Exploratory framework on EEG signals for the development of BCIs

Martin Perez-Guevara
Supervisor: Christopher James

ABSTRACT: Finding appropiate spatio-temporal features of electroencephalography (EEG) signals to build a brain computer interface (BCI) is an extremely complex and challenging problem. In this study, motivated by new perspectives on the brain computer interface's research community and new methods developed in topological data analysis, we propose a framework to explore features in the EEG signal domain in an unsupervised way, such that a subject aided by machine learning and self labeling could establish the basis of a completely personalized and accurate BCI. In this study, we show the viability of an exploratory framework methodology based mainly on the detection instead of classification perspective and the Mapper method that belongs to the domain of persistent homology, and discuss further research and development necessary to implement the framework.

1. INTRODUCTION

Electroencephalography (EEG) is a recording of the electrical activity around the scalp. Specifically, as explained in Piotr Olejniczak review [1], it is a graphic time series representation of the difference in voltage between two different cerebral locations. The obtained signal is influenced by diverse factors like the electrical conductive properties of the tissues that lie in the middle of the electrical source and the electrode employed to measure the potentials, the conductive properties of the electrode itself and the orientation of the potential source in the cortex.

The EEG is made possible thanks to the current flow that passes through the tissues between the source of electrical activity inside the brain and the recording electrode. But then EEG only provides a two-dimensional picture of a three dimensional process, which degenerates into the inverse problem, since the electrical source can not be uniquely determined from EEG and so different tasks extracted from brain activity might not be easy to differentiate. Also the bone tissue and skin tissue with the influence of the environment add additional noise to the signal that further complicates analyzing EEG recordings.

Even though EEG presents the inverse problem and a low signal to noise ratio, it is still employed for its extremely high time resolution, its non invasive nature and recent developments in commercial products, like the Emotiv headset , that lower the costs of the hardware. Turning EEG into an accessible solution for diverse applications like the NeuroPhone system proposed by researchers in Darthmouth College [2].

In this study, EEG will be employed as the main tool to explore the construction of a Brain-computer Interface (BCI), which consists on interpreting the signals obtained from the brain to control commands in a computer program, for example selecting characters to write a word.

Most BCI applications developed so far are aimed towards aiding disabled people. In a similar fasion to eye glaze devices, BCI applications allow writing, selecting items on a screen or controlling the direction of a wheelchair. Moreover the most successful algorithms relay on motor imagery and event based potentials, since particularly paraplegic people can use these intentions without the interference of muscle activity on the EEG signal.

The P300 Wave, as presented by Picton [3], is a great example of a context specific feature that can be extracted from the waveforms of EEG activity signals to detect the conscious intention of a subject to select an improbable target. Donchin et all [4], show how the P300 wave effect can be employed to implement a BCI for spelling with high accuracies that only depends on visual stimuli and the intention of a subject. This particular implementation commonly known as the the P300 speller in the BCI community is one of the promising illustrations of the potential and possibilities of BCI employing only EEG to record brain activity.

However, Schalk et all [6], identify in their study two important issues regarding the construction of a BCI, the signal identification problem and the signal identification paradox.

The signal identification problem explains that the selection of EEG signals and their location, originated by specific brain activity, is not completely understood yet (there is essentially no theoretical basis), since even the fundamental processes behind brain activity are not well comprehended.

The BCI signal identification problem is fundamentally different to a normal classification problem in the sense that data classes are not easily defined a priori to consequently select the features. It has been only empirically shown sometimes that particular mental tasks have particular effects on specific brain signals, and still the definition of the tasks and signal features to implement some kind of classification is difficult, suboptimal and ill defined. Moreover the identified signal features are normally subject-dependent and non-stationary.

So it is expected that multiple alternative features like the P300 wave could still be identified and employed for BCI applications. Furthermore, motor imagery and tasks normally defined to implement a BCI for disabled people could not be that useful when considering completely healthy individuals in diverse contexts aimed at machine control, critical applications and augmented reality.

This unexplored domain of signal features and defined tasks pose an interesting challenge that might be addressed by personalized unsupervised learning. In this study, topological data analysis will be presented as an approach to aid the exploration of the features of EEG signals under different brain activity states. Particularly the Mapper method [5] will be implemented as a tool to explore and visualize the signals in conjunction with hard and soft clustering algorithms.

On the other hand, the signal identification paradox is due to the fact that there is no a priori basis for selecting mental tasks and signal features, so the possible choices increase with increasing signal fidelity and the latter improves by defining and discriminating more classes of brain activity (signal specificity), which means that also the identification procedure and algorithmic training increase in complexity.

Moreover signals might change in time and under learning and interaction conditions in ways that are difficult to identify to retrain the algorithmic classifiers. Under this scenario of increasing complexity in the dynamics of the feature/task space, it is possible that the BCI performance may degrade even with better signal recording.

As an anwer to the mentioned paradox, Schalk et all [6] propose the SIGFRIED (SIGnal modeling For Real-time Identification and Event Detection) methodology. Which consist on the perspective of detecting events that are unlikely to belong to a general class which is uninteresting for control, like a resting state, to then use the unlikelihood of events as a control parameter in the construction of a BCI. This approach greatly simplifies the collection of labeled information on an exploratory framework and will be employed in this study in conjunction with the Mapper method to try to reveal interesting areas on the feature space that might be discriminated to build control mechanisms on a BCI.

Furthermore there is an interesting phenomena appreciated during the development of BCI with different subjects under similar experimental conditions called BCI illiteracy. This consist on the inability of the algorithms that were able to build accurate classifiers based on the identified relation between signal features and tasks to work on a non-negligible portion of the subjects population (between 15% and 30%), as explained by Vidaurre et all. [7]. It is possible that employing personalized unsupervised learning to detect relevant signal features for each subject will address this problem, as there seems to be in many cases no completely universal solution to relate specific brain states and feature signals.

Considering the potential of EEG and all the mentioned challenges to build a BCI, the main proposal of this study will consist on developing an exploratory framework based on personalized unsupervised learning, topological data analysis and the detection instead of classification perspective to tackle the emphasized problems while at the same time looking for new insights in the identification of useful EEG signal features.

2. TOPOLOGICAL DATA ANALYSIS

Nowadays data is being produced at increasing rates thanks to new experimental methods and developments in high power computing. Furthermore the nature of data is changing, now it is more high-dimensional and noisier with more missing parts then ever. As explained by Gunnar Carlsson [8], developments in geometry and topology might be employed to bring to light many informative features of this kind of datasets for which conventional methods that depend on specific metrics or low dimensional spaces might fail.

According to Gunnar [8] there are several key points regarding data analysis that justify the use of geometric and topological methods. For instance: That to obtain knowledge about how data is organized in a large scale is desirable, like finding out patterns or clusters that show something about the dataset; That metrics are not theoretically justified in many domains of problems like in biology; That spaces and their coordinates are not natural in any sense and are also commonly unjustified; And that summaries over the whole range of parameters when analyzing data can be more valuable than individual choices, like keeping the whole dendogram when realizing hierarchical clustering procedures to analyze data.

Then topology turns out to be a good candidate to address the mentioned issues, since, as stated by Gunnar [8], it is the branch of mathematics which deals with qualitative geometric information. In general it is the study of connectivity information, so topological methodologies like homology can help study the datasets. Also the geometric properties studied by topology are less sensisitive to the choice of metrics and do not depend on chosen coordinates on specific spaces.

Furthermore, the idea of building summaries over complete domains of parameters, when analyzing datasets, involves the notion of functoriality that is at the heart of algebraic topology and allows the computation of homological invariants from local information. Also in general it is known that information about topological spaces can be learned by simplicial approximation.

A great example of how topology can be applied to data analysis can be appreciated in the research done by Gunnar et all [9], where they found a new type of breast cancer from microarray data, with a significant biological signature, that was completely ignored by previously employed clustering algorithms. This discovery was done in a completely unsupervised way by looking at the shape of the data obtained with the application of the Mapper method [5] and the selection of functions representative of the problem at hand, like the abnormality of the cancer tissue. Then the discovery was confirmed by going back to the specific clustered points of the dataset and studying their relation.

Moreover the mentioned research was one of the main motivations to apply topological data analysis and particularly the

Mapper method in this study. Since EEG signal features can constitute a very high dimensional and still quite unexplored dataset, like the microarray dataset employed to understand cancer tissue.

2.1 Details on the application of topological data analysis for this study

Point clouds, understood as a finite set of points for which a distance function applies, are the main objects to which the geometric and topological techniques are applied. As explained by Gunnar [8], one can think of point clouds as finite samples taken from a geometric object, perhaps with noise. The notion of point clouds is quite abstract and therefore any set of points defined in an n-dimensional space for which a distance function can be defined is a candidate for topological data analysis.

Then the features of a finite segment of an EEG signal, which is a discrete multidimensional time series, can be represented as a point of data. This implies that by considering consecutive segments of the EEG signal, one can obtain point clouds representative of brain activity and so of sets of tasks behind that brain activity. In this study the features of overlapped segments of the same length of an EEG signal will be considered to form the point clouds.

Once a point cloud is defined, one needs to represent somehow its topology to be able to apply any of the tools of homology to find homological invariants that will give us an insight on the properties of the underlying object analyzed. As explained by Gunnar [8], intuitively, a simplicial complex structure on a space is a representation of the space as a union of points, intervals, triangles and higher dimensional analogues. And it turns out that a simplicial complex provides a particularly simple combinatorial way to represent topological spaces.

There are several methods to build simplicial complexes to approximate the topology of the space represented by the point cloud, like the Cech, Vietoris-Rips and Witness complexes methods. But in this study clustering algorithms will be employed as explained in the Mapper method [5].

Clustering algorithms take a finite metric space as an input and produce as output a partition of the underlying space, where the subspaces defined by the partition are considered as clusters. In the context of a metric space, this means that points inside a cluster are nearer to each other than to points in different clusters.

As stated by Gunnar [8], clustering should be thought of as the statistical counterpart to the geometric construction of the path-connected components of a space, which is the fundamental block upon which algebraic topology is based. This justifies the use of clustering as an alternative tool for the construction of simplicial complexes as outlined in the Mapper method [5].

Nonetheless an important problem in the construction of the simplicial complexes under any method is the selection of the values of the parameters that in each case will result on the simplicial complex that best approximate the true topology of the object underlying the point cloud. This problem can be addressed by the ideas behind persistent homology. Specifically by looking at all the simplicial complexes defined by the whole range of values of the parameters considered and analyzing how the topological properties vary as the parameters' values change.

Robert Ghrist [10] surveys throughly how persistent homology can be employed on diverse point cloud datasets, by employing a representation of the induced algebraic characterizations called barcodes. Moreover Balakrishnan et all [11] introduce some statistical ideas to persistent homology to separate short lived topological properties considered as "topological noise" from the real approximated "topological signal" with measures of statistical confidence.

This opens up the possibility of exploring the brain processes behind the EEG signals in this study in an unsupervised way under general considerations of time (by progressively considering new datapoints in time) and space (by considering features of multiple combinations of electrode locations on the scalp at a given time).

2.2 Details on the Mapper method

The Mapper method, presented by Singh et all [5], allows the representation of a point cloud dataset as a simplicial complex. It is based on the idea of partial clustering of the data guided by a set of meaningful functions defined on the data.

After applying the method to the dataset, the obtained simplicial complexes can be analyzed with the techniques of persistent homology to reveal qualitative information about the underlying object represented by the point cloud dataset. Moreover the simplicial complexes generated by this method can be used to visualize and interpret the dataset directly thanks to the meaning assigned to the functions that guide the partial clustering.

The Mapper method steps, represented in Figure 1, can be summarized as:

1. Determining the point cloud dataset.
2. Selecting a small set of meaningful functions to map the points to a low dimensional space.
3. Segmenting the low dimensional space into intervals of length "l" overlapped with percentage "o".
4. Generating the subdatasets corresponding to the defined intervals in the low dimensional space.
5. Applying a clustering algorithm with automatic detection of the number of clusters in each subdataset to obtain the nodes of the simplicial complex.
6. Evaluating the intersections of clusters belonging to consecutive overlapped intervals to obtain the connections of the simplicial complex.
7. Building the simplicial complex for further analysis
8. In the case of further visualization of the simplicial

complex, define additional visual properties to analyze the dataset, like color or size of the nodes.

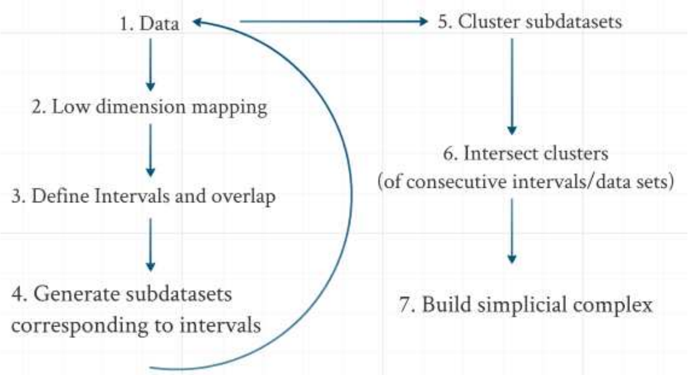

1. Data
2. Low dimension mapping
3. Define Intervals and overlap
4. Generate subdatasets corresponding to intervals
5. Cluster subdatasets
6. Intersect clusters (of consecutive intervals/data sets)
7. Build simplicial complex

Fig. 1. Diagram of the Mapper method

The visual properties that can be stablished to have a more informative representation of the dataset when visualizing the resulting simplicial complex are numerous. For example one can define the color to the nodes to represent the average value of the meaningful functions used for the low dimensional mapping or define the size of the nodes to represent the proportion of points belonging to each cluster with respect to the whole dataset.

In the case of this study, coloring the nodes, using pie charts to replace the nodes and maintaining spatial configurations of the nodes, when drawing the simplicial complex, are the visual properties that will be defined to get an informative view on the structure of the brain processes and tasks underlying the EEG signals.

Furthermore, a more comprehensive explanation of the mapper method along with its application on multiple examples of trivial and non trivial point cloud datasets belonging to diverse problems on shape and object recognition can be found in the paper published by Singh et all [5].

3. SIGFRIED (SIGnal modeling For Real-time Identification and Event Detection)

As was pointed out in the introduction, the BCI classification problem is peculiar in the sense that not only the features that represent classes but also the classes themselves have to be found and defined. This is an important problem, since there is no fundamental theoretical basis to define the tasks that are expected to generate changes on the signals recorded from brain activity and as greater signal fidelity is desired then more complex task definitions are also required.

Schalk et all [6] explains the signal identification problem and paradox and how they greatly increase the time cost of developing and implementing brain computer interfaces.This delay the adoption of BCI technology even though there has been important advances in the production of mass consumption inexpensive devices for EEG.

The time and effort that a subject requires to train the algorithmic classifiers grows fast with the complexity of the task definitions even when details about the relationship between the signal and the task are known. Moreover the procedures based on specific features of EEG signals might not work for all subjects. Furthermore, to complicate even more the problem, signal features have shown to be non stationary and highly sensitive to feedback conditions imposed by interacting with the digital interfaces.

SIGFRIED constitutes an anwer to the mentioned difficulties because it only needs a small sample of only one reference class to be able to discriminate other classes. It might also use more than one class as reference and the resulting output in any case is a continuous feature that can be employed as an input for computer commands. In addition the methods behind it are not expensive computationally and easy to implement.

Schalk et all [6] propose to define a rest category that could be used as a main reference to detect and analyze non rest activity that might be used as input for a BCI. In this study this suggestion will be taken into account to create a meaningful function to map the point cloud of the EEG signal to the low dimensional space in the Mapper method. This will allow to interpret the structure of the simplicial complex in terms of non rest or extreme events and to see if some distinction can be made between labeled tasks with respect to specific EEG signal features.

3.1 Details on SIGFRIED

SIGFRIED can be summarized in the following steps:

1. Specify signal features that will constitute the representation of brain activity in the segments of EEG signals.
2. Retrieve a labeled sample of the desired reference class. In this study a class representing an approximated rest state will be employed as reference.
3. Fit a Gaussian mixture model to the reference class.
4. Compute loglikelihood of each data point with respect to the fitted Gaussian mixture model.
5. Employ the loglikelihood of points as a continuous detection signal. In this study this measure will be employed to give meaning to the structure of the simplicial complex obtained from the application of the Mapper method.

In the original proposal of Schalk et all [6], the Gaussian mixture model (GMM) is fitted to the reference class by employing and Expectation-Maximization procedure complemented with the Akaike Information Criterion to automatically determine the number of gaussian distributions in the mixture.

But in this study a more promising approach called free split/merge expectation maximization (FSMEM), presented by Wagenaar [12] and developed as an extension to the work of Ueda et all [13], will be used to fit the GMM with an automatic detection of the number of Gaussian distributions in the mixture. In addition this approach will also be employed as a soft clustering alternative when determining the clusters of the subdatasets corresponding to intervals of the low dimensional mapping when implementing the Mapper method.

4. EXPLORATORY FRAMEWORK (PARTIALLY DEVELOPED)

The main idea behind this study is to propose a framework that would allow any subject to set up a BCI by exploring the personalized and dynamic relationship between his self defined tasks on specific contexts of action and the EEG signals features' space derived from brain activity.

There are many challenges behind this idea. The first one is to counteract the signal identification problem and paradox. This is the main motivation to adopt the detection instead of classification paradigm and apply some of the ideas behind SIGFRIED.

The second challenge is determining the moment at which a fundamental change on a feature of the EEG signal has taken place in a high time resolution and continous signal setting. This can be addressed by considering multiple time scales on the segments of the EEG signal with a relative time point in common. That translates into two possibilities: an even higher dimensional space characterizing a point in time by the features of different segment lengths in the signal or studying the persistent topological properties of different time scales. Then topological data analysis can play an important role in the development of the framework.

The third challenge is that labeling can not be exact because of the fast and noisy changes on the EEG signal. Even in the current most carefully set up experiments with labeling, a big segment is considered in which the action that should generate a change in a signal feature takes place. But the exact segment of the signal that should correspond to the realized task is quite difficult to define. This motivates considering multiple time scales, unsupervised learning and a probabilistic perspective on the likelihood of the cloud points to try to find patterns in the EEG signal.

The fourth challenge is that the feature space can become very high dimensional thanks to spatial resolution (electrodes distributed around the scalp) and the huge amount of possible features that can be obtained from a signal. This again motivates employing topological data analysis as a way to explore the high dimensional nature of this dataset without trying to make too many assumptions about which features are important from the start.

The fifth challenge would be to implement computationally efficient unsupervised learning algorithms to assist the subject. Since the feedback of the framework and the desired BCI should work as fastest as possible to give the subject an intuitive and viable experience. This motivates the exploration of clustering algorithms and methods optimized for the specific nature of the EEG signal dataset, like a stream collaborative clustering.

Finally it is of most importance to create an interface intuitive enough so the subject can easily learn to navigate and comprehend the feedback from the machine learning algorithms that assist him to be able to successfully discriminate and choose areas of the feature space to stablish the BCI control structure. It is also important for the framework to allow the implementation, in the future, of any fundamental breakthrough on neuroscience or on the condition and activities of the subject. This can be achieved by exploiting the visualization capacities of the mapper method and minimizing assumptions during the exploration process.

The framework steps can be summarized as:

1. Settting up the online/offline exploration data generation
2. Exploring as much fundamentally different features of the EEG signal as possible.
3. Exploring the possible meaningful functions that can be defined for detection of unlikely events which can constitute the BCI control structure and the visualization tools.
4. Implementing the Mapper method with considerations on the partition of the low dimensional mapping and the clustering algorithms that would be optimal for the nature of the EEG signal and the BCI construction problem.
5. Implementing bayesian methods to asses any task classification and confidence statistics on the topological persistent properties of the dataset.
6. Computing the persistent topological properties of the dataset.
7. Visualizing and interacting with the exploration framework to build the BCI.

The main objective of this study is to confirm, up to step four, the viability of the framework to find structure in the EEG signal features. So the development of the rest of the framework and its optimization remain as justified and necessary further research.

4.1 Online/Offline exploration data setup

In this study, we employ a dataset taken from the BCI competition IV[1] that is actually part of a bigger dataset from a study of Blankertz et all [14]. The selected dataset is labeled and contains a category that can be understood as rest before intructions to execute a task like moving the right or left hand or foot are presented to subjects.

We considered the training dataset belonging to subject A, consisting of 200 trials for which the subject was randomly asked to move his left hand or foot after a period of rest. In each trial, the first 2 seconds consist of a white screen, then a cross appear in the center of the screen for the next 2 seconds and finally the instruction to move the left hand or foot appears over the cross and remains for 4 seconds. An example of a trial with the time segments categorized can be appreciated

[1] Dataset published on: http://www.bbci.de/competition/iv/desc_1.html

in Figure 2.

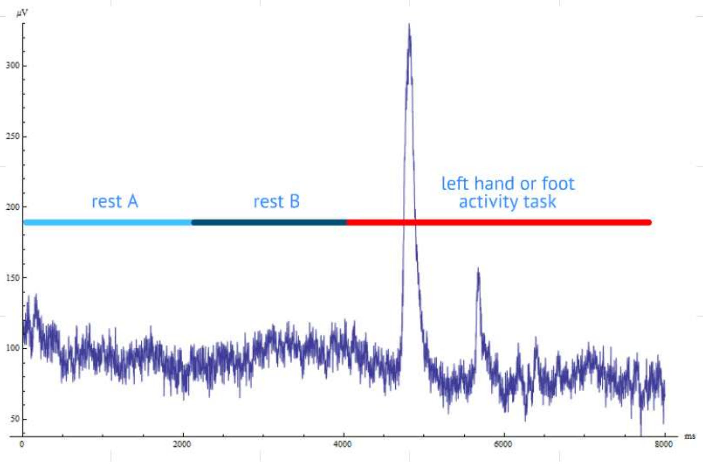


Fig. 2. Trial example with categories of time segments indicated

The EEG was setup with 59 channels (electrodes around the scalp), the signal have a time resolution of 1000Hz, which means a 1000 samples per second and was band-pass filtered between 0.5 and 200 Hz.There are in total four categories that are accurately labeled, two types of rest, left hand movement and foot movement.

Then we considered time segments of 300ms sampled consecutively from the signal of every trial, every 40ms in the case of the rest categories and every 20ms in the case of the movement categories. The length was selected to be 300ms since it is long enough to contain the big perturbations observed in the signals during the movement task. Nonetheless more time scales should be considered in future research.

So we ended up with a point cloud of 50200 points. Of these, 15200 points belonged to rest categories, 17500 points to left hand movement and 17500 points to foot movement. With a dimensionality of 17700 that correspond to the 300 time measures of the time series of each of the 59 channels (electrodes).

We did not consider a point every millisecond in this first approach of the framework for practical computational constraints in time and memory, since we would end up clustering datasets with millions of points in that case. Also signals are not expected to change radically from 1ms to the next so we would just be oversampling segments of signals with almost identical features.

4.2 Exploring features of EEG signals

The list of spatio-temporal features that have been considered for EEG signals in the literature is numerous. McFarland et all [15] makes a good review of the most popular ones and their application on BCI. From these the Fourier Transform is number one on the category of temporal features and will be employed in this study to show the capacity of the framework to capture structure in the EEG dataset.

To test the framework, the average of the power spectrum, computed with the Fast Fourier Transform in Mathematica, in the frequency bands Beta (13-30Hz), Gamma A (30-100Hz) and Gamma B (100-200Hz) was considered for each datapoint. An example of the power spectrum of a datapoint and the corresponding band frequencies can be appreciated in Figure 3.

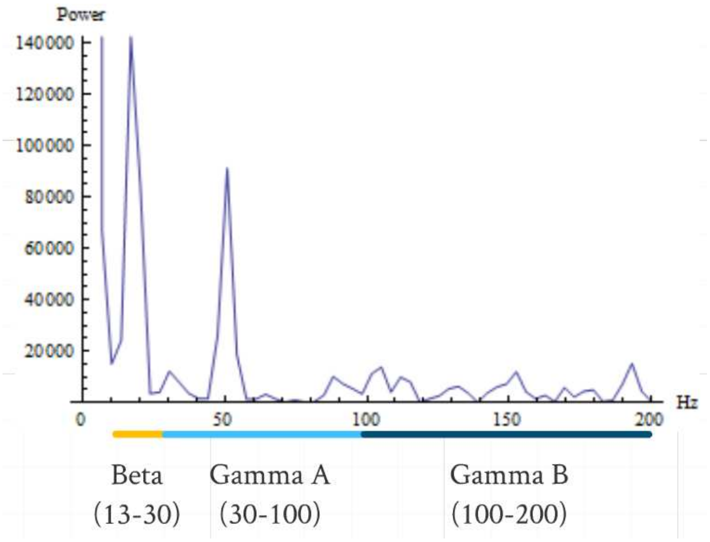


Fig. 3. Power Spectrum example with frequency bands indicated

Moreover we considered the averaged signal of the trials for each movement task.Then a subset of 12 channels was selected from the 59 channels set, based on the greatest difference between the average of the power spectrum in the mentioned frequency bands of the two tasks. The selected channels were "AF3", "Fz", "CFC5", "C5", "CCP5", "CP5", "P4", "P6", "PO1", "PO2", "O1" and "O2". How the difference on the average of power for the different tasks' averaged signal in the Beta band can be appreciated in Figure 4, as an example of the channel

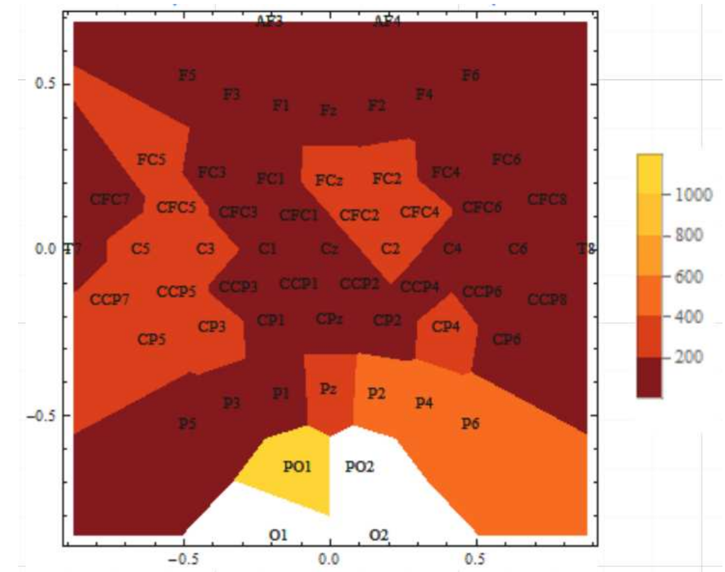


Fig. 4. Difference in the averaged power spectrum of the two movement tasks' averaged signal for the Beta frequency band (The channels are arranged as the 2d projection of their positions in the scalp) / (One can see the big difference in the channels PO1, PO2, O1 and O2, so these were part of the selected set of channels)

selection, although the Gamma bands were also considered.

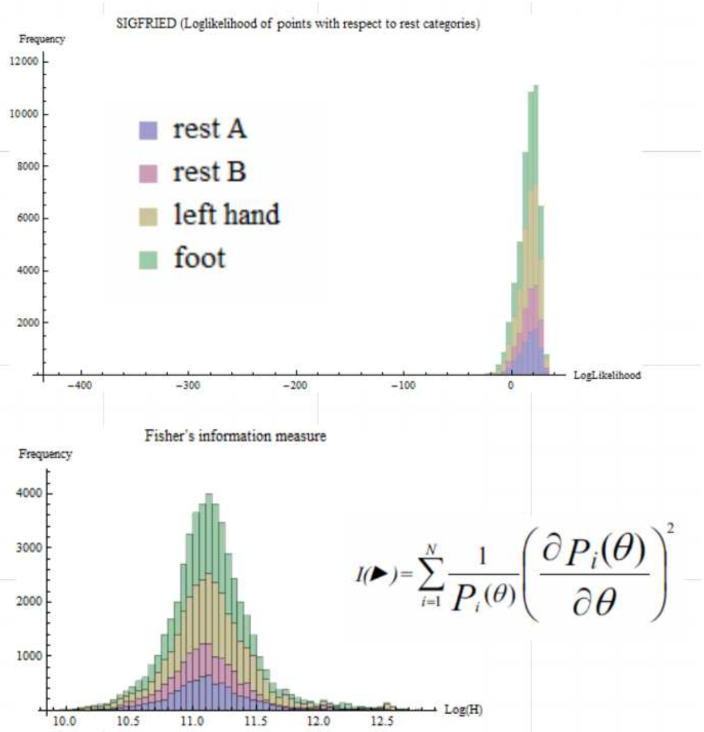
Finally each datapoint is represented by the average of the three frequency bands for each of the 12 selected channels, which results in a 36 dimensional representation of the cloud points. However in future research multiple additional features should be considered simultaneously. We are limiting the application of the framework to this feature extraction technique and to a smaller number of preselected channels, for practical computational constraints in the dimensionality of the data, particularly for working with soft clustering based on fitting a gaussian mixture model.

4.3 Exploring meaningful functions for detection

In this study we will explore the EEG signal point cloud from two different perspectives. The first one is implementing Schalk et all [6] proposal to fit a GMM to rest categories to then be able to discriminate samples of other classes based on their loglikelihood. In this way we would be characterizing extreme events of activity as a continuous function.

The second one consists on considering notions of complexity as defined by Christopher James et all [16] to characterize the brain activity. Particularly we will present the Fisher's information measure and contrast it with the perspective of SIGFRIED when analyzing the existence of structure in the EEG signal dataset.

In figure 5 the frequency of the loglikelihood of the points in the case of SIGFRIED and of the logarithm of the complexity measure can be appreciated. It is noticeable that the loglikelihood of SIGFRIED greatly separates a small portion of the points from the mountain of points likely to belong to rest activity, the loglikelihood of some points is even lower than -400 in contrast with the main distribution centered above 0. This makes a big contrast with the distribution of the values of the Fisher's information measure that seems to approximate more a normal distribution with short tails for all categories, which means that the complexity captured by the measure does not necessarily relate to the idea of rest vs non rest activity.



$$I(\blacktriangleright)=\sum_{i=1}^{N}\frac{1}{P_i(\theta)}\left(\frac{\partial P_i(\theta)}{\partial\theta}\right)^2$$

Fig. 5. Histogram of distribution of point categories for SIGFRIED and the information measure

As an additional clarification, the Fisher's information measure is obtained by: First constructing a matrix with consecutive overlapped subsegments of length 100 ms belonging to the segment of 300ms that represents the EEG signal point for each channel; Then stacking the matrices of all channels to create a global matrix for the point; After a singular value decomposition (SVD) is employed to get the singular values; Finally the singular values are used as probabilities in the information measure formula. The formula for the Fisher's information measure can be seen in Figure 5 and an illustration of the computation of complexity measures taken from the work of Christopher James et all [16] can be appreciated in Figure 6.

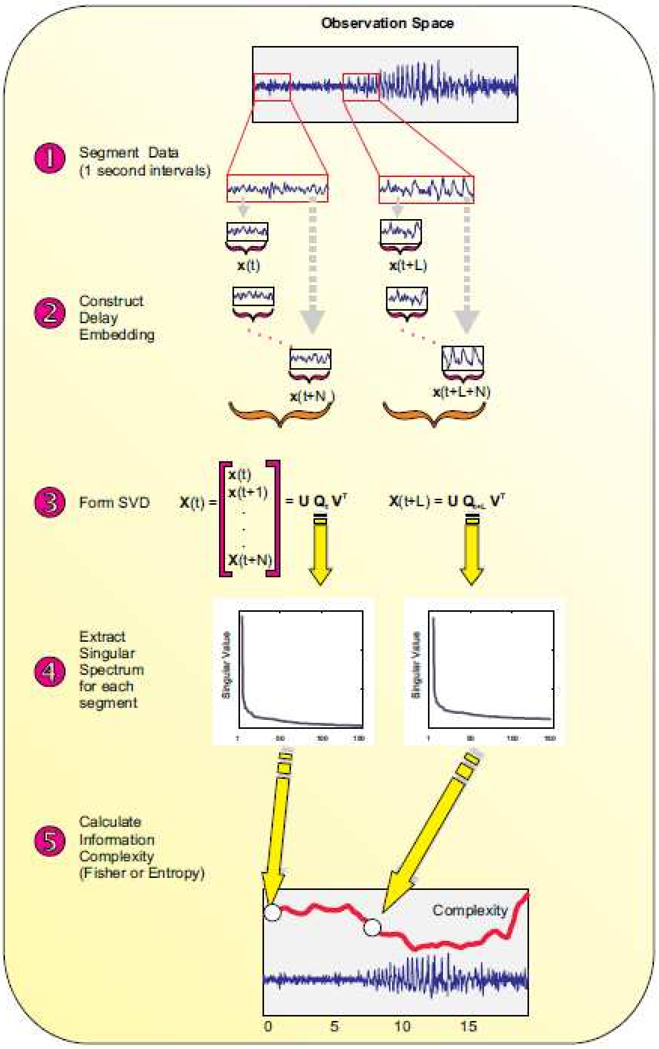

Observation Space
1 Segment Data (1 second intervals)
x(t)
x(t+L)
2 Construct Delay Embedding
x(t+N)
x(t+L+N)
3 Form SVD
X(t) = [x(t) x(t+1) . . . x(t+N)] = U Q_t V^T
X(t+L) = U Q_{t+L} V^T
4 Extract Singular Spectrum for each segment
Singular Value
5 Calculate Information Complexity (Fisher or Entropy)
Complexity
0
5
10
15

Fig. 6. Illustration of the computation of complexity measures

4.4 Implementing the Mapper method

The implementation of the mapper method realized in this study, considering the same structure of steps presented before, can be summarized as:

1. The dataset if formed by the cloud of points of segments of the EEG signal, also represented by a 36 dimensional vector that encodes the information of the average power spectrum for 3 different band frequencies for 12 different electrode locations in the scalp.
2. The SIGFRIED methodology and the Fisher's information measure will be employed as the meaningful functions to map the point cloud to a low dimensional space.
3. The low dimensional mapping will be partitioned by different number of intervals and varied overlaps to appreciate changes on the data structure due to resolution considerations.
4. The subdatasets corresponding to the defined intervals in the low dimensional space are generated.
5. Two different clustering algorithms will be employed with the Mapper method. The first one, as proposed by the Mapper method [5], is single linkage clustering [17] with the addition of an automatic detection on the number of clusters employing the Silhouettes statistic [18]. The second one is fitting a GMM with FSMEM[2] [12] as in the SIGFRIED methodology to explore the application of soft clustering for further Bayesian analysis in posterior research. In this way we obtain the nodes of the simplicial complexes.
6. The intersections of clusters belonging to consecutive overlapped intervals are evaluated to obtain the connections of the simplicial complexes.
7. The simplicial complexes are built for further analysis.
8. Pie charts will be employed instead of nodes to represent the proportion of the categories of points inside every node. Also a coloring of nodes will be employed as an alternative representation to show the value of the meaningful functions associated with the node's intervals. Moreover the simplicial complexes will be spatially arranged sometimes in such a way that it can be interpreted from the perspective of the meaningful function and the pie charts at the same time.

4.5 About the obtained simplicial complexes

The final result of the mapper method are the simplicial complexes on which persistent homology can be applied to reveal the most relevant and persistent topological properties of the underlying object, which in this case is the brain processes and tasks represented by the EEG signal features.

But before applying persistent homology, it is important to confirm that some interesting structure is being captured by the Mapper method and the functions defined to implement it. So considering that the power spectrum of white noise approximates a constant value we can model how the simplicial complex of completely unstructured white noise would look like and compare it with the structures of the simplicial complexes that we get with the proposed functions and features of the EEG signal.

In figure 7, an important difference between the structure of white noise and the methods employed to analyze the EEG signal dataset can be observed. Moreover there is also an important difference between the information measure and the SIGFRIED perspective on the structure of the generated simplicial complex. The nodes in figure 7 are represented as pie charts showing the proportion of the labeled categories that were clustered inside each node, and the spatial arrangement of the simplicial complexes is aligned with the notion of increaing complexity or decreasing loglikelihood as corresponds to the underlying perspective. In this case all the low dimensional mappings where partitioned in 20 intervals with a 50% overlap.

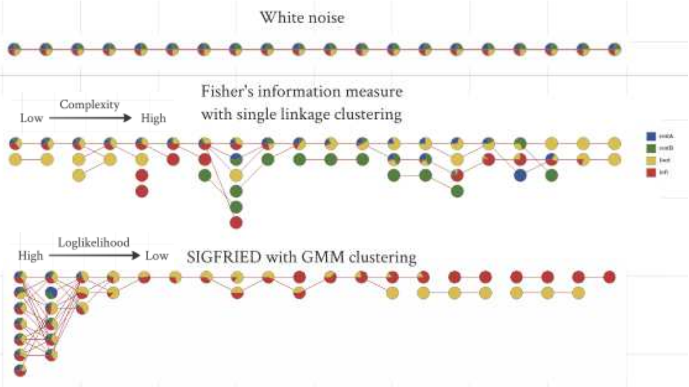

White noise
Low
Complexity
High
Fisher's information measure
with single linkage clustering
High
Loglikelihood
Low
SIGFRIED with GMM clustering

Fig. 7. Comparison of the structure of data in the simplicial complexes produced under the fisher's information and SIGFRIED perspective against white noise.

It seems that SIGFRIED effectively improves the discrimination capacity of the framework to separate the rest from non rest tasks and then also improves the capacity to further find important differences between the non rest tasks in an unsupervised way. But on the other hand the information measure is showing more patterns and structures that although might not be directly connected with the defined tasks, might give some important insight into some brain processes or different tasks in a different context, which is also important if we are exploring the feature space on an unsupervised way.

This might imply that both functions, SIGFRIED and the information measures, can be useful and perhaps complementary. Suggesting the possibility of combining them on a two dimensional mapping of the dataset when projecting the point cloud on a low dimensional space to partition it during the implementation of the Mapper method

Furthermore, in Figure 8, it is possible to see the rich structures that arise at a different level of resolution (considering 40 intervals with 50% overlap) in the case of the information

[2] Matlab code obtained from : http://www.mathworks.com/ matlabcentral/fileexchange/22711-free-split-and-merge-expectation-maximization-for-multivariate-gaussian-mixture

measure perspective. This confirms the potential of the proposed methods in the exploratory framework to discover structures in the EEG signal, possibly allowing the desired development of a BCI and a better comprehension of particular brain processes for a specific subject.

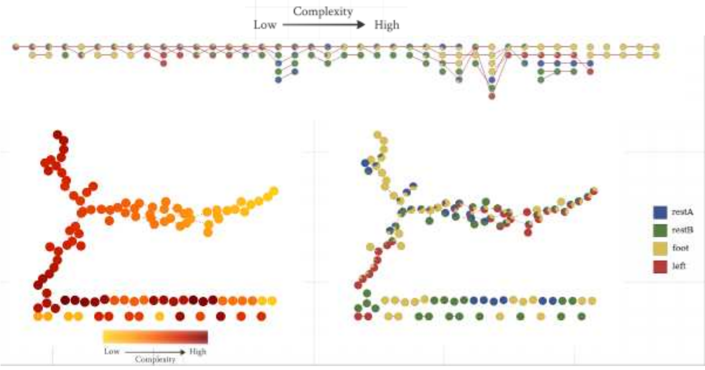


Fig. 8. Simplicial complex represented visually in different ways, produced under the Fisher's information measure and single linkage clustering.

In addition to Figure 8, Figure A1, in Appendix A, shows the important structural changes that can be seen in the generated simplicial complexes at different resolutions (different partitions on the low dimensional mapping and percentage of overlap). This shows the importance of applying the tools of persistent homology to be able to establish which of the appreciated topological features of the complexes are actually approximating the underlying object to the point cloud.

5. NECESSARY FURTHER DEVELOPMENT OF THE FRAMEWORK

After checking the potential of the Mapper method and in general of the notions of persistence homology, and of the perspective of detection instead of classification to explore EEG signals to build a BCI, there is still plenty to develop to complete a preliminary full implementation of the framework.

As part of the topics that need further development we can consider mainly:

1. Extending the use of Bayesian methods for detection and classification in framework, so that confidence measures can be taken at the different stages of the methodology.
2. Implementing clustering mechanisms optimized for the nature of the EEG signal dataset and the persistent homology techniques. Lets consider the high throughput nature of the EEG signal, the need to cluster different time scales and the possibility of understanding channels as diferent populations with similar features. Then it would be reasonable to propose the implementation of a stream collaborative soft clustering inspired on the works of Song et all [19] and Pedrycz et all [20]. In addition it would be interesting to consider the ideas of Chazal et all [21] on clustering also based on persistent homology.
3. Implementing the persistent homology analysis to characterize the persistent topological properties of the EEG signals in time, space and resolution.
4. Implementing additional techniques to extract information from the simplicial complexes, like the appearance and persistence of branching structures that are not captured by homology.
5. Extending the number of meaningufl features for detection and for the clustering of EEG signal's point cloud.
6. Defining the visualization methods that will be employed to receive feedback and interact with the framework once it is implemented. To employ a two dimensional meaningful mapping on the Mapper method seems like a good alternative in contrast with the current one dimensional mapping.
7. Finally most of the algorithms and computations at different scales can run in parallel, so it would be crucial to exploit GPU parallelization to turn the framework into a fast and responsive interface for a BCI.

6. FINAL REMARKS

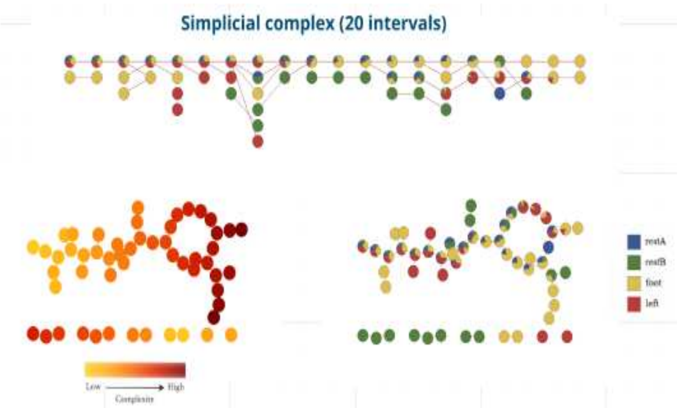
This study have presented important challenges on the construction of BCI and at the same time proposed a framework that might encode a solution to them, based on new mathematical methodologies and new paradigms in the BCI community.

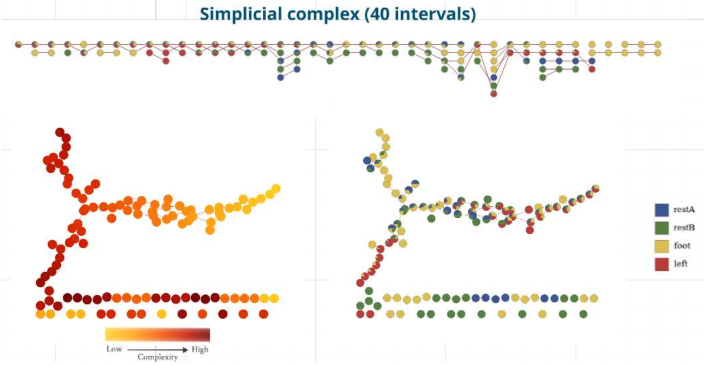
The preliminary results of the potential of the proposed framework and its methods to characterize EEG signals, to understand the relationship between tasks and brain activity and very likely to construct in an easier and more accurate way a brain computer interface, seem very promising.

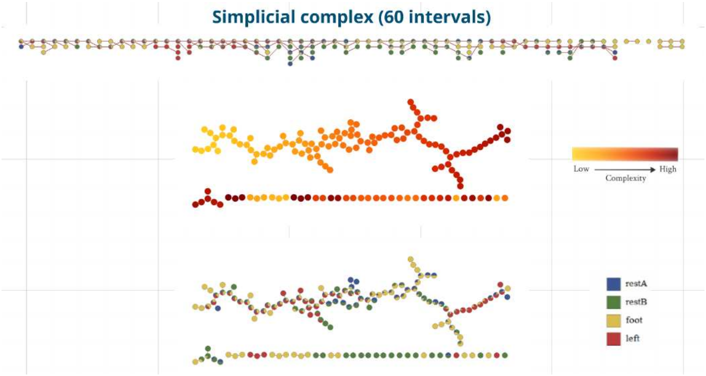
Nonetheless there is still quite a lot ahead to develop and explore before claiming the usefulness of the framework or settling down for the specific algorithms, methods and perspectives that assisted the different steps of the proposed exploratory framework.

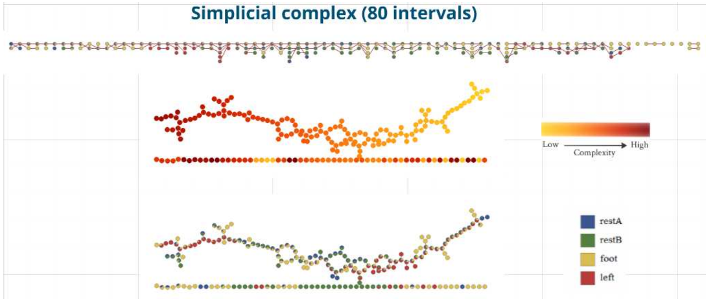
7. APPENDIX A (SIMPLICIAL COMPLEXES)

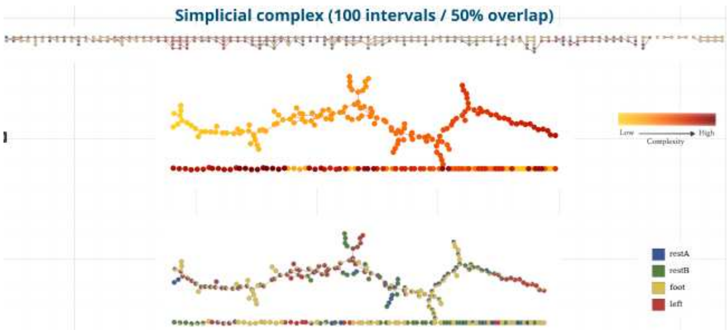
In the following Figure A1, the effect of different levels of resolution due to changes on the number of intervals and the percentage of the overlap during the implementation of the Mapper method can be appreciated.

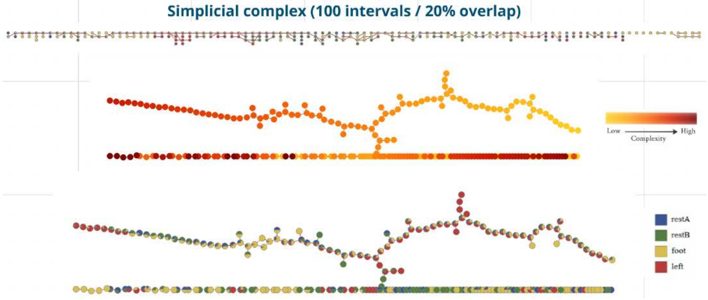












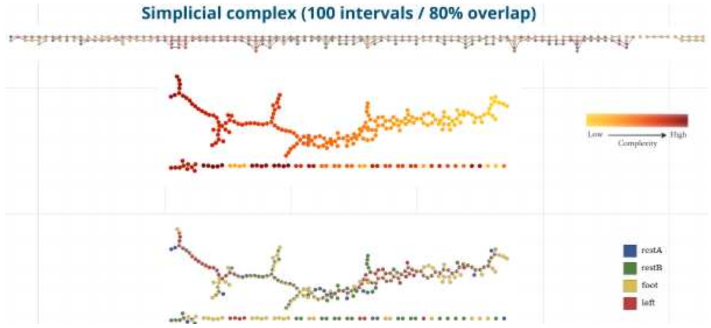


Fig. A1. Simplicial complexes based on the Fisher's information measure and single linkage clustering, with different resolutions of intervals with a 50% overlap followed by different resolutions of overlap with a 100 intervals.

8. ACKNOWLEDGEMENT

First of all, thanks to Christopher James for his dedication and commitment as supervisor of this miniproject. Thanks to Simon Davies and Sylvester Rozario from the Neuroengineering lab for their support. And thanks to the professors of the complexity centre for their time and help, particularly to professors Robert Mackay, Charo I del Genio and Colm Connaughton.

9. REFERENCES

[1] Olejniczak, P. (2006). Neurophysiologic basis of EEG. Journal of clinical neurophysiology, 23(3), 186-189.

[2] Campbell, A., Choudhury, T., Hu, S., Lu, H., Mukerjee, M. K., Rabbi, M., & Raizada, R. D. (2010, August). NeuroPhone: brain-mobile phone interface using a wireless EEG headset. In Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds (pp. 3-8). ACM.

[3] Picton, T. W. (1992). The P300 wave of the human event-related potential.Journal of clinical neurophysiology, 9(4), 456-479.

[4] Donchin, E., Spencer, K. M., & Wijesinghe, R. (2000). The mental prosthesis: assessing the speed of a P300-based brain-computer interface. Rehabilitation Engineering, IEEE Transactions on, 8(2), 174-179.

[5] Singh, G., Mémoli, F., & Carlsson, G. (2007, September). Topological methods for the analysis of high dimensional data sets and 3d object recognition. InEurographics Symposium on Point-Based Graphics (Vol. 22). The Eurographics Association.

[6] Schalk, G., Brunner, P., Gerhardt, L. A., Bischof, H., & Wolpaw, J. R. (2008). Brain–computer interfaces (BCIs): detection instead of classification. Journal of neuroscience methods, 167(1), 51-62.

[7] Vidaurre, C., & Blankertz, B. (2010). Towards a cure for BCI illiteracy. Brain topography, 23(2), 194-198.

[8] Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.

[9] Nicolau, M., Levine, A. J., & Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences, 108(17), 7265-7270.

[10] Ghrist, R. (2008). Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society, 45(1), 61-75.

[11] Balakrishnan, S., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., & Wasserman, L. (2013). Statistical Inference For Persistent Homology. arXiv preprint arXiv:1303.7117.

[12] Daniel Wagenaar (2000). FSMEM for MoG. http://www.danielwagenaar.net/res/papers/00-Wage2.pdf

[13] Ueda, N., Nakano, R., Ghahramani, Z., & Hinton, G. E. (1998). Split and merge EM algorithm for improving Gaussian mixture density estimates. In Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop (pp. 274-283). IEEE.

[14] Blankertz, B., Dornhege, G., Krauledat, M., Müller, K. R., & Curio, G. (2007). The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects. NeuroImage, 37(2), 539-550.

[15] McFarland, D. J., Anderson, C. W., Muller, K. R., Schlogl, A., & Krusienski, D. J. (2006). BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation. Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 14(2), 135-138.

[16] Lowe, D., James, C. J., & Germuska, R. (2001). Tracking complexity characteristics of the wake brain state.

[17] Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

[18] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[19] Song, M., & Wang, H. (2005, March). Highly efficient incremental estimation of gaussian mixture models for online data stream clustering. In Defense and Security (pp. 174-183). International Society for Optics and Photonics.

[20] Pedrycz, W., & Rai, P. (2008). Collaborative clustering with the use of Fuzzy C-Means and its quantification. Fuzzy Sets and Systems, 159(18), 2399-2427.

[21] Chazal, F., Guibas, L. J., Oudot, S. Y., & Skraba, P. (2011, June). Persistence-based clustering in Riemannian manifolds. In Proceedings of the 27th annual ACM symposium on Computational Geometry (pp. 97-106). ACM.