Geodesic Clustering of Positive Definite Matrices For Classification of Mental Disorder Using Brain Functional Connectivity

Muhammad Abubakar Yamin
Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy.
Department of Electrical, Electronics and Telecommunication Engineering, Università degli Studi di Genova.
Genova, Italy.
muhammad.yamin@iit.it

Jacopo Tessadori
Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia,
Genova, Italy

Muhammad Usman Akbar
Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia,
Genova, Italy.
Department of Electrical, Electronics, and Telecommunication Engineering Università degli Studi di Genova.
Genova, Italy.

Michael Dayan
Human Neuroscience Platform, Fondation Campus Biotech Geneva,
Geneva, Switzerland.

Vittorio Murino
Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy
Department of Computer Science, Università di Verona,
Verona, Italy

Diego Sona
Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia,
Genova, Italy.
Neuroinformatics Laboratory, Fondazione Bruno Kessler, Trento, Italy

***Abstract*—Functional Magnetic Resonance Imaging (fMRI) is a commonly used technique to evaluate brain activity, and can be used to distinguish patients from healthy controls in a variety of diseases. In this work, we present a two-step approach to discriminate healthy subjects against those affected by either Autism Spectrum Disorder or Schizophrenia on the basis of their connectivity patterns. We exploited the property that connectivity patterns described by positive definite matrices define a Riemannian manifold. In this framework, to generate a vector representation used in the classification task, we performed a geodesic clustering of the connectivity matrices. Cluster centroids were then used as a dictionary allowing to encode all subjects graphs as vectors of geodesic distances. A linear Support Vector Machine was then used to classify subjects. To show the advantage of using geodesic distances for this problem, the same analysis was conducted using a Euclidean metric. Experiments show that employing Euclidean distances leads to a lower classification performance and possibly to the definition of the wrong number of clusters, whereas geodesic clustering results in a significantly improved accuracy.**

***Index Terms*—static functional connectivity, geodesic clustering, k-means clustering, connectomes, SVM, SPD matrices, Riemannian manifold**

I. Introduction

The study of connectivity between different regions of the brain is known as "connectomics", which is a relatively recent research field that allows neuroscientists to investigate the interplay between different regions of the brain by modeling it as a network or "connectome". At the whole-brain scale, functional processes are commonly investigated through resting-state functional Magnetic Resonance Imaging (rs-fMRI). In particular, a functional connectome is usually constructed by computing Pearson correlation between averaged time-series of brain regions defined through an appropriate atlas. A network can be obtained by defining nodes as atlas regions and links as functional connectivity (FC) between pairs of these regions during rest. Nowadays FC plays a vital role to characterize brain connectivity in many psychiatric and neurodegenerative disorders. In turn, this representation can be used to discriminate healthy controls (HC) from those affected by such disorders.

To this aim, recently, different methods have been proposed to classify groups of subjects using the geometrical properties of symmetric positive definite (SPD) matrices. The set of all SPD matrices of the same size forms a Riemannian manifold, so several approaches have been developed to leverage this manifold structure during the analysis. In [1], a probabilistic model for covariance matrices was used to distinguish post-stroke patients from HC. In [2] manifold transportation of covariance matrices was applied in longitudinal studies to determine changes in FC after a task. In [3] a kernel based classification approach has been deployed which analyzed the FC matrices using Log-Euclidean Gaussian kernel and Stein Gaussian kernels. In [4], an approach based on Grassmanian geometry and low-rank graph Laplacian has been used for a classification task exploiting a set of sub-networks that was then used to identify connectivity biomarkers.

Support Vector Machines (SVM) can be used to directly perform classification of vectorized brain graphs, selecting time-series correlation as features [5], [6]. However, this ap-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

proach partly misrepresents the real geometry of the problem, as it attempts to adopt Euclidean metrics to describe data which, in fact, lie in a Riemannian manifold. Correctly taking into account the properties of positive semi-definite matrices allowed to classify sub-connectivity patterns [7], functional states generated from auditory stimuli [8] or mild cognitive impairment [9].

In this work, we employ a geodesic clustering algorithm which uses geodesic metrics on a Riemannian manifold to cluster FC matrices. The computed centroids are then used to generate a representation allowing to discriminate between classes. More specifically, using a two-fold approach, functional connectivity matrices of brain activity during rest are clustered and, in a second step, the geodesic distances of the connectivity matrices from the cluster centroids are used as features to train a linear-SVM.

The proposed method has been tested on two different problems: HC vs. subjects affected by Autism Spectrum Disorder (ASD), and HC vs. subjects affected by Schizophrenia (SCHZ). To show the benefit of using the Riemannian properties, the same experiments have been done using the Euclidean metrics, comparing the results in terms of both clustering and classification performance. Moreover, we have also compared the results from our approach versus state-of-the-art methods [3], [4]. Results showed that analysis of FC matrices by using our approach gives better results in term of classification accuracy.

Organization: Following the brief introduction and background information given above, Section II describes the public dataset used for this experiment and the proposed method for the classification. Section III reports the results of experiments done, while Section IV concludes with a discussion on the proposed method and the results obtained.

II. Material and Method

A. Data Acquisition and Pre-Processing

In this work, to test our method, we have used two publicly available functional connectivity datasets. The first dataset is from the ASD connectome database released by UCLA [10]. This dataset is composed of the rs-fMRI of 37 HC and 42 ASD patients. Further details of acquisition and pre-processing are described in [11]. FC matrices were obtained from the Power atlas, which defines 264 regions of interest (ROIs) in the brain [12]. These 264x264 FC matrices are estimated for each subject by computing the pairwise Pearson correlation between average time-series of brain ROIs. Furthermore, we analyzed the FC dataset released by the Network Based Statistic (NBS) toolbox. It is composed of 15 HC and 12 SCHZ subjects [13]. In this dataset, FC matrices were built using a subset of regions from the AAL atlas (90 ROIs without cerebellum) using the same pairwise Pearson's correlation approach. In our method, we are considering the whole connectivity matrix including negative values.

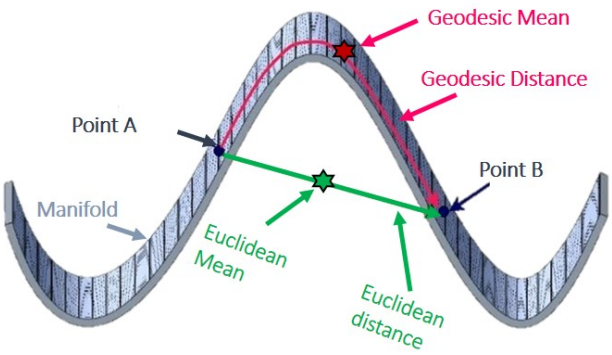

Fig. 1. The difference between Euclidean distance and Euclidean mean of two points (green straight line and star) and the corresponding geodesic distance and geodesic mean (red curve and star along the manifold).

B. Manifold Representation of SPD Matrices

Let $\mathcal{X}_\rho = \{\rho_1, \ldots, \rho_n\}$ be the set of correlation matrices describing the brain functional connectivity of all N subjects. The correlation matrices are symmetric and positive semi-definite in nature and can be easily regularized into SPD matrices by adding a small constant to the main diagonal ($\rho_i = \rho_i + \lambda I$, with λ very small, e.g., $\lambda = 10^{-5}$). The set of all SPD matrices of the same size form a Riemannian manifold, which allows the analysis of such matrices on a manifold space. To take the full advantage it is recommended to use the notion of geodesic distances which allows a description of this data better than using Euclidean metrics [3], [14].

Intuitively, a geodesic distance computes the shortest path between two points over a smooth and curved manifold [15]. There are several possible alternative geodesic distances on the Riemannian manifold of SPD matrices [14], [16]; we decided to adopt the Log-Euclidean (Log-E) distance, which is simple and fast to compute. Equations 1 and 2 describe, respectively, the log-E distance formula between two SPD matrices ρ_i and ρ_j and the closed form formula to compute the mean [17] of two or more SPD matrices with this metric. The conceptual difference between geodesic distance, Euclidean distance and corresponding means is illustrated in Figure 1.

$$d_L(\rho_i, \rho_j) = \| \log(\rho_i) - \log(\rho_j) \|_F. \tag{1}$$

$$\begin{aligned} \rho_L &= \exp\left\{ \arg\inf_{\rho} \sum_{i=1}^{n} \| \log(\rho_i) - \log(\rho) \|^2 \right\} \\ &= \exp\left\{ \frac{1}{n} \sum_{i=1}^{n} \log(\rho_i) \right\}, \end{aligned} \tag{2}$$

C. Geodesic Clustering Analysis

In the proposed method, we have used geodesic k-means clustering algorithm [16] to cluster the FC matrices into different groups. The aim is to divide FC matrices into homogeneous groups of subjects presenting similarities in their connectivity.

The underlying assumption is that there are some alterations in brain connections of the patients [18] that can be grasped by the clusters. K-means was implemented using the Log-E distance [14] as defined in eq.(1), with the centroids computed as the geodesic mean, which can be computed in the closed form by eq.(2).

D. The DB Index

In order to choose the optimal number of clusters (K) we used the Davies-Bouldin (DB) index as criterion [19]. This index evaluates the consistency using the distance of all points within a cluster to the corresponding centroids and the separation between clusters using the distance between centroids. The lower is the index, the better are the clustering results. In this work, the DB index is computed for every considered number of clusters (i. e. K=[2,3,4,5,6]) and the minimum value suggests the natural partition of data.

Consider a set of correlation matrices $\mathcal{X}_\rho = \{\rho_1, \ldots, \rho_n\}$ and a set of clusters $\mathcal{C} = \{c_1, \ldots, c_k\}$ partitioning $\mathcal{X}_\rho$ in K groups. Cluster representatives are defined as

$$\overline{c_k} = \frac{1}{|c_k|} \sum_{\rho_i \in c_k} \rho_i \tag{3}$$

and the distance between matrices $d(\rho_i, \rho_j)$ used in our analysis between items is the Log-E distance. The equation for the DB index is given as follow

$$S(c_k) = \frac{1}{|c_k|} \sum_{\rho_i \in c_k} d(\rho_i, \overline{c_k}) \tag{4}$$

and

$$DB(\mathcal{C}) = \frac{1}{K} \sum_{c_i \in \mathcal{C}} \max_{c_j \in \mathcal{C} \setminus c_i} \left\{ \frac{S(c_i) + S(c_j)}{d(\overline{c_i}, \overline{c_j})} \right\} \tag{5}$$

E. Feature Extraction and Classification

The working hypothesis is that we can cluster the FC matrices preserving the alteration of brain connectivity characterizing the groups. This would allow therefore a compression of graphs into a smaller vectorized representation retaining the group differences while filtering the intrinsic variability of subjects in the same group. Indeed, using the cluster representatives as a dictionary, we built a vector representation for each subject, computing the features as the distances of the subject FC matrix from all cluster centroids.

In our experiments, we performed geodesic clustering multiple times with a variable number of clusters ranging from $K = 2$ to $K = 6$ in order to find the best K. Once convergence was achieved we computed the Log-E distance between the samples in the training set and all K centroids (e.g. for $K = 2$ each sample was described by 2 distance values and for $K = 4$ each sample was represented by 4 features). These distance values were used as feature vectors to train a linear-SVM. In the test phase, each sample in the test set was described by the distances of the corresponding FC matrix from all cluster centroids computed during training.

To avoid double dipping we made all the experiments using 5-fold cross-validation, randomly selecting the samples and preserving the proportion between the classes in each fold. For statistical reasons we repeated this cross-validation process 100 times with randomized selection of folds.

In the end, we evaluated the results in terms of average accuracy and confusion matrix averaging over all 100 iterations. In our experiments, all distances were computed using the Log-E distance (eq. 1) and the corresponding geodesic mean (eq. 2). In addition, to show the advantage of using the geodesic distance on the manifold containing the data, we performed identical computations using Euclidean metrics, allowing to evaluate the differences in performance. In order to check the significance level of the performance of our classifier we performed a permutation test on labels. For this purpose, we generated a null distribution by randomly changing the labels 1000 times and in each iteration we performed L-SVM classification using 5-fold cross validation and computed the mean cross fold accuracy.

III. Results

Figure 2 depicts the box plots showing the classification accuracy over 100 iterations with the proposed method on the ASD dataset (Fig. 2A) and on the SCHZ dataset (Fig. 2B). Blue and orange bars represent results obtained with geodesic and Euclidean metrics respectively. The gray line shows the average over 100 iterations of the DB index for the geodesic clustering. For the ASD dataset (Fig. 2A) it can be seen that the highest mean accuracy (67.12%) is achieved with Log-E distance for $K = 4$ clusters, whereas with Euclidean distance the maximum obtained mean accuracy is 61.59% with $K = 6$ clusters. Figure 2A also shows that the DB index (line plot) has a minimum value in $K = 4$, suggesting that this is the optimal number of clusters. This is reinforced by the fact that this is the same number of clusters with peak accuracy for geodesic clustering.

For the Schizophrenia dataset, Figure 2B shows that, with the geodesic metric, maximum mean accuracy (75.33%) is achieved with K=2. Similarly, in Figure 2B, the DB index (silver line) for this dataset also has the minimum value for K=2. On the other hand, for Euclidean metrics, the maximum accuracy is achieved (70.03%), on this dataset, for K=4. The embedded tables in both figures summarize these results. Table 1 shows the average confusion matrix for geodesic clustering results for both datasets between HC and pathological subjects.

In order to assess the statistical significance of our obtained results we implemented a permutation test. The results of permutation test are represented in form of pvalues in Figure 2. The p-value is computed as the ratio between the number of accuracy values greater than the tested accuracy and the total number of permutations (1000 in our case). These results strongly support the principle according to which the use of geodesic metric on SPD matrices, which form a Riemannian manifold, gives better results in term of accuracy, whereas the use of Euclidean metric on SPD matrices is suboptimal.

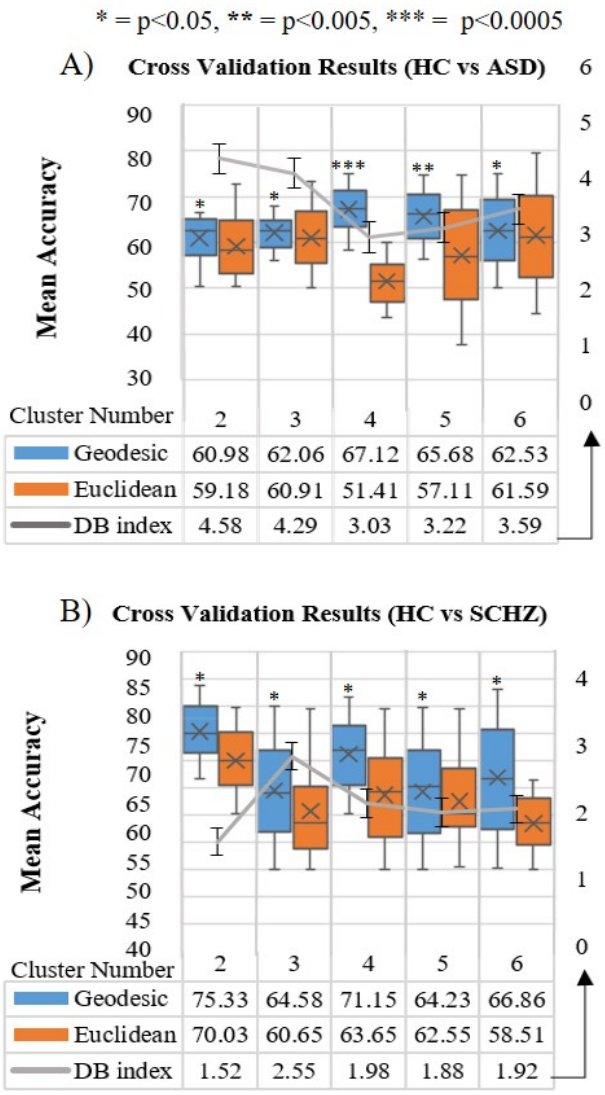

Fig. 2. Boxplot representing the mean classification accuracy for A) HC vs ASD and B) HC vs SZ dataset with geodesic (blue box) and Euclidean (orange box) metrics based k-means clustering. Line plot shows the mean DB index value for each cluster of geodesic k-means clustering. selection. Stars on the bar shows the significance level obtained through permutation test.

Results from the proposed methodology also outperform the results presented in [3,4] using the same dataset.

IV. Discussion and Conclusion

In this work, we have presented a novel computational framework, which allows the classification of HC and patients using static FC matrices obtained from rs-fMRI. To achieve this goal, we performed k-means clustering by taking advantage of the properties of SPD matrices: in this context, using geodesic metrics proved to be superior to the Euclidean approach.

In particular, classification features have been constructed with a subject-wise graph similarity representation by using a geodesic metric based on k-means clustering. This algorithm adopted log-Euclidean distance on the Riemannian manifold

TABLE I
Confusion matrix of average classification results for the proposed approach based on geodesic clustering for HC vs. ASD and HC vs. SCHZ datasets

Mean Confusion Matrix of HC vs ASD (for K=4)			
		Predicted Class	
		HC	Pathological Subjects
Actual Class	HC	22	15
	Pathological Subjects	12	30

Mean Confusion Matrix of HC vs SCHZ (for K=2)			
		Predicted Class	
		HC	Pathological Subjects
Actual Class	HC	10	5
	Pathological Subjects	2	10

space of SPD matrices. To avoid double dipping, both clustering and classifier tuning occurred on training folds: we applied the geodesic k-mean clustering algorithm to compute the centroids on data independent from the testing folds. Furthermore, we computed the distance of each training sample from each centroid and used this distance vector as feature set to train the L-SVM classifier. In testing, we computed the distance of each test sample from the centroids defined on the training set and then used this distance value to test the performance of the trained classifier. To reduce the impact of fold selection, this process was repeated for 100 iterations and results were summarized as the average of all these iterations.

In order to evaluate our proposed algorithm, we made a similar experiment but using the Euclidean metric base k-means clustering distance instead of the geodesic metric. The results of this study noticeably reveal that the use of Euclidean metrics on the manifold of SPD matrices is suboptimal, as it is causing a data representation leading to decreased accuracy. Indeed, the classification performance improved when using the geodesic metric which computes the shortest possible distance along the curvature of the manifold, thus offering an optimal data representation. Hence to compare and analyze FC SPD matrices it is suggested to consider a geodesic metric exploiting the properties of the Riemannian space on which these matrices lie. This study also reveals that a specific encoding of the FC matrices, describing them according to their distances from cluster centroids, allows good performance in distinguishing between HC and patients.

References

[1] G. Varoquaux, F. Baronnet, A. Kleinschmidt, P. Fillard, and B. Thirion, "Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling," in MICCAI 2010, pp. 200–208. Springer, 2010.

[2] B. Ng, M. Dressler, G. Varoquaux, J. B. Poline, M. Greicius, and B. Thirion, "Transport on Riemannian manifold for functional connectivity-based classification," in MICCAI 2014, pp. 405–412. Springer, 2014.

[3] L. Dodero, H. Q. Minh, M. S. Biagio, V. Murino and D. Sona. "Kernel-based classification for brain connectivity graphs on the Riemannian manifold of positive definite matrices". ISBI 2015. 16-19 April

[4] L. Dodero., F. Sambataro, V. Murino, D. Sona. "Kernel-Based Analysis of Functional Brain Connectivity on Grassmann Manifold". In MICCAI 2015. Lecture Notes in CS, vol. 9351.

[5] Satterthwaite, T.D., Wolf, D.H., Ruparel, K., Erus, G., Elliott, M.A., Eickho, S.B., et al.: Heterogeneous impact of motion on fundamental patterns of developmental changes in functional connectivity during youth. Neuroimage 83 (2013) 45-57
[6] Craddock, R.C., Holtzheimer, P.E., Hu, X.P., Mayberg, H.S.: Disease state prediction from resting state functional connectivity. Magnetic resonance in Medicine 62(6) (2009) 1619-1628
[7] Eavani, H., Satterthwaite, T.D., Filipovych, R., Gur, R.E., Gur, R.C., Davatzikos, C.: Identifying sparse connectivity patterns in the brain using resting-state fmri. NeuroImage 105 (2015) 286-299
[8] S. Vega-Pons, P. Avesani, M. Andric, and U. Hasson, "Classification of inter-subject fMRI data based on graph kernels," in Pattern Recognition in Neuroimaging, 2014 International Workshop on, June 2014, pp. 1–4.
[9] B. Jie, D. Zhang, Chong-Yaw Wee, and D. Shen, "Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification," Human brain mapping, 2014.
[10] J A Brown, J D Rudie, A Bandrowski, John D Van 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.
[11] JD Rudie, JA Brown, D Beck-Pancer, LM Hernandez, EL Dennis, PM Thompson, SY Bookheimer, and M Dapretto, "Altered functional and structural brain network organization in autism," NeuroImage: clinical, vol. 2, pp. 79–94, 2013.
[12] J. D Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A.C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al., "Functional network organization of the human brain," Neuron, vol. 72, no. 4, pp. 665–678, 2011.
[13] A. Zalesky, A. Fornito, and E.T. Bullmore, "Network-based statistic: identifying differences in brain networks," Neuroimage, vol. 53, no. 4, pp. 1197–1207, 2010.
[14] Yamin A., Dayan M., Squarcina L., Brambilla P., Murino V., Diwadkar V., and Sona D. "Comparison of brain connectomes using geodesic distance on manifold: A twin's study" International Symposium on Biomedical Imaging 2019 Venice (April 8-11 2019).
[15] O. Tuzel, F. Porikli, P. Meer, "Pedestrian detection via classification on Riemannian manifolds", PAMI, vol. 30, no. 10, pp. 1713-1727, 2008.
[16] Yamin A., Dayan M., Squarcina L., Brambilla P., Murino V., Diwadkar V., and Sona D. (2019) Analysis of Dynamic Brain Connectivity Through Geodesic Clustering. Image Analysis and Processing – ICIAP 2019. Lecture Notes in Computer Science, vol 11752. Springer, Cham
[17] Dryden, i.l., Koloydenko, a. and Zhou, d., 2009. Non-euclidean statistics for covariance matrices with applications to diffusion tensor imaging. Annals of applied statistics, 3 (3), pp. 1102 - 1123.
[18] Zhou Y, Zeidman P, Wu S, Razi A, Chen C, Yang L, Zou J, Wang G, Wang H, Friston KJ. Altered intrinsic and extrinsic connectivity in schizophrenia. Neuroimage Clin. 2017 Dec 5;17:704-716. doi: 10.1016/j.nicl.2017.12.006. PMID: 29264112; PMCID: PMC5726753.
[19] V. M. Vergara, A. Abrol, F. A. Espinoza and V. D. Calhoun, "Selection of Efficient Clustering Index to Estimate the Number of Dynamic Brain States from Functional Network Connectivity*," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 632-635.