An Object Oriented Approach to Multimodal Imaging Data in Neuroscience

Andrea Cappozzo, Federico Ferraccioli, Marco Stefanucci and Piercesare Secchi

Abstract We propose a methodological framework for exploring complex multimodal imaging data from a neuroscience study with the aim of identifying a data-driven group structure in the patients sample, possibly connected with the presence/absence of lifetime mental disorder. The functional covariances of fMRI signals are first considered as data objects. Appropriate clustering procedures and low dimensional representations are proposed. For inference, a Frechet estimator of both the covariance operator itself and the average covariance operator is used. A permutation procedure to test the equality of the covariance operators between two groups is also considered. We finally propose a method to incorporate spatial dependencies between different brain regions, merging the information from both the Structural Networks and the Dynamic functional activity.

Keywords Data objects · Functional data analysis · Principal components Multimodal Imaging · Neuroscience

1 Introduction

The following work arises from the StartUp Research experience, a workshop held at Certosa di Pontignano on June 25–27 2017. Seven groups formed by early-career

A. Cappozzo
Department of Statistics and Quantitative Methods,
University of Milano-Bicocca, Milan, Italy
e-mail: andrea.cappozzo@unimib.it

F. Ferraccioli (✉)
Department of Statistical Sciences, University of Padova, Padua, Italy
e-mail: ferraccioli@stat.unipd.it

M. Stefanucci
Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
e-mail: marco.stefanucci@uniroma1.it

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

A. Canale et al. (eds.), *Studies in Neural Data Science*, Springer Proceedings in Mathematics & Statistics 257, https://doi.org/10.1007/978-3-030-00039-4_4

researchers and a senior mentor acting as group leader were challenged to develop novel methods for analysing a common dataset.

Both researchers and practitioners involved in the field of data analysis are nowadays increasingly challenged in confronting with data structures that lie outside the classical Euclidean framework. That is, thanks to the technological advancements of measurement machineries, not only datasets are becoming massive in terms of size (the way-too-exploited buzzword big data is a living proof of the concept) but also substantial in terms of data complexity. As a consequence, statisticians are encouraged to sharpen their mathematical and programming skills for tackling the enormous knowledge-discovery opportunities that lie within these complex datasets. Object oriented data analysis (OODA) is a framework, firstly introduced in [24], for approaching data challenges in which the *object* of the analysis (i.e., the observation or statistical unit) possesses distinctive features that would not be exploited by performing a classical multivariate analysis after data dimension reduction. Examples of data objects that are considered by OODA include (but are not limited to) curves, images, tree structured data and positive semi-definite matrices [14]. In such a context mathematics plays a fundamental role in rigorously defining the embedding space and properties of the objects under study, and consequently fostering the development of new statistical methodologies. Two central notions are the base-ground for understanding the conceptual framework of OODA:

- *Object Space*: is the set in which the mathematical representation of the data lie. For example, the employed object space for the dynamic functional activity (see Sect. 2) is the Hilbert space $\mathbb{L}^2$ of square-integrable functions.
- *Feature Space*: is the set of features that numerically represent the data object. The feature space for the scan-rescan dynamic functional activity of the 24 subjects in the study (see Sect. 2) is a digitized $70 \times 404 \times 24 \times 2$ array.

The OODA framework is particularly appropriate when applied to neuroscience, where the large use of Magnetic Resonance Imaging (MRI) in the study of brain connectivity and activity has recently created new challenges for statisticians. The nature and complexity of data coming from electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and diffusion tensor imaging (DTI) have favoured the development ofz ad-hoc methodologies greatly expanding the statistical neuroscience literature [8, 18]. During the StartUp Research workshop our group attempted to analyse the provided dataset employing mathematical tools coming from OODA, with the aim of exploring the connectivity structure within subject brains and across groups of subjects with different traits in order to identify possible meaningful and significant patterns.

The data comes from a pilot study of the Enhanced Nathan Kline Institute-Rockland Sample project; it comprises multimodal imaging data and subject-specific covariates for $n = 24$ subjects, for 12 of which 2 scan-rescan imaging sessions are available. A detailed description of the project, scopes, and technical aspects can be found at http://fcon_1000.projects.nitrc.org/indi/enhanced/. The pilot study includes three data sources:

- **Structural networks**: These data measure the anatomical interconnections—made by white matter fibers—among brain regions of interest, and are collected from DTI.
- **Dynamic functional activity**: These data measure the dynamic activity of each brain region through changes in the blood-oxygen-level dependent (BOLD) signal during resting state fMRI (R-fMRI) sessions.
- **Functional networks**: These data measure synchronization in brain activity for each pair of brain regions, and are obtained from the correlation in dynamic functional activity.

Some missing data are present in the dataset: the Dynamic functional activity for 2 subjects and the Structural networks for 4 subjects were not collected. Additionally, subject-specific information related to age, whether she/he is left-handed, right-handed or ambidextrous and her/his current and lifetime mental disorder were available only for 20 samples, impacting the performance evaluation of the method proposed in Sect. 7.

In Sect. 2 the necessary framework is introduced and Functional Data Analysis methods [19] are employed for obtaining the main data object of our analysis: a set of 22 functional networks numerically represented as correlation matrices. Subsequently, a proper distance metric for the aforementioned objects is considered for performing cluster analysis, as reported in Sect. 3. Section 4 considers a low dimensional representation of the data objects, and comparison with the results obtained by the clustering method is addressed. Section 5 reports a formal permutation procedure to test the equality of the mean functional networks between the two groups determined in Sect. 3. In order to identify possible different sources of variation a thorough study of the eigenstructure for the two mean functional networks is reported in Sect. 6. Section 7 considers a possible solution to account for spatial dependence between Dynamic functional activity of different regions, performing data fusion for the subset of subjects for which both Structural networks and Dynamic functional activity are available.

2 Curves and Correlation Matrices as Data Objects

Let us first consider the fMRI signal from the first scan. The data consist of 70 signals for each of the 24 subjects, corresponding to the BOLD activity of the 70 brain regions described by the Desikan Atlas [5]. Over the past decades the number of fMRI studies has increased exponentially [20], fostering the development of several methods for the analysis and interpretation of resting-state fMRI data, such as seed-based correlation analysis, independent component analysis and network-based models [4]. We propose to employ a Functional Data Analysis approach for performing the analysis, considering each signal as a realization of a stochastic process $X(t_i)$ sampled at times t_i, where $i = 1, \ldots, 403$; the last instant of time was not recorded for several patients and therefore it was not considered in the analysis. Subjects are

sampled at the same time schedule, so that registration is not deemed to be necessary [19]. Two subjects are not considered in the following analysis because of missing data, namely patient with ID 1 and patient with ID 21.

When dealing with functional data the usual starting point is to represent the data observed on a finite grid of points as functions. This part of the analysis is called smoothing, and there are several approaches to do it. Two important classes of smoothers are represented by kernel smoothing and orthogonal basis [19]. Both approaches sharing the idea of filtering out the short-time variation while keeping the global shape of the signal: we employ the latter for pre-processing the fMRI data. Orthogonal basis smoothing relies on the fact that, given an orthogonal basis for the space of interest, every function can be represented as an infinite linear combination of bases. A truncated version of the infinite sum provides a continuous representation of the discrete signal and reduces the dimensionality of the problem. The following analyses are based on a Fourier expansion, a standard choice in signal processing literature, with 100 bases. As it happens, it is not clear whether the short-time oscillations can be treated as noise or they might be related to some specific conditions of the brain. Future work might consider more appropriate bases such as wavelets [13] or Hierarchical Component Analysis [23]. An example of a smoothed function and its residuals for a given brain area and subject is reported in Fig. 1. A first interesting question that arises from the smoothing process would be to understand whether the residuals of the smoothing have some kind of clinical interpretation. We now have 70 functions for each of the 22 subjects, each function related to a different brain region. We used these functions to construct a correlation matrix between regions for each of the subjects. More in detail, we can compute the correlation between pairs of functions for every subject

$$\mathrm{Cor}(f_i, f_j) = \frac{\langle f_i, f_j \rangle}{||f_i|| ||f_j||} = \cos(\theta_{ij}), \quad \text{for } i, j = 1, \ldots, 70. \tag{1}$$

Here $\langle \cdot, \cdot \rangle$ denotes the inner product

$$\langle f_i, f_j \rangle = \int_{\Omega} f_i f_j d\mu \tag{2}$$

on the Hilbert space $\mathbb{L}^2(\Omega, \mathscr{B}, \mu)$, where $\Omega = [0, 403]$, $\mathscr{B}$ is the Borel σ-algebra of $[0, 403]$ and μ the Lebesgue measure. The norm $|| \cdot || = \langle \cdot, \cdot \rangle^{1/2}$ is induced by the inner product in (2). For a more detailed treatment of the underlying Hilbert space theory for functional data analysis, see [9, 10]. Processing the functional signal through the operator defined in (1) results in a 70×70 correlation matrix for each of the 22 subjects in the study. In Fig. 2 a subset of the so computed correlations matrices are graphically represented as heatmaps.

There is a clear difference in terms of correlation magnitude amongst subjects. Particularly, it seems that a subgroup of patients (ID 7, 10, 15, 17, 20, 22 and 23) present a much higher positive correlation between brain regions than the ones recorded for the rest of the subjects, visible by the overall darker blue areas in the correlation

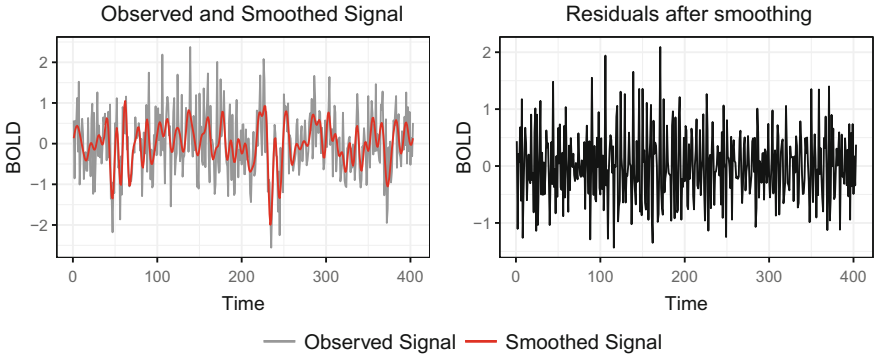

Fig. 1 Observed and smoothed signals (left plot) and residuals after the Fourier basis approximations (right plot) for a given subject and brain area

Fig. 2 Heatmaps of the first scan fMRI signal correlation matrices for 3 patients in the study, each belonging to a different subgroup identified by visual exploration of the magnitudes in the matrices

plots. Another interesting pattern visible in some patients (ID 6, 12, 14, 17 and 19 primarily) is given by the presence of a specific brain region, namely *rh-frontalpole*, that is negatively correlated to the remaining areas identified by the Desikan Atlas. A third behaviour that emerges from the visual exploration of the plots in Fig. 2 is the mild negative correlation and/or almost absence of correlation for two brain areas with the others for some subjects (ID 7, 8, 11, 14, 15, and 23). Particularly, these two areas are *lh-frontalpole* and *lh-temporalpole*. Lastly, there are two subjects (ID 2 and 23) that present almost individual patterns in the correlation structure between brain regions.

3 Clustering of Functional Networks

It is of interest to verify the presence of groups of subjects with similar brain activity, employing appropriate statistical methods given the complex structure of the objects under analysis. That is, the aim is to define a suitable distance concept in order to

characterize proximity amongst objects and subsequently perform cluster analysis according to the provided metric.

Given the considered context we cannot embed our objects of interest, i.e., the aforementioned correlation matrices, in a classical Euclidean space. Particularly, the correlation matrices represented in Fig. 2 are finite-dimensional approximations of a rescaled covariance operator for functional random processes, and therefore a suitable inference framework must be considered. Given a random function f taking values in $\mathbb{L}^2(\Omega)$ we define the covariance operator C_f for $g \in L^2(\Omega)$:

$$C_f g(t) = \int_\Omega E([f(t') - E(f(t'))][f(t) - E(f(t))]g(t')dt'. \tag{3}$$

For a review of definitions and theoretical properties of operators on $L^2(\Omega)$ see [2].

Denote with $PD(p)$ the space of positive semi-definite symmetric matrices of dimension p, that is the set of real symmetric matrices having non-negative eigenvalues [1]. We recall that $PD(p)$ is not a vector space and an inner product is not defined; it is however a Riemaniann manifold in which we can define a distance. For a detailed list of non-Euclidean distances for covariance matrices, see for example [6]. However, in a context of functional data, infinite dimensional extension of metrics for positive-semidefinite matrices must be used. Employing the inferential framework for covariance operators introduced in [17] we are able to extend the matrix-based distances to the functional case.

With the aim of measuring synchronization in brain activity and their respective dissimilarities among patients we consider the functional extension of the square root distance between variance covariance matrices, firstly defined in [6]. That is, given two covariance operators S_1 and S_2 their square root distance is defined as

$$d_R(S_1, S_2) = ||S_1^{1/2} - S_2^{1/2}||_{HS} \tag{4}$$

where $||\cdot||_{HS}$ denotes the Hilbert-Schmidt norm, generalization of the Frobenius norm for finite-dimensional matrices. Among the available matrix-based distances extendable to the functional case we decided to consider (4) since it takes into account the full eigenstructure of the covariance operator [17]. The definition of a proper distance is directly linked to the introduction of a mean value concept, given the chosen distance. Particularly, letting $S_1, \ldots, S_n$ be a sample of independent covariance operators we define its sample *Fréchet mean* based on the square root distance (4) as

$$\hat{\Sigma} = \hat{\Delta}\hat{\Delta}^\top \tag{5}$$

where

$$\hat{\Delta} = \underset{\Delta}{\operatorname{arginf}} \left\{ \sum_{i=1}^{n} ||S_i^{1/2} - \Delta||_{HS}^2 \right\} = \frac{1}{n} \sum_{i=1}^{n} S_i^{1/2}. \tag{6}$$

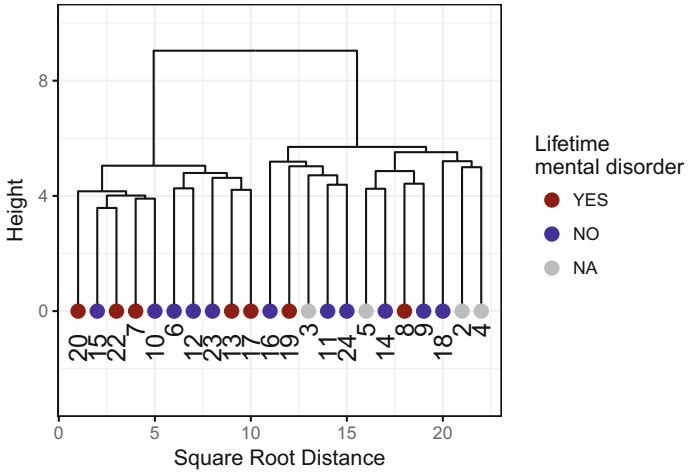

Fig. 3 Dendrogram from hierarchical clustering with Ward agglomeration method to the functional network data. The dendrogram highlights the presence of two main clusters not seemingly related to patients' mental health status

For proofs and discussion related to the consistency of the sample Fréchet mean based on the square root distance, refer to [11].

Making use of the square root distance defined in (4) we proceed in trying to identify possible presence of groups amongst the data objects employing a distance-based clustering algorithm. Particularly, the analysis was carried out considering hierarchical clustering with Ward agglomeration method [15].

The result of the clustering algorithm is graphically presented in Fig. 3: the dendrogram clearly highlights the presence of two different clusters in our sample of patients. The groups however do not seem to be separated along the additional information on the subjects provided in the study. Therefore, even though the difference between the mean correlation matrices of the two groups results to be statistically significant (see Sect. 5), interpretation explaining the groupings remains still unclear. A clinician assessment, together with a thorough consideration of the medical history of each patient involved in the study would provide insight on groups interpretability and classification.

In the upcoming section, the problem of finding homogeneous groups amongst functional networks is differently tackled employing a non-linear dimensionality reduction technique. Both methodologies agree in terms of identified number of groups and groups structure.

4 Low Dimensional Representation

In order to obtain a low dimensional representation of the correlation matrices a Local Linear Embedding (LLE) algorithm [21] is considered. This method is based on a simple geometric intuition. Suppose the data consist of N real-valued vectors X_i, each of dimensionality D, sampled from some smooth underlying manifold. We expect each data point and its neighbours to lie on or close to a locally linear patch of the manifold. We can characterize the local geometry of these patches by linear coefficients that reconstruct each data point from its neighbours. In the first step of the algorithm one identifies K nearest neighbours per data point, as measured by Euclidean distance. In the second step the weights W_{ij} that best reconstruct each data point X_i from its neighbours are computed, minimizing

$$\sum_{i=1}^{n} \left(X_i - \sum_{j=1}^{K} W_{ij} X_j \right)^2 .$$

The weights W_{ij} summarize the contribution of the j-th data point to the i-th reconstruction. Finally we can compute the vectors Y_i of low dimensional coordinates, $d < n$, best reconstructed by the weights W_{ij}, minimizing

$$\sum_{i=1}^{d} \left(Y_i - \sum_{j=1}^{K} W_{ij} X_j \right)^2 .$$

This cost function—like the previous one—is based on locally linear reconstruction errors, but here we fix the weights W_{ij} while optimizing the coordinates Y_i. In Fig. 4 we can see the two dimensional representations ($d = 2$), for different number of neighbours (from 3 to 11, starting from the left upper corner). The triangles represent the patients with lifetime disease, while the colour represents the groups identified by hierarchical clustering. We firstly note that the algorithm is robust with respect to the choice of the hyper-parameter K. Secondly, and more relevant for the scope of our analysis, we recognize that in all the considered representations 5 out of the 7 subjects with lifetime disease are in the red group and the remaining 2 in the black group (these are patients labelled with ID 8 and 19 respectively). We can also note that the low dimensional representation preserves the structure of the original space and the separation performed by the hierarchical clustering is still clearly visible: an average Adjusted Rand Index of 0.96 between the groupings found with the two methods, varying K from 3 to 11 in the LLE, is obtained. A formal permutation test for statistically assessing the significant difference between the two sub-populations identified by both hierarchical clustering and LLE is developed in the upcoming section.

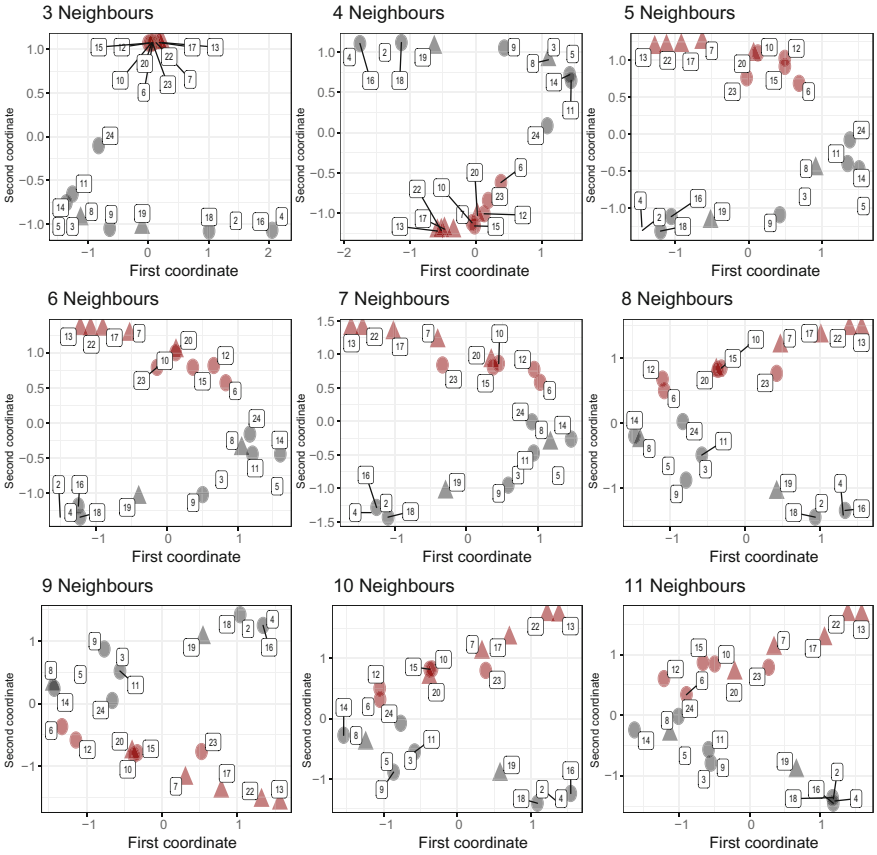

Fig. 4 2-dimensional representation of the correlation matrices through Local Linear Embedding algorithm for different number of neighbours K. Triangles represent patients with lifetime disease; colours represent the groups identified by hierarchical clustering of Sect. 3

5 Hypothesis Testing for Correlation Structures

Let us consider the two groups of patients identified in the previous Sections. We want to verify whether the functional activity, recorded in terms of 70×70 correlation matrices for each of the 22 subjects, is significantly different in the 2 groups. We assume that our two samples are such that $S_1^{(1)} \cdots S_{n_1}^{(1)}$ are random $PD(p)$ matrices with expectation $E(S_i) = \Sigma_1$, $i = 1, \ldots, n_1$ and $S_1^{(2)} \cdots S_{n_2}^{(2)}$ are random $PD(p)$ matrices with expectation $E(S_j) = \Sigma_2$, $j = 1, \ldots, n_2$. $S_1^{(1)} \cdots S_{n_1}^{(1)}$ and $S_1^{(2)} \cdots S_{n_2}^{(2)}$ are the sample correlation matrices belonging to the first and second group respectively. Particularly, in our context $n_1 = 12$ and $n_2 = 10$ with patients

(2, 3, 4, 5, 8, 9, 11, 14, 16, 18, 19, 24) belonging to the first group and patients (6, 7, 10, 12, 13, 15, 17, 20, 22, 23) to the second group (see Fig. 3). We would like to test

$$H_0 : \Sigma_1 = \Sigma_2 \quad \text{versus} \quad H_1 : \Sigma_1 \neq \Sigma_2 \,.$$

To test these hypotheses we follow a permutational approach along the methods advanced in [16, 17]. We reformulate the test in terms of square root distances between covariance objects: the considered test statistic is $d(\hat{\Sigma}_1, \hat{\Sigma}_2)$ where $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ denotes the Fréchet mean as defined in (6) for the samples in the two groups. H_0 is rejected for large values of $d(\hat{\Sigma}_1, \hat{\Sigma}_2)$. The test is simply a two way ANOVA, but equipped with a proper metric and consequently with a proper definition of sample mean. If H_0 is true, complete exchangeability of the random variables generating the sample observations holds and therefore, in order to approximate the distribution of the test statistic under H_0, the two samples are pooled together and randomly assigned to the two groups preserving sample sizes. The test consists in a comparison of $d(\hat{\Sigma}_1, \hat{\Sigma}_2)$ with M random permutations computed via Monte Carlo of $d(\hat{\Sigma}_1^{(m)}, \hat{\Sigma}_2^{(m)})$, $m = 1, \ldots, M$; where $\hat{\Sigma}_i^{(m)}$ is the sample mean correlation matrix for group i in permutation m. The p-value with $M = 100$ permutations is less than 0.01, with a difference of the two sample means of 2.41. Thus, we conclude that the two sub-populations have statistically different correlation matrices, confirming and validating the results previously highlighted by the clustering and LLE methods. The same permutation test had been initially applied to groups clustered by subjects characteristics; notwithstanding, none of the additional information available for the subjects under study (age, handedness, current/lifetime mental disorder) have been proved significant in distinguishing different groups. Figure 5 shows the heatmaps of the sample mean correlation matrices of the two considered groups. The difference between the two is clear, with higher correlation values in the second group.

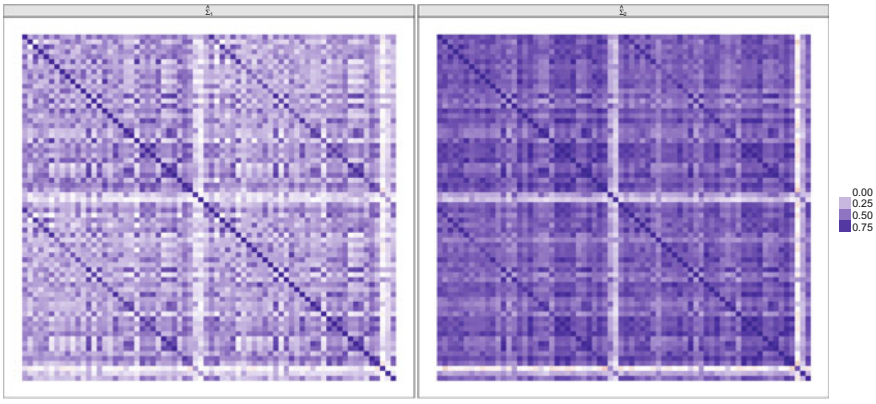

Fig. 5 Heatmap of the sample mean correlation matrices in the two identified groups

6 Eingenstructure of the Mean Correlation Matrices

In this Section a comprehensive analysis of the main sources of variability for the two sample mean correlation matrices identified in Sect. 3 will be addressed. Figure 6 displays the generalized variance in subspaces of increasing dimension for the two correlation matrices, defined as the cumulative product of their eigenvalues [7].

The generalized variance is proportional to the square of the volumes of the hyper-ellipsoids projected onto the principal components subspaces. It is clearly visible that the first group is characterized by a much larger generalized variance, supporting the significant difference between the two groups highlighted by the permutation test.

A spectral decomposition of the two sample mean correlation matrices is reported in Fig. 7. Since we are considering correlation matrices, the employed terminology comes from the Principal Components Analysis literature [12]. Particularly, the variance denotes the magnitude of the different eigenvalues whereas the contribution to the total variability is calculated dividing the cumulative sum of the eigenvalues by their total. The magnitude of the eigenvalues in the first group decreases more slowly than in the second group, as it was already apparent in Fig. 6. Five components account for 80% of the total variability in $\hat{\Sigma}_2$, whereas for $\hat{\Sigma}_1$ nine components are necessary for achieving the same contribution.

In order to check whether the source of variability is different in the two groups, the components (loadings) of the first six eigenvectors for the two sample mean covariance matrices are plotted in Fig. 8. As it can be seen from the graphs, the source of variability seems different, especially considering the first three loadings. This is further highlighted by the graphical representation of the 3-D spatial coordinates for

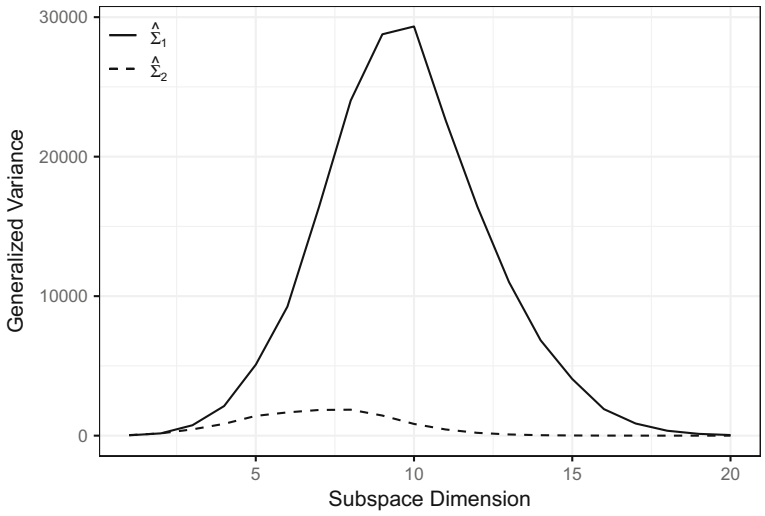

Fig. 6 Generalized variance for the two sample mean correlation matrices, defined as the cumulative product of their eigenvalues

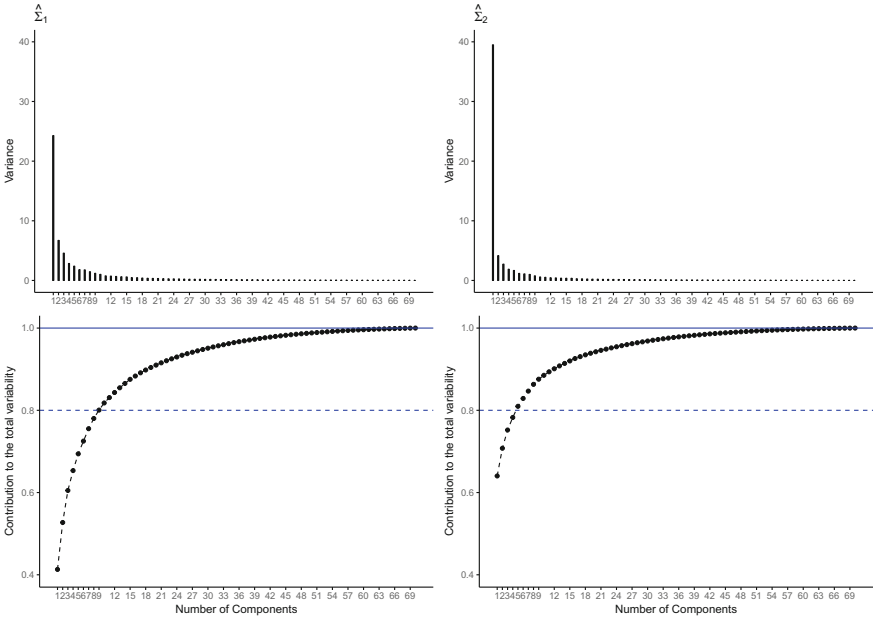

Fig. 7 Eigenvalues and relative cumulative sum of eigenvalues for the mean correlation matrices of the two groups identified in the patients sample

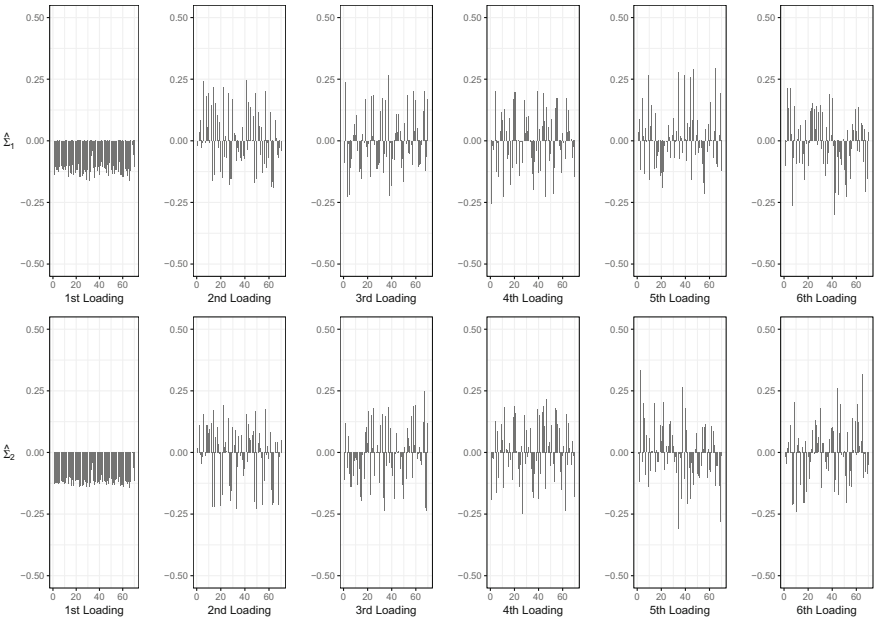

Fig. 8 Entries of the first 6 loadings vectors (eigenvectors) for the two sample mean correlation matrices of the two groups identified in the patients sample

the centroids of the brain regions reported in Fig. 9, where centroids are coloured according to the first, second and third loading vectors entries respectively. A spatial pattern seems to be present in the matrices eigenstructure. The present analysis motivates and justifies the novel approach introduced in the upcoming section, where we attempt to account for the spatial dependence employing a re-weighted version of the functional network. Particularly, the structural network (i.e., the count of number of white fibers that connect each brain region) is interpreted as a measure of proximity between the brain regions.

7 Spatial Dependence for Functional Networks

So far, we have only considered the fMRI data corresponding to the 70 regions of the Desikan atlas as independent. Nonetheless, the spatial dependence has not been filtered out during pre-processing and it is therefore reasonable to suppose that some sort of spatial dependence is still present in the registered signals, as it can be graphically seen in Fig. 9. A possible procedure for incorporating the spatial dependence within our analysis framework would be to exploit the information contained in the structural networks available for each patient. The structural networks contain the total number of white matter fibers connecting each pair of brain regions for each subject. The aforementioned structure can be interpreted as an adjacency matrix: the intuition behind this definition is that the more white fibers connecting a pair of brain regions the closer the two brain regions can be considered. Particularly, the functional networks (identified by the correlation matrices employed in the previous Sections) can be re-weighted according to the magnitude enclosed in the structural networks, subject-wise. Considering only the first scan, indicate with d_{uv} the count of how many white matter fibers are found to connect brain regions u and v for a specific subject. We define the symmetric 70×70 weight matrix W induced by the structural network for each subject having entries as follows:

$$w_{uv} = \begin{cases} 1 & u = v \\ d_{uv} / \left(\sum_{u=1}^{70} \sum_{v=1}^{70} d_{uv} \right) & u \neq v \end{cases} \tag{7}$$

Subsequently, we define the *re-weighted functional networks* R as the Hadamard product between W and the functional networks computed in (1). For obtaining R both structural and functional networks must be available, therefore it was possible to calculate the re-weighted functional networks only for 18 out of 24 patients present in the study. Notice that R is still a symmetric and positive semi-definite matrix thanks to Schur product theorem [22]. Employing the same methodology described in Sect. 3 we perform hierarchical clustering on the re-weighted functional networks: the dendrogram of the clustering procedure is reported in Fig. 10. Likewise in the previous analysis the dendrogram highlights the presence of two different clusters, with a significant difference in their mean correlation matrices (the test in Sect. 5 was

Fig. 9 Graphical representation of the 3-D spatial coordinates for the centroids of the brain regions, under different 2-D views. Colour intensity is associated with the entries of the first (top), second (middle) and third (bottom) loading of $\hat{\Sigma}_1$ (first rows) and $\hat{\Sigma}_2$ (second rows) respectively. Shapes describe hemisphere membership

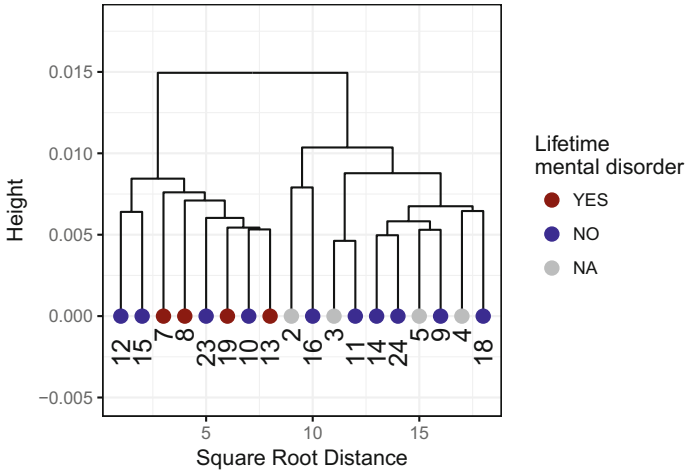

Fig. 10 Dendrogram from hierarchical clustering with Ward agglomeration method to the re-weighted functional network data. The dendrogram highlights the presence of two main clusters, partially related to patients' mental health status

repeated and the null hypothesis rejected). In addition, the two identified groups seem at least partially related to the presence or absence of lifetime mental disorder for the available set of patients. Although the sample size is very small, we empirically evaluate the main source of dissimilarity between the two groups of patients with and without lifetime mental disorder considering their mean re-weighted functional networks, computed using (5). The preeminent differences are due to the higher weighted correlations found for patients with lifetime mental disorder between areas *lh - posteriorcingulate* and *lh - corpuscallosum*, *rh - posteriorcingulate* and *lh - posteriorcingulate*, *rh - superiorfrontal* and *rh - caudalmiddlefrontal*, *rh - corpuscallosum* and *lh - posteriorcingulate*, compared to the weighted correlations in these areas for patients with absence of lifetime mental disorder.

8 Conclusions and Future Research Directions

The present work is the result of a 48 h workshop during which the authors, guided by their senior group leader Piercesare Secchi, were asked to propose original statistical methods for data analysis in neuroscience [3]. We applied several techniques from Object Oriented Data Analysis literature for exploring data coming from the Enhanced Nathan Kline Institute-Rockland Sample project. Three different clustering methods are proposed for the fMRI data, with the last and most promising one involving the processing of both structural and functional network for each patient.

Our approach began with the identification of two clusters in the space of the correlation matrices of the smoothed fMRI signals. These two groups corresponded only partially to the labelling concerning the presence/absence of mental disease. A non-linear dimensional reduction technique helped us to visualize the clusters: the two sub-populations structure is clear in the identified subspace. The difference between the two groups, formalized through a statistical test in which the null hypothesis was the equality of the two mean correlation matrices, is highly significant. A deeper analysis of the eigenstructure of the two mean correlation matrices highlighted the differences in the sources of variability in the two groups, together with a possible spatial dependence in the data objects. Lastly, an attempt at performing data fusion weighting the functional networks with the structural networks is addressed: promising initial results seem to have been achieved. In particular, employing re-weighted functional networks, subjects with confirmed presence of mental disease are more clearly separated from patients with absence of mental disease, fostering the employment of the aforementioned procedure whenever functional and structural networks are available. Nevertheless, both a larger sample size as well as knowledge domain would be necessary for establishing and interpreting the described discoveries.

The StartUp Research workshop has been a challenging yet enriching and unforgettable experience, in which we had the chance to meet, connect and learn from our peers, colleagues and senior mentors. We early-career researchers had a direct experience on the essential importance of interaction and knowledge-sharing which, ultimately, lead to knowledge creation.

Acknowledgements We acknowledge Greg Kiar and Eric Bridgeford from NeuroData at Johns Hopkins University, who pre-processed the raw DTI and R-fMRI imaging data available at http://fcon_1000.projects.nitrc.org/indi/CoRR/html/nki_1.html. We would like to deeply thank the StartUp Research Scientific Committee for efficiently and flawlessly organizing such a motivating experience. We thank Professor Francesca Greselin and Doctor Mauro Ceroni for their support and help throughout the drafting of this manuscript.

References

1. Amari, S.I.: Differential-geometrical methods in statistics. Lecture Notes in Statistics, vol. 28. Springer, New York (1985)
2. Bosq, D.: Linear Processes in Function Spaces. Lecture Notes in Statistics. Springer, New York (2000)
3. Canale, A., Durante, D., Paci, L., Scarpa, B.: Connecting statistical brains. Significance **15**(1), 38–40 (2018)
4. Cole, D.M., Smith, S.M., Beckmann, C.F.: Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. Front. Syst. Neurosci. **4**, 8 (2010)
5. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage **31**(3), 968–980 (2006)
6. Dryden, I.L., Koloydenko, A., Zhou, D.: Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann. Appl. Stat. **3**(3), 1102–1123 (2009)

7. Friendly, M., Monette, G., Fox, J.: Elliptical insights: understanding statistical methods through elliptical geometry. Stat. Sci. **28**(1), 1–39 (2013)
8. Heuvel, M.P.V.D., Pol, H.E.H.: Exploring the brain network: a review on resting-state fMRI functional connectivity. Eur. Neuropsychopharmacol. **20**(8), 519–534 (2010)
9. Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer Series in Statistics. Springer, New York (2012)
10. Hsing, T., Eubank, R.: Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics. Wiley, Chichester, UK (2015)
11. Huckemann, S.: Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. Ann. Stat. **39**(2), 1098–1124 (2010)
12. Jolliffe, I.T.: Principal Component Analysis and Factor Analysis, pp. 115–128. Springer, New York (1986)
13. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press (1999)
14. Marron, J.S., Alonso, A.M.: Overview of object oriented data analysis. Biom. J. **56**(5), 732–753 (2014)
15. Murtagh, F., Legendre, P.: Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm. J. Classif. **31**(3), 274–295 (2011)
16. Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data: Theory, Applications and Software. Wiley (2010)
17. Pigoli, D., Aston, J.A.D., Dryden, I.L., Secchi, P.: Distances and inference for covariance operators. Biometrika **101**(2), 409–422 (2014)
18. Plis, S., Meinecke, F.C., Eichele, T.: Analysis of multimodal neuroimaging data. IEEE Rev. Biomed. Eng. **4**, 26–58 (2011)
19. Ramsay, J., Silverman, B.W.: Functional Data Analysis. Springer Series in Statistics. Springer, New York (2005)
20. Roalf, D., Gur, R.: Functional brain imaging in neuropsychology over the past 25 years. Neuropsychology **31**(8), 954–971 (2017)
21. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
22. Schur, J.: Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. J. Reine Angew. Math. **140**, 1–28 (1911)
23. Secchi, P., Vantini, S., Zanini, P.: Hierarchical independent component analysis: a multi-resolution non-orthogonal data-driven basis. Comput. Stat. Data Anal. **95**, 133–149 (2016)
24. Wang, H., Marron, J.S.: Object oriented data analysis: sets of trees. Ann. Stat. **35**(5), 1849–1873 (2007)