Consensus clustering approach to group brain connectivity matrices

Javier Rasero[1,2,3], Mario Pellicoro[2], Leonardo Angelini[2,3,4],
Jesus M Cortes [1,5], Daniele Marinazzo[6], and S. Stramaglia[2,3,4]

[1] *Biocruces Health Research Institute. Hospital
Universitario de Cruces. E-48903, Barakaldo, Spain.*

[2] *Dipartimento di Fisica, Universitá degli Studi "Aldo Moro" Bari, Italy*

[3] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy*

[4] *TIRES-Center of Innovative Technologies for Signal Detection and Processing,
Universitá degli Studi "Aldo Moro" Bari, Italy*

[5] *Ikerbasque, The Basque Foundation for Science,
E-48011, Bilbao, Spain.*

and

[6] *Faculty of Psychology and Educational Sciences,
Department of Data Analysis, Ghent University,
Henri Dunantlaan 1, B-9000 Ghent, Belgium*

(Dated: December 13, 2016)

Abstract

A novel approach rooted on the notion of *consensus* clustering, a strategy developed for community detection in complex networks, is proposed to cope with the heterogeneity that characterizes connectivity matrices in healthy and disease. The method can be summarized as follows: (i) define, for each node, of a distance matrix for the set of subjects (ii) cluster the distance matrix for each node, (iii) build the consensus network from the corresponding partitions and (iv) extract groups of subjects by finding the communities of the consensus network thus obtained. Applications on a toy model and two real data sets, show the effectiveness of the proposed methodology, which represents heterogeneity of a set of subjects in terms of a weighted network, the consensus matrix.

PACS numbers: 42.30.Sy,87.57.-s,87.19.L-,87.19.lf

In the supervised analysis of human connectome data [1, 2], subjects are usually grouped under a common umbrella corresponding to high-level clinical categories (e.g., patients and controls), and typical approaches aim at deducing a decision function from the labeled training data, see e.g. [3]. However, the population of healthy subjects (as well as those of patients) is typically highly heterogeneous: clustering algorithms find natural groupings in the data, and therefore constitute a promising technique for disentangling heterogeneity that is inherent to many diseases and to the cohort of controls. Such an unsupervised classification may also be used as a preprocessing stage, so that the subsequent supervised analysis might exploit the knowledge of the structure of data. Some studies dealt with similar issues: semi-supervised clustering of imaging data was considered in [4, 5], other recent approaches cope with the heterogeneity of subjects using multiplex biomarkers techniques [6] and combinations of imaging and genetic patterns [7], whilst a strategy to overcome inter-subject variability while predicting predicting behavioral variable from imaging data has been proposed in [8]. Connectivity features have been used in data-driven approaches for analysis and classification of MRI data in [9, 10] The purpose of this work is to introduce a novel approach that is rooted on the notion of *consensus* clustering [11], a strategy developed for community detection in complex networks [12].

To introduce our method, let us assume that a connectivity matrix is associated to each item to be classified (usually a subject, but also individual scans for the same subject as in the example illustrated below). The goal of supervised analysis is to mine those features of matrices which provide the best prediction of available environmental and phenotypic factors, such as task performance, psychological traits, and disease states. When it comes to using unsupervised analysis of matrices to find groups of subjects, the most straightforward approach would be to extract a vector of features from each connectivity matrix, and to cluster these vectors using one of the commonly used clustering algorithms. The purpose of the present work is to propose a new strategy for unsupervised clustering of connectivity matrices. In the proposed approach the different features, extracted from connectivity matrices, are not combined in a single vector to be fed into the clustering algorithm; rather, the information coming from the various features are combined by constructing a *consensus* network [11]. Consensus clustering is commonly used to generate stable results out of a set of partitions delivered by different clustering algorithms (and/or parameters) applied to the same data; here, instead, we use the consensus strategy to combine the information, about the data structure, arising from different features so as to summarize them in a single consensus matrix.

The unsupervised strategy that we propose here to group subjects, without using phenotypic measures, may be summarized as follows, see figure (1): (i) definition, for each node, of a distance matrix for the set of subjects (ii) clustering the distance matrix for each node, (iii) build the consensus network from the corresponding partitions and (iv) extract groups of subjects by finding the communities of the consensus network thus obtained . We remark that the proposed approach not only provides a partition of subjects in communities, but also the consensus matrix, which is a geometrical representation of the set of subjects. In the next section we describe in detail the method and apply it to a toy model, then we show the application on two real MRI data sets. Finally, some conclusions are drawn.

I. METHOD

Let us consider m subjects whose functional (structural) $N \times N$ connectivity matrix [13], where N is the number of nodes, will be denoted by $\{\mathbf{A}_\alpha\}$, $\alpha = 1, \ldots, m$. For each node, we calculate the distance between connectivity patterns (i.e., each node's connectivity with the rest of the brain, given by the corresponding column in matrix $\mathbf{A}$). As the distance between a pair of connectivity patterns (i.e., between two subjects for a given node) we use $\sqrt{2(1-r)}$, where r is the corresponding Pearson correlation (other choices might be used). The $m \times m$ distance matrix corresponding to node i will be denoted by $\mathbf{D_i}$, with $i = 1, \ldots, N$. The set of $\mathbf{D}$ matrices may be seen as corresponding to layers of a multilayer network [14], each brain node providing a layer.

Each distance matrix $\mathbf{D_i}$ is then clustered using the k-medoids method [15] which performs a partition of subjects into k groups. Subsequently, an $m \times m$ consensus matrix $\mathbf{C}$ is evaluated: its entry $C_{\alpha\beta}$ indicates the number of partitions in which subjects α and β were assigned to the same group, divided by the number of partitions N. The number of clusters k may be kept fixed, thus rendering the consensus matrix depending on k; a better strategy, however, is to average the consensus matrix over k ranging in an interval, so as to fuse, in the consensus matrix, information about structures at different resolutions.

The consensus matrix, obtained as explained before, is eventually partitioned in communities by modularity maximization, or by any other algorithm for community detection in complex networks.

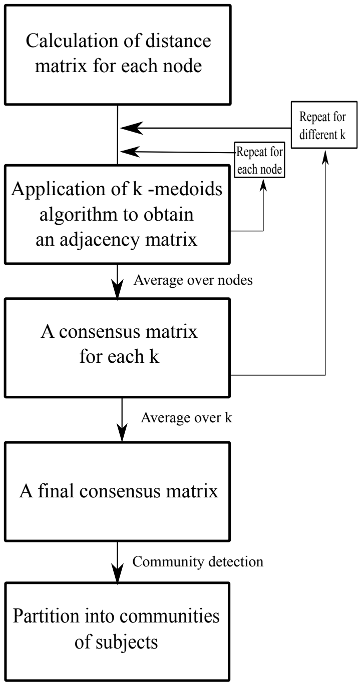
Calculation of distance matrix for each node
Repeat for different k
Repeat for each node
Application of k -medoids algorithm to obtain an adjacency matrix
Average over nodes
A consensus matrix for each k
Average over k
A final consensus matrix
Community detection
Partition into communities of subjects

FIG. 1: The flowchart of the proposed methodology.

We like to remark that the proposed approach has similarities with the approach in [16] where the techniques from genome wide association studies, copying with the problem of an huge number of comparisons, were applied to connectome data, thus identifying nodes whose whole-brain connectivity patterns vary significantly with a phenotypic variable. The approach in [16] consists in two steps. First, for each node in the connectome, a whole brain functional connectivity map is evaluated, and then the similarity between the connectivity maps of all possible pairings of participants, using spatial correlation, is calculated. Then, in the second stage, a statistics for each node is evaluated, indicating the strength of the relationship between a phenotypic measure and variations in its connectivity patterns across subjects. The main similarity with the proposed approach is that in both methods, for each node in the connectome, the comparison between the connectivity maps yields a distance matrix in the space of subjects.

II. A TOY MODEL

As a toy model to describe the application of our method, we simulate a set of 100 subjects, divided in four groups of 25 each. The subjects are supposed to be described by 30 nodes. We will compare our proposed approach with a standard procedure such as averaging the distance matrices and then applying the clustering algorithm to the average distance matrix.

The distance matrices corresponding to the first ten nodes are constructed in the following way: the distance for pairs belonging to the same group is sampled uniformly in the interval $[0.1, 0.4]$, whilst the distance for pairs belonging to different groups is sampled uniformly in the interval $[0.2, 0.4]$. The distance matrices corresponding to the twenty remaining nodes have all the entries sampled uniformly in the interval $[0.2, 0.4]$. It follows that in our toy model only 10 nodes, out of 30, carry information about the presence of the four groups.

First of all, we evaluate the distance matrix among subjects, averaged over the 30 nodes, and apply the k-medoids algorithm to this matrix , searching for $k = 4$ clusters (thus exploiting the knowledge of the number of classes present in data); this procedure leads to an efficiency of 0.89, measured as the fraction of subjects which are correctly assigned to groups.

Subsequently, we run the proposed approach by applying, separately to each distance matrix for each of the 30 nodes, the k-medoids algorithm with varying k. We then build the corresponding consensus matrix. For example in figure (2) the consensus matrix among subjects is depicted as obtained applying k-medoids with $k = 10$ separately to each of the 30 layers. Then, the communities of the consensus matrices have been estimated using the modularity optimization[17]

In figure (3) the efficiency of the partition, provided by modularity maximization on the consensus matrix, is depicted versus k, showing that the proposed method performs better that the partition of the average distance matrix for this example, for k greater than five. For k smaller than six the modularity maximization typically yields two or three groups; we remark that the efficiency 0.89 is reached by k-medoids on the average distance using $k = 4$ i.e. exploiting the knowledge of the number of groups present in the data set, whilst the modularity optimization algorithm determines both the number of clusters and the partition. Intuitively, the proposed approach works better in this example for large k, because in the

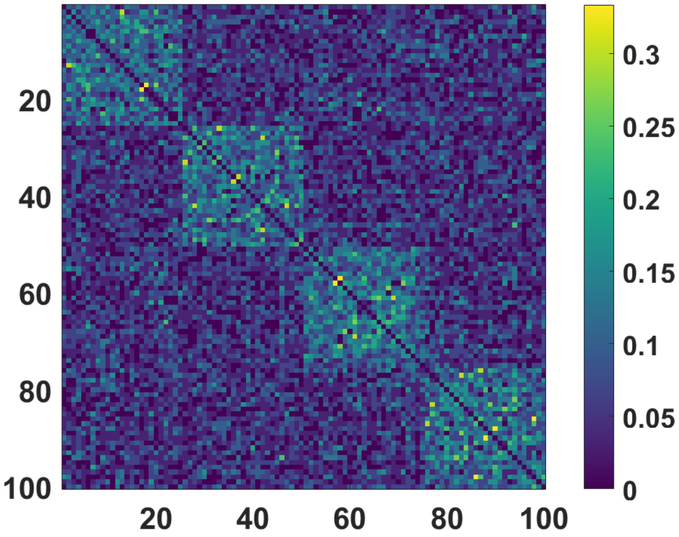

FIG. 2: Concerning the toy model, the consensus matrix among subjects, obtained applying k-medoids with $k = 10$ separately to each of the 30 layers.

distance matrix corresponding to a given node, due to the high level of randomness, the block corresponding to a group is seen as fragmented in smaller pieces; those pieces can be retrieved using k-medoids with larger k. On the other hand when the consensus is made across the different nodes, all those pieces merge in the consensus matrix and build the block corresponding to the four groups.

It is also worth noting that the efficiency by clustering the averaged consensus matrix (over the values of k) is one, i.e. perfect group reconstruction. Averaging over the values of k appears then to be a convenient strategy. Moreover, averaging over values of parameters is a common strategy for consensus clustering, hence building the consensus matrix while joining several values of k is in line with the philosophy of consensus clustering [11].

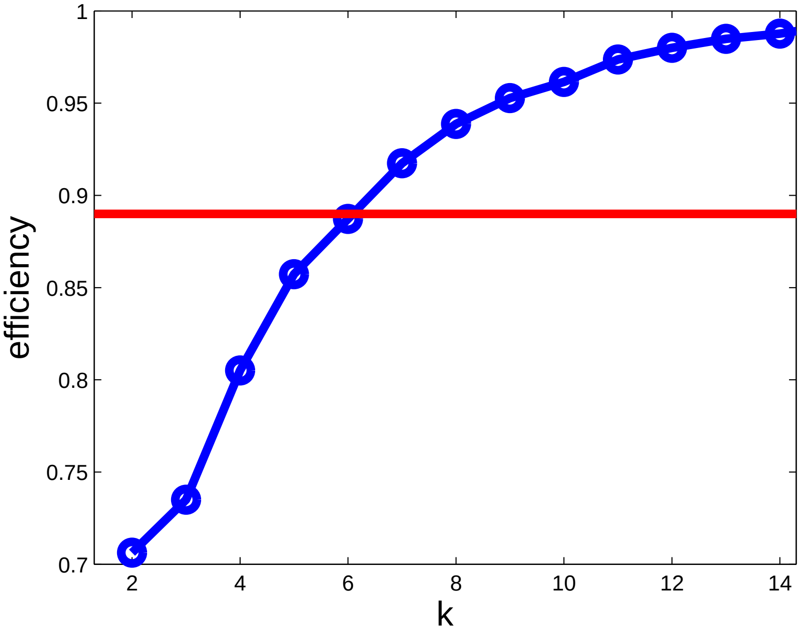

FIG. 3: The efficiency of the partition, provided by modularity maximization on the consensus matrix, is depicted versus k. The horizontal line represents the efficiency obtained by clustering the average distance matrix using k-medoids and $k = 4$.

III. APPLICATION TO REAL DATA SETS

A. Longitudinal data set

Growing interest is devoted to longitudinal phenotyping in cognitive neuroscience: accordingly we consider here data from the MyConnectome project [18, 19], where functional connectivity from a single human, over a period of 18 months, was recorded. In [20] the presence of two distinct temporal states has been identified, that fluctuated over the course of 18 months. These temporal states were associated with distinct patterns of time-resolved blood oxygen level dependent (BOLD) connectivity within individual scanning sessions and also related to significant alterations in global efficiency of brain connectivity as well as differences in self-reported attention. The data we process consists of 89 sessions each resulting in a 268×268 functional network. This data was obtained from the OpenfMRI database. Its accession number is ds000031. The functional MRI (fMRI) data was preprocessed with FSL (FMRIB Software Library v5.0). The first 10 volumes were discarded for correction of the magnetic saturation effect. The remaining volumes were corrected for motion, after which slice timing correction was applied to correct for temporal alignment. All voxels were spatially smoothed with a 6mm FWHM isotropic Gaussian kernel and after intensity normalization, a band pass filter was applied between 0.01 and 0.08 Hz. In addition, linear and quadratic trends were removed. We next regressed out the motion time courses, the average CSF signal and the average white matter signal. Global signal regression was not performed. Data were transformed to the MNI152 template, such that a given voxel had a volume of 3mm x 3 mm x 3mm. Finally we obtained 268 time series, each corresponding to an anatomical region of interest (ROI), by averaging the voxel signals according to the functional atlas described in [21].

Each of the 89 sessions resulted in a 268×268 matrix of Pearson correlation coefficients.

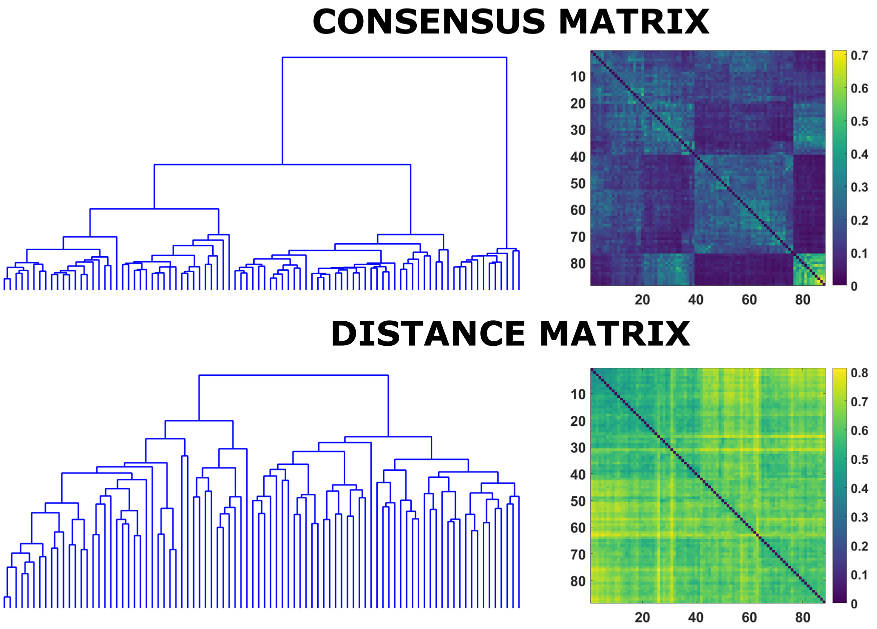

FIG. 4: (Top) Concerning the *MyConnectome* data set, the consensus matrix, obtained averaging over k, by the proposed approach is displayed with nodes ordered according to hierarchical clustering, with the corresponding dendrogram. (Bottom) The average distance matrix, among the different sessions of the same subject, and the corresponding dendrogram.

We treated the sessions as if they were connectivity matrices of different subjects, and applied the proposed methodology. In figure (4) we depict the distance matrix, among the different sessions of the same subject, and the consensus matrix, obtained averaging over ten values of k. Sessions are ordered, in both cases, according to hierarchical clustering; the corresponding dendrogram are also shown in the figure. It is clear that the consensus matrix shows a hierarchical structure. Maximization of the modularity provides two communities with modularity equal to 0.175. As depicted in (5), the two communities are significantly different (Wilcoxon rank sum test, Bonferroni correction for multiple comparisons) for several PANAS scores associated to tiredness, thus confirming the presence of two distinct temporal states. However the hierarchical structure of the consensus matrix that we obtained suggests that longer longitudinal recordings are needed to further evidence the richness of distinct functional states for single subjects.

B. Resting healthy subjects, functional and structural connectivity

We consider 171 healthy subjects from the NKI Rockland dataset [22]; for each subject we use both the structural Diffusion Tensor Imaging DTI network and the functional network, already obtained from processed data as described in [23]. In this case the networks have 118 nodes. In figure (6) we depict the consensus matrix for both DTI and fMRI networks; modularity maximization yields two communities in each case, and the communities are characterized by different age with probability 9×10^{-10} for the structural and 7×10^{-4} for the functional. We remark that using k-medoids over the average distance, we obtain two groups with different age, t-test with probability 10^{-3} using the functional distance, whilst no significant difference in age using the structural connectivity.

Inspired by the results found by our method, we also performed a multivariate distance regression [16], that allowed us to build a pseudo F-statistics to test whether age correlates with the differences observed in the distance matrix for each node. We have achieved this by comparing the observed F-statistic with the pseudo F-distribution (that is not normal) after 10^5 data permutations. As expected, for both structural and functional data, we found 124 and 76 nodes statistically related with age respectively, thus suggesting that age might be one of the variables responsible of the community structure found by our method.

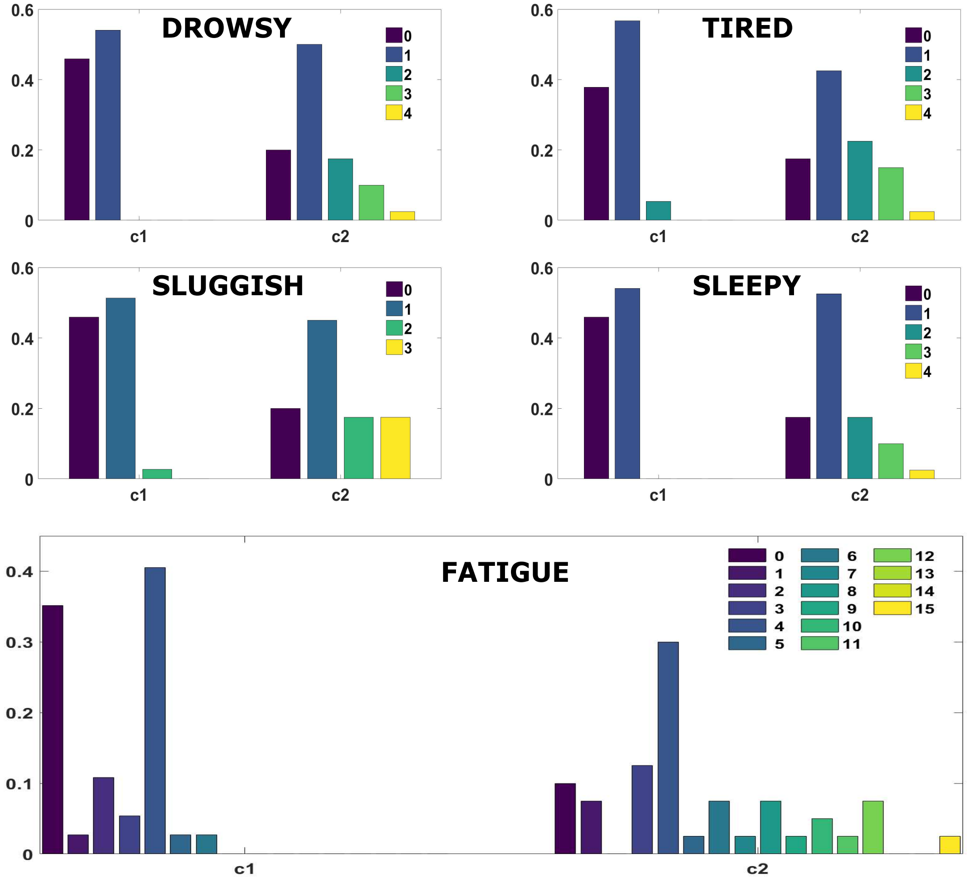

FIG. 5: Concerning the *MyConnectome* data set, the distributions of the score of the variables which are significantly different are depicted in the two communities found by modularity optimization on the consensus matrix provided by the proposed approach: *drowsy* (Bonferroni corrected p-value = 0.028), *tired* (Bonferroni corrected p-value = 0.041), *sluggish* (Bonferroni corrected p-value = 0.026), *sleepy* Bonferroni corrected p-value = 0.012), *fatigue* (Bonferroni corrected p-value = 0.022)

IV. CONCLUSIONS

An important issue such as dealing with the heterogeneity that characterizes healthy conditions, as well as diseases, requires the development of effective methods capable to highlight the structure of sets of subjects at varying resolutions. The approach that we propose here is applied to sets of subjects each described by a connectivity matrix; we propose a strategy, rooted in complex networks theory, to obtain a consensus matrix which

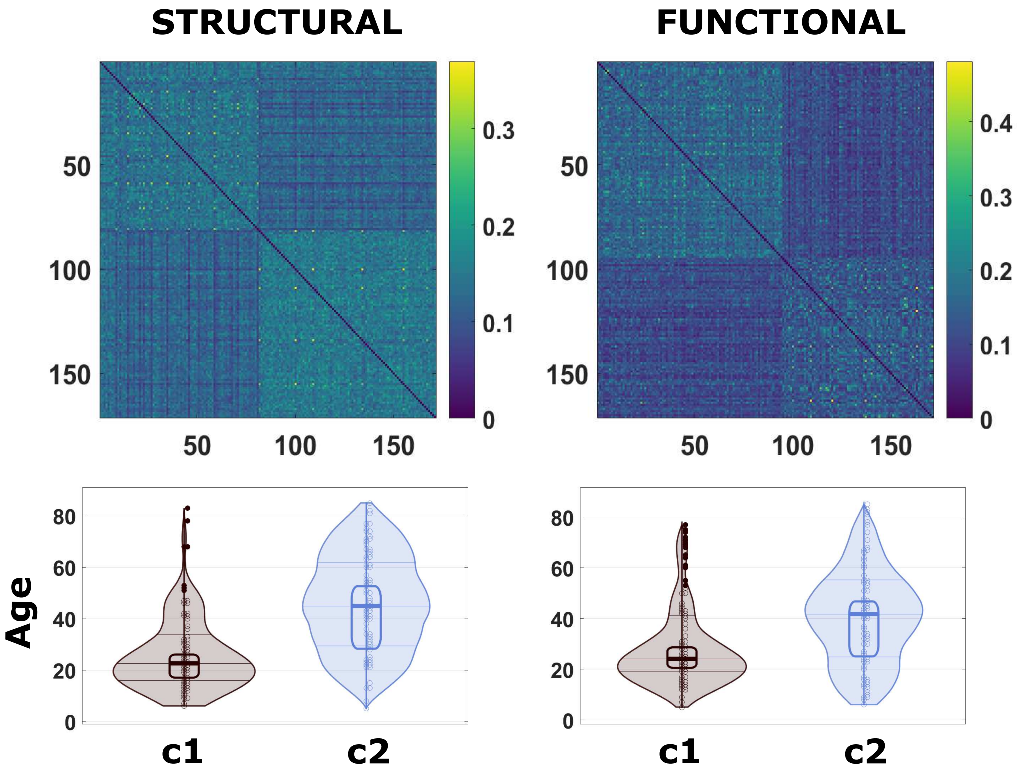

FIG. 6: (Up) Concerning the NKI data set, the consensus matrix found by the proposed approach is shown for structural (up-left) and functional (up-right) connectivity. (Down-left) Age differences in structural connectivity data (p-value $= 9 \times 10^{-11}$) and (down-right) functional connectivity data (p-value $= 7.00 \times 10^{-4}$) respectively. The rectangles indicate the estimator with 95 percent high density interval, calculated by Bayesian bootstrap. The shaded areas indicate random average shifted histograms, with a kernel density estimate. The code for these plots is available at `https://github.com/CPernet/Robust_Statistical_Toolbox/`, courtesy of Cyril Pernet

describes the geometry of the data-set providing at different resolutions groups of similar subjects. Obviously the choice of k-medoids as the clustering algorithm for the individual layers, is not mandatory, other algorithms can be used as well. Moreover, in the present work the features that we considered are the connectivity maps of all nodes, however other features can be considered as well and the same strategy can applied to fuse the different layers, each corresponding to a feature, and produce a consensus matrix.

Acknowledgements

JR acknowledges financial support from the Minister of Education, Language Policy and Culture (Basque Government) under Doctoral Research Staff Improvement Programme.

[1] O. Sporns, *Networks of the Brain.* 2011.

[2] R. C. Craddock, S. Jbabdi, C.-G. Yan, J. T. Vogelstein, F. X. Castellanos, A. D. Martino, C. Kelly, K. Heberlein, S. Colcombe, and M. P. Milham, "Imaging human connectomes at the macroscale," *Nature Methods*, vol. 10, pp. 524–539, may 2013.

[3] A. Fornito and E. T. Bullmore, "What can spontaneous fluctuations of the blood oxygenation-level-dependent signal tell us about psychiatric disorders?," *Current Opinion in Psychiatry*, vol. 23, pp. 239–249, may 2010.

[4] R. Filipovych, S. M. Resnick, and C. Davatzikos, "Semi-supervised cluster analysis of imaging data," *NeuroImage*, vol. 54, pp. 2185–2197, feb 2011.

[5] R. Filipovych, S. M. Resnick, and C. Davatzikos, "JointMMCC: Joint maximum-margin classification and clustering of imaging data," *IEEE Transactions on Medical Imaging*, vol. 31, pp. 1124–1140, may 2012.

[6] J. Steiner, P. Guest, H. Rahmoune, and D. Martins-de Souza, "The application of multiplex biomarker techniques for improved stratification and treatment of schizophrenia patients," in *Multiplex Biomarker Techniques* (P. C. Guest, ed.), no. 1546 in Methods in Molecular Biology, pp. 19–35, Springer New York. DOI: 10.1007/978-1-4939-6730-8_2.

[7] E. Varol, A. Sotiras, and C. Davatzikos, "HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework," *NeuroImage*, feb 2016.

[8] S. Takerkart, G. Auzias, B. Thirion, and L. Ralaivola, "Graph-based inter-subject pattern analysis of fMRI data," *PLoS ONE*, vol. 9, p. e104586, aug 2014.

[9] E. Amico, D. Marinazzo, C. DiPerri, L. Heine, J. Annen, C. Martial, M. Dzemidzic, S. Laureys, and J. Goñi, "Mapping the functional connectome traits of levels of consciousness," tech. rep., may 2016.

[10] A. Iraji, V. D. Calhoun, N. M. Wiseman, E. Davoodi-Bojd, M. R. Avanaki, E. M. Haacke, and

Z. Kou, "The connectivity domain: Analyzing resting state fMRI data using feature-based data-driven and model-based methods," *NeuroImage*, vol. 134, pp. 494–507, jul 2016.

[11] A. Lancichinetti and S. Fortunato, "Consensus clustering in complex networks," *Scientific Reports*, vol. 2:336, mar 2012.

[12] A.-L. Barabasi and J. Frangos, *Linked: The New Science of Networks*. Perseus Books Group, 1st ed.

[13] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, pp. 1059–1069, sep 2010.

[14] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, pp. 1–122, nov 2014.

[15] P. Brito, P. Bertrand, G. Cucumel, and F. D. Carvalho, *Clustering by means of Medoids. Selected Contributions in Data Analysis and Classification*. Springer Science & Business Media, 2007.

[16] Z. Shehzad, C. Kelly, P. T. Reiss, R. C. Craddock, J. W. Emerson, K. McMahon, D. A. Copland, F. X. Castellanos, and M. P. Milham, "A multivariate distance-based analytic framework for connectome-wide association studies," *NeuroImage*, vol. 93, pp. 74–94, jun 2014.

[17] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577–8582, may 2006.

[18] T. O. Laumann, E. M. Gordon, B. Adeyemo, A. Z. Snyder, S. J. Joo, M.-Y. Chen, A. W. Gilmore, K. B. McDermott, S. M. Nelson, N. U. Dosenbach, B. L. Schlaggar, J. A. Mumford, R. A. Poldrack, and S. E. Petersen, "Functional system and areal organization of a highly sampled individual human brain," *Neuron*, vol. 87, pp. 657–670, aug 2015.

[19] R. A. Poldrack, T. O. Laumann, O. Koyejo, B. Gregory, A. Hover, M.-Y. Chen, K. J. Gorgolewski, J. Luci, S. J. Joo, R. L. Boyd, S. Hunicke-Smith, Z. B. Simpson, T. Caven, V. Sochat, J. M. Shine, E. Gordon, A. Z. Snyder, B. Adeyemo, S. E. Petersen, D. C. Glahn, D. Reese Mckay, J. E. Curran, H. H. H. Göring, M. A. Carless, J. Blangero, R. Dougherty, A. Leemans, D. A. Handwerker, L. Frick, E. M. Marcotte, and J. A. Mumford, "Long-term neural and physiological phenotyping of a single human.," *Nature communications*, vol. 6, p. 8885, 2015.

[20] J. M. Shine, O. Koyejo, and R. A. Poldrack, "Temporal metastates are associated with differential patterns of time-resolved connectivity, network topology, and attention," *Proceedings*

of the National Academy of Sciences, vol. 113, pp. 9888–9891, aug 2016.

[21] X. Shen, F. Tokoglu, X. Papademetris, and R. T. Constable, "Groupwise whole-brain parcellation from resting-state fMRI data for network node identification.," *NeuroImage*, vol. 82, pp. 403–15, nov 2013.

[22] `http://fcon_1000.projects.nitrc.org/indi/pro/nki.html`.

[23] J. A. Brown, J. D. Rudie, A. Bandrowski, J. D. V. Horn, and S. Y. Bookheimer, "The UCLA multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis," *Frontiers in Neuroinformatics*, vol. 6, 2012.