$\alpha^2 = \frac{\|\mathbf{C} - \mu \mathbf{I}\|_F^2}{P}, \quad \beta^2 = \frac{\mathbb{E}\ \|\mathbf{S} - \mathbf{C}\|_F^2}{P}, \quad \delta^2 = \frac{\mathbb{E}\ \|\mathbf{S} - \mu \mathbf{I}\|_F^2}{P}, \tag{15}$$

and note that these are subjected to $\alpha^2 + \beta^2 = \delta^2$, we can perform the explicit minimization in equation (14). The expressions of $\mu^*$ and $\gamma^*$ are:

$$\mu^* = \frac{\langle \mathbf{C}, \mathbf{I} \rangle_F}{P} = \frac{\mathrm{Tr}\,(\mathbf{C})}{P}, \qquad \gamma^* = \frac{\beta^2}{\delta^2}, \tag{16}$$

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

$$\mathbf{S}^* = \mu^* \frac{\beta^2}{\delta^2} \mathbf{I} + \frac{\alpha^2}{\delta^2} \mathbf{S}. \tag{17}$$

Of course, estimator (17) depends on the unknown exact covariance matrix $\mathbf{C}$, even though only through four scalar functions. In (Ledoit and Wolf, 2004) authors solve this problem by proposing the following estimators for $\alpha$, $\beta$, $\delta$ and $\mu^*$:

$$\widehat{\mu}^* = \frac{\mathrm{Tr}\,(\mathbf{S})}{P}, \qquad \widehat{\delta^2} = \frac{\|\mathbf{S} - \widehat{\mu}^* \mathbf{I}\|_F^2}{P}, \tag{18}$$

$$\widehat{\beta^2} = \min \left( \widehat{\delta^2}; \frac{1}{N^2} \sum_{k=1}^{N} \frac{\|X_k\ X_k^T - \mathbf{S}\|_F^2}{P} \right), \tag{19}$$

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

$$\widehat{\mathbf{S}}^* = \widehat{\mu^*} \frac{\widehat{\beta^2}}{\widehat{\delta^2}} \mathbf{I} + \frac{\widehat{\alpha^2}}{\widehat{\delta^2}} \mathbf{S}. \tag{20}$$

In (Ledoit and Wolf, 2004) authors show how estimates (18) 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 Ledoit and Wolf, 2004 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 $\widehat{\mathbf{S}}^*$. For a detailed description of how and when shrinkage estimation of covariance is recommended over the sample estimation, see for instance (Ledoit and Wolf, 2004).

## 4. Case Studies

In this section we provide three simulations involving our proposed clustering method. In Subsection 4.1 we show a first example, regarding standard bivariate data, in order to give a clear geometric idea of clustering 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 clustering arising from Max-Swap algorithm is assessed against the true identities of data. In Subsection 4.3, instead, we apply the clustering algorithm 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 clustering arising from using Max-Swap algorithm. In order to ease the geometrical interpretation, we chose to focus on two bivariate data sets, 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 data sets having the same means but different variance-covariance structures. In particular, a generic clustering 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:

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

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 $\mathbf{C}_x$ and $\mathbf{C}_y$ 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 data set $D$ 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 Figure 1.
We considered also another data set, 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:

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

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

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

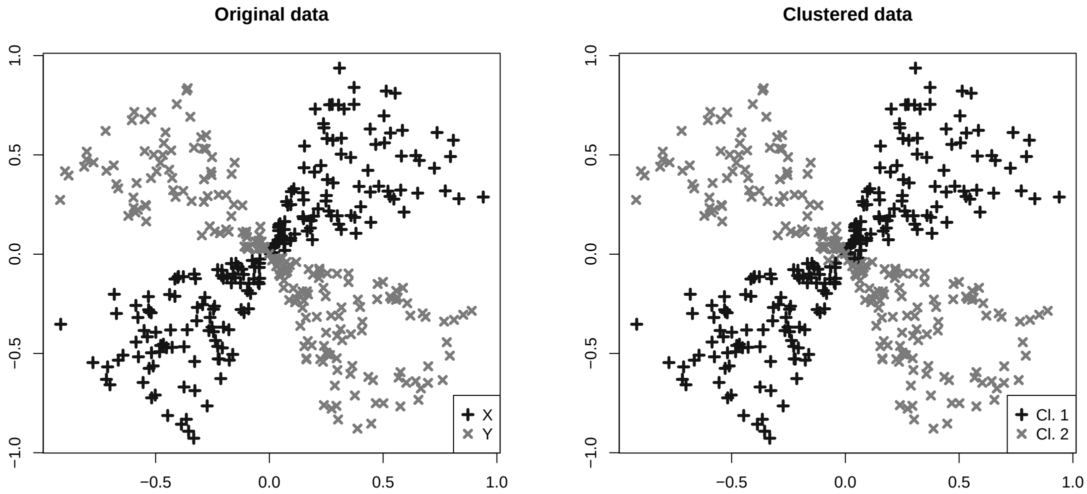

Figure 1: *Left*: Hourglass data set used in the first multivariate experiment, collecting $N = 400$ points subdivided into family X ($K = 200$ points, marked by $+$) and family Y ($K = 200$ points, marked by $\times$). *Right*: Outcome of clustering via Max-Swap algorithm.

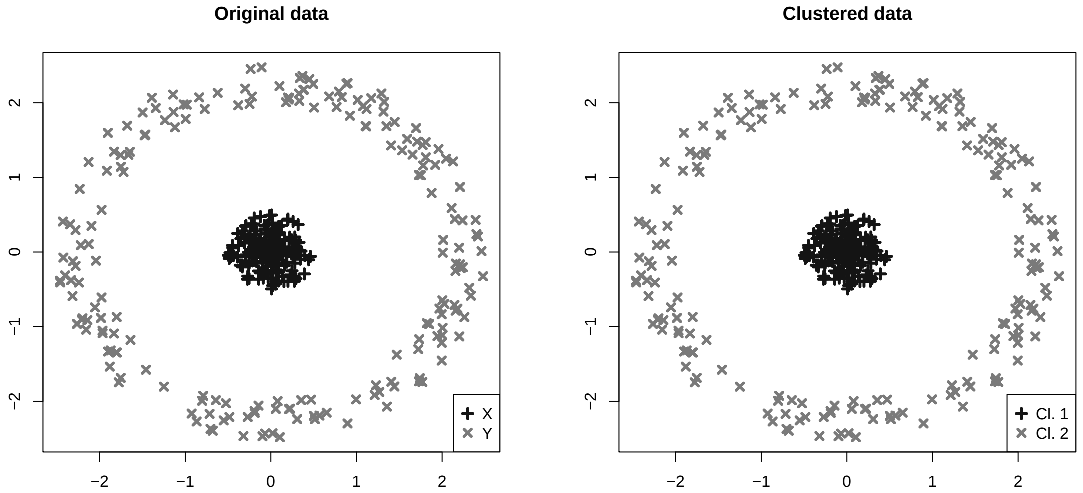

Figure 2: Bull's eye data set used in the second multivariate experiment, collecting $N = 400$ points subdivided into family X ($K = 200$ points, marked by $+$) and family Y ($K = 200$ points, marked by $\times$). *Right*: Outcome of clustering via Max-Swap algorithm.

that clearly differ only in terms of their variability's magnitude. *Bull's eye* data set 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 Figure 2.

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 was run for 10 times, keeping the result for which the objective function was highest. 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, $\mathbf{S}$. The results of the clustering procedure are also shown in Figure 1 and Figure 2 (right panels), where it is clear how the clustering method is able to detect the observations belonging to the two populations, whose difference is is only in their covariances. Since in this simulated scenario we know the law generating observations, and therefore the labels of the generated data, we are able to assess the performances of the clustering procedure by comparing the two identified groups of data with the original labels. In particular, in the case of hourglass data, only 3 observations out of 200 in each cluster belong to the other population, and are all located very close to the data centre. In the bull's eye example, instead, the two clusters are composed of elements coming from only one population.

The outcome of standard clustering algorithms, like K-means or hierarchical clustering, show the complete inefficacy of location-based clustering for such data sets, where data are mostly different in terms of their variability. In particular, 20 K-means ($K = 2$) runs on Bull's eye and Hourglass data sets yield, if compared with the true labelling of observations, a mis-classification rate of $0.27 \pm 0.01$ and $0.47 \pm 0.01$, respectively, while 20 runs of a hierarchical clustering with euclidean distance and Ward linkage give a mis-classification rate of $0.29 \pm 0.05$ and $0.46 \pm 0.04$, respectively.

### 4.2 Synthetic functional data

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

$$\mathcal{C}_x = \sum_{i=1}^{L} \sigma_i \; e_i \otimes e_i, \qquad \mathcal{C}_y = \sum_{i=1}^{L} \eta_i \; e_i \otimes e_i, \tag{21}$$

where $\{e_i\}_{i=1}^{L}$ are the first $L$ elements of the orthonormal Fourier basis on the interval $I = [0, 1]$, save for the constant, i.e.,:

$$e_{2k-1} = \sqrt{2} \sin\left(2k\pi x\right), \quad e_{2k} = \sqrt{2} \cos\left(2k\pi x\right), \quad x \in I,$$

for $k = 1, \ldots, L/2$, and the eigenvalues are chosen as:

$$\sigma_i = 1, \qquad \eta_i = \frac{\sigma_i}{\sqrt{5}}, \quad \forall\; i = 1, \ldots, L. \tag{22}$$

In what follows we considered $L = 30$. A visual representation of the related covariance functions is in Figure 3. It is clear from (21) and from Figure 3 that covariances $\mathcal{C}_x$ and

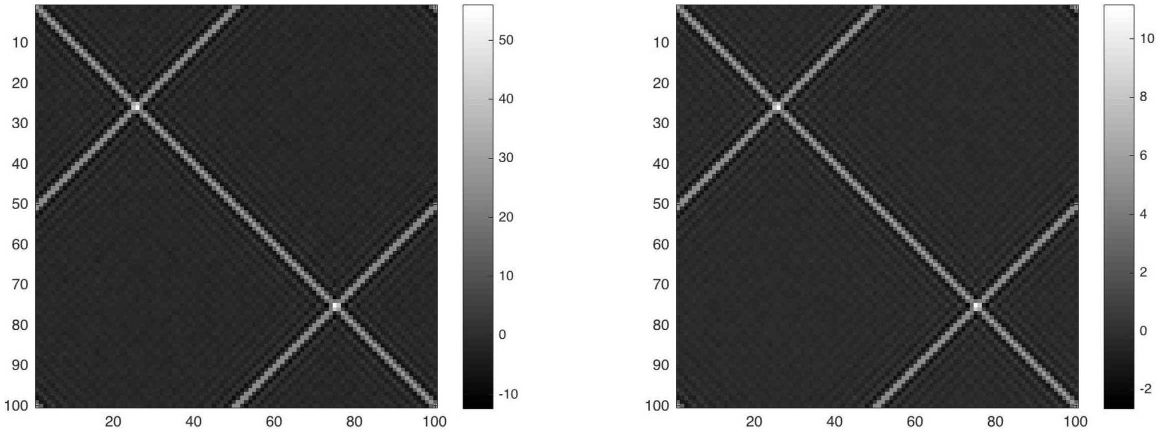

Figure 3: Contour plot of covariances $\mathcal{C}_x$ and $\mathcal{C}_y$ for the experiment with synthetic functional data. The different scales of contour plots show that the difference between the variance-covariance structures is only in magnitude.

$\mathcal{C}_y$ 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 (21):

$$X_i = \sum_{j=1}^{L} \xi_{ij} \sqrt{\sigma_j}\ e_j, \qquad Y_i = \sum_{j=1}^{L} \zeta_{ij} \sqrt{\eta_j}\ e_j, \tag{23}$$

for $i = 1, \ldots, K$, 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 $K \in \{20, 25, 30,\ 35,\ 40, 45, 50\}$, corresponding to total cardinalities of $N \in \{40, 50, 60, 70, 80, 90, 100\}$.

We applied our clustering algorithm to each synthetic data set. The different values of $K$ (i.e., of $N$) allow to study the performances of clustering 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 performances tends to improve.

We used Max-Swap algorithm both with the standard sample covariance estimator, $\mathbf{S}$, and with the shrinkage covariance estimator $\widehat{\mathbf{S}}^{*}$.

Like in the case of multivariate data, we know the laws generating observations, hence their original labels, therefore we can analyse the identified clusters in terms of the composition of units from $X$ and $Y$. This allows us to understand whether the algorithm is able to detect groups of observations that we know *a priori* are different only in their variability. We report the results of the clustering procedure in Tab. 1. Similarly to the case of multivariate synthetic data of Subsection 4.1, each of them is related to the one trial in a set of 10 for which the distance between covariances was highest. This was done in order to take

| | | **Sample covariance** $(\mathbf{S})$ | | **Shrinkage estimator** $(\widehat{\mathbf{S}}^*)$ | |
|---|---|---|---|---|---|
| N | K | **Miscl.** $(\mathbf{1},\mathbf{2})$ | **Err.** | **Miscl.** $(\mathbf{1},\mathbf{2})$ | **Err.** |
| 40 | 20 | $2\ (1,1)$ | 5% | $0\ (0,0)$ | 0% |
| 50 | 25 | $2\ (1,1)$ | 4% | $0\ (0,0)$ | 0% |
| 60 | 30 | $0\ (0,0)$ | 0% | $0\ (0,0)$ | 0% |
| 70 | 35 | $6\ (3,3)$ | 8.5% | $2\ (1,1)$ | 3% |
| 80 | 40 | $2\ (1,1)$ | 2.5% | $2\ (1,1)$ | 2.5% |
| 90 | 45 | $0\ (0,0)$ | 0% | $0\ (0,0)$ | 0% |
| 100 | 50 | $0\ (0,0)$ | 0% | $0\ (0,0)$ | 0% |

Table 1: Clustering performances results for the application with synthetic functional data.

account of the heuristic nature of the algorithm.

Results undoubtedly highlight that covariance-based clustering is effective, yielding groups which can easily be interpreted as the original ones up to an error always lower than 10%, also for challenging cases of scarce data. In addition, for reasonable sample sizes, the error tends to get lower. 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 clustering 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. 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 clustering method to a real data set belonging to a biomedical context. In particular, we deal with data produced by functional near-infrared spectroscopy (fNIRS), an optical technique able to noninvasively monitor the cerebral hemodynamic at cortical level. Exploiting the relatively low absorption of biological tissues, light in the red and near-infrared wavelength range can penetrate the human head down to some centimeters and reach the cerebral cortex. Therefore, fNIRS can provide a measure of oxy- and deoxy-hemoglobin, the main chromophores contributing to light absorption at this wavelength range. In particular we study the measurements along time of the concentration of deoxygenated hemoglobin in the brain of a group of six right-handed healthy subjects (male, 44 years old) while they are carrying out a motor task (i.e., squeezing a soft ball in the right hand) at a rate of 2 Hz guided by a metronome. 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

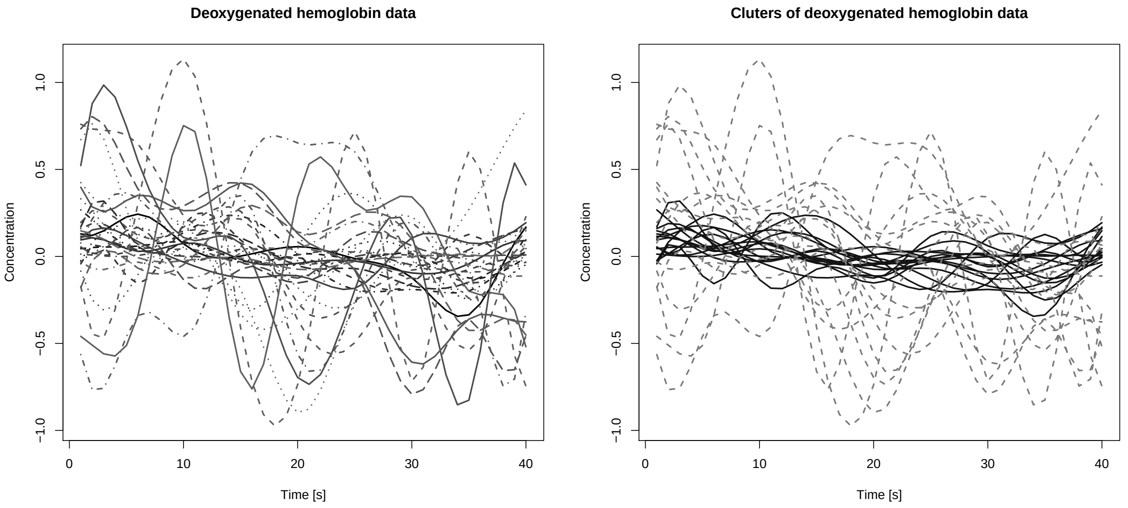

Figure 4: Deoxygenated hemoglobin's concentration data. In the left panel are represented the preprocessed data on which the clustering is carried out. In the right panel is shown the output.

measurement and preprocessing techniques as well as the experimental instruments used to collect data are described in (Torricelli et al., 2014; Zucchelli et al., 2013; Re et al., 2013).

Each statistical unit of the data set consists of a sampling along 40 seconds of deoxygenated hemoglobin's concentration at the related location on the brain. Our clustering 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. The aim of the study was in fact to detect different behaviours corresponding to activated vs. non activated cerebral areas. This activation is reflected in difference in covariance operators more than in difference in the mean level of deoxygenated hemoglobin concentration, thus we wish to apply our clustering algorithm in order to detect the two clusters.
In a pre-processing stage, signals affected by artifacts due to the measurement procedure were removed, while the others were de-trended 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 Figure 4.

We run the Max-Swap clustering algorithm on these data to perform clustering, both using $\mathbf{S}$ and $\widehat{\mathbf{S}}^*$ estimators, finding equal partitions of initial data. The results are shown in Figure 4, 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 activated versus non activated brain regions and our results are in agreement with those obtained in (Bonomini et al., 2015).

## 5. Conclusions

In this paper we have studied the problem of performing clustering on two groups of 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 clustering 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 clustering 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 clustering performances.

## References

V. Bonomini, R. Re, L. Zucchelli, F. Ieva, L. Spinelli, D. Contini, A. M. Paganoni, and Torricelli A. A new linear regression method for statistical analysis of fnirs data. *Biomedical optics express*, 2(6):615–630, 2015.

D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*, volume 149 of *Lecture Notes in Statistics*. Springer, 2000.

T. T. Cai and A. Zhang. Inference for high-dimensional differential correlation matrices. *Journal of Multivariate Analysis*, 143:107–126, 2016.

J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1975.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*, volume 200 of *Springer Series in Statistics*. Springer, 2012.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

R. Mitra, P. Müller, and Y. Ji. Bayesian graphical models for differential pathways. *Bayesian Analysis*, 11(1):99–124, 2016.

J. O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

R. Re, D. Contini, M. Turola, L. Spinelli, L. Zucchelli, M. Caffini, R. Cubeddu, and A. Torricelli. Multi-channel medical device for time domain functional near infrared spectroscopy based on wavelength space multiplexing. *Biomedical optics express*, 4(10):2231–2246, 2013.

J. Schafer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1175–1189, 2005.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956.

A. Torricelli, D. Contini, A. Pifferi, M. Caffini, R. Re, L. Zucchelli, and L. Spinelli. Time domain functional nirs imaging for human brain mapping. *Neuroimage*, 85:28–50, 2014.

M. Watson. Coxpress: differential co-expression in gene expression data. *BMC bioinformatics*, 7(1):1, 2006.

L. Zucchelli, D. Contini, R. Re, A. Torricelli, and L. Spinelli. Method for the discrimination of superficial and deep absorption variations by time domain fnirs. *Biomedical optics express*, 4(12):2893–2910, 2013.