# UNSUPERVISED CLUSTERING OF

# SCHIZOPHRENIA PATIENTS, COCAINE USERS,

# AND NORMAL CONTROLS

# BASED ON AUDITORY EVOKED POTENTIALS

---

A Thesis

Presented to

the Faculty of the Department of Engineering Technology

University of Houston

---

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in Engineering Technology

---

By
Ainhoa Bastarrika Iriarte

August, 2012

*Objective:* In this Master Thesis, we employed auditory evoked responses to identify features from the time domain signals, Granger Causality matrixes and graph theory that provide maximum separation among schizophrenic patients, cocaine addicts and normal controls.

*Methods:* We analyzed data from 12 schizophrenia, 12 cocaine addicts and 12 normal control subjects. Responses were obtained in a paired-stimulus paradigm, in which auditory stimulus $S_1$ is followed by an identical $S_2$. Amplitude and latency of the N100 component were measured from the averaged evoked potential, Granger Causality matrixes were computed and a graph was constructed from there. Amplitude, latency, Granger Causality values and a group of characteristics form the graph were used as features to cluster responses in three groups. Several methods were used for clustering, while their performance was quantified in a 10-fold cross validation approach.

*Results:* We found that the most important features come from Granger Causality matrix values, amplitude and latency coming from the average evoked responses appear as insignificant features for the clustering. Influence of electrode C3 to Pz appear as the most significant feature, separating schizophrenic patients from normal controls and cocaine addicts with a 100% accuracy. In order to separate cocaine addicts from normal controls at least 31 features coming from Granger Causality matrixes were needed.

*Conclusions:* Our results demonstrate that Granger Causality values can accurately separate schizophrenia patients, cocaine addicts and normal controls and suggest that the Pz-C3 region plays a significant role in information processing in human brain.

*Significance:* The proposed technique may have a significant impact as a clinical tool in the quest for identifying physiological markers of schizophrenia.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Schizophrenia is a chronic, debilitating mental disorder that affects about 1 percent of the population in the United States, which corresponds to more than 3 million people [31]. It is a psychotic mental illness and is characterized by symptoms of thought, behavior, and social problems. The thought problems associated with schizophrenia are described as psychosis, in that some times the person's thinking does not match with reality. For example, the sufferer may hear voices or see people that are in no way present. The individual with this disorder may also have disorganized speech, disorganized behavior, physically rigid or lax behavior (catatonia), significantly decreased behaviors or feelings, as well as illusions, which are ideas about themselves or others that have no basis in reality (for example, the individual might experience paranoia, in that he or she thinks others are plotting against them when they are not).

On the other hand, over the last few decades the use of drugs has increased all over the world [35][33]. Many cocaine user develop both physiological and psychological dependency on a regular use of this drug. The consequences of an excess of cocaine may be physiological problems, lethargy, psychosis, depression, akathisia, and fatal

overdose. This work presents a study towards finding biological markers based on brain activity analysis that would help classify schizophrenia patients and cocaine addicts from healthy subjects.

## 1.1 Technology and illness recognition systems

Over the years, technology and science have helped medicine obtain more accurate diagnostic tools for a wide range of pathologies. Nevertheless, the human body still has lots of black boxes full of mysterious. Step by step, technology and science try to unlock them to obtain a better understanding of ourselves.This is the case of the brain, a human organ with lots of unknown features and processes waiting to be studied and understood.

It is desirable to develop a non-invasive neurophysiological test that can help physicians diagnose and differentiate various subtypes of these deseases, which share several symptoms.

For the moment, automated classification of Normal Control (NC) and schizophrenic (SZ) have been researched using Auditory Evoked Potentials (AEP) single trial responses [38] [14]. AEP are brain responses generated due to auditory stimuli. Recent research has done some work on NC and SZ classification based on ICA and single trial analysis[38] [14].

## 1.2 Problem statement

This Master Thesis presents a typical clustering-classification problem: given a data set of brain signals recorded from 36 subjects classify each subjects as schizophrenic, cocaine addict, or healthy. As a new approach, features extracted from Granger

Causality (GC) Networks and the graph theory combined with features extracted from the time domain AEP responses, like amplitude and latency will be used.

## 1.3 Research objectives

The long term goal of this line of research is to define biological markers based on brain activity signals that describe different mental states and disorders and highlight the differences between groups of subjects, in order to get a better classification. This thesis focuses on the next specific objectives:

- Identify a set of features that best separates the 3 groups. Previous experiences will be relayed with a similar group of data that consisted of two groups only and explore wether the three groups are still separable.
- Identify the minimum set of features and simplest algorithms that provide maximum performance as in specific objective 1.
- Explore the effects of various parameters on the accuracy of the results obtained.

# Chapter 2

# Background

## 2.1 Brain activity signals

The brain is the core of the nervous system. Moreover, the nervous system is responsible for emotions, actions, and thoughts. To wire an analogy, the nervous system is the controlling system of our body and the brain the microprocessor of that control system. Brain tissue can be divided in white and the gray matter. The gray matter is a network of cells, called neurons, which is the part of the brain involved responsible for processing and control. Continuing with the analogy, it would be the cores that a multicore microprocessor system. The difference is that instead of having 4 cores like the 2nd Generation Intel CoreTM i7, the gray matter has about $10^{11}$ neurons. The white matter is the tissue from where the messages of the gray matter pass. In the previous analogy it would represent the connections among the cores of the distributed multiprocessor.The neurons are electrically excitable cells that process and transmit information [26].

The neurons communicate via synapses, that can be, via electrical or chemical signaling but almost all the time is via chemical synapses. However, before explaining

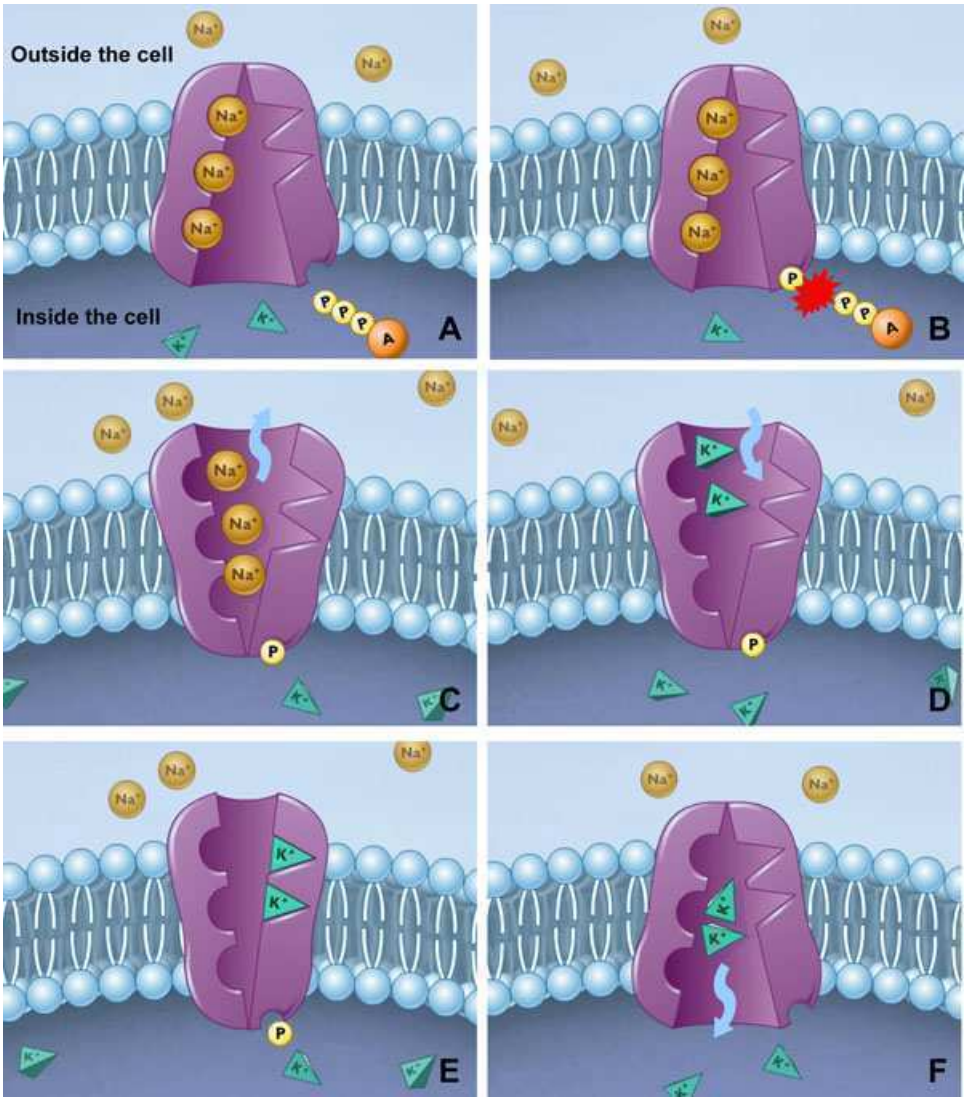



Figure 2.1: Na-K Exchange in a cell membrane [2].

how this communication occurs, is better to explain how the signals are generated. In the quiescent state, sodium [Na+] ions cannot diffuse through the plasma membrane and go in to the cell. However, when a neuron is externally stimulated, the sodium channels in the membrane are opened. Due to the fact that sodium concentration is much higher outside the cell, the Na+ ions will diffuse into the neuron [26].

This diffusion changes the potential of the cell. Now, the inside of the cell is more positive and the outside more negative. This process is called depolarization. If the stimulus is strong enough, the polarity of the cell is reversed and a signal called nerve impulse (or action potential is initiated). Then, it propagates over the entire axon. Almost immediately, the membrane becomes again impermeable to [Na+] ions but permeable to potassium [K+] ions. Consequently, [K+] ions diffuse out of the cell rapidly. This [K+] outflow establishes the electrical potential into a resting state. This state is called repolarization. After that, the sodium-potassium pump is activated and the initial concentrations of [Na+] and [K+] are re-established.

Once the nerve impulse is generated it should be transmitted from one neuron to the other. First of all, it should be stated that the impulse itself does not pass from one neuron to the next one. When the action potential reaches an axon terminal, the vesicles containing the neurotransmitters diffuse with the membrane. This causes porelike openings that form and release the neurotransmitter. The neurotransmitter molecules are diffused across the synapse and they reach the receptor of the next neuron. If a sufficient amount of neurotransmitter molecules arrive to the receptor, the process explained before (depolarization, action potential generation, repolarization) will occur and the nerve signal will be transmitted to the next neuron.

It is important to note that the transmission of an impulse is an electrochemical event. The transmission through the length of a neuron is a electrical transmission, however, the next neuron is stimulated chemically. All of the ion movements explained in this section are equivalent to a current flow (action potential). When an amount of cells signals synchronously, a combined current is obtained. This current can be modeled as a current dipole.

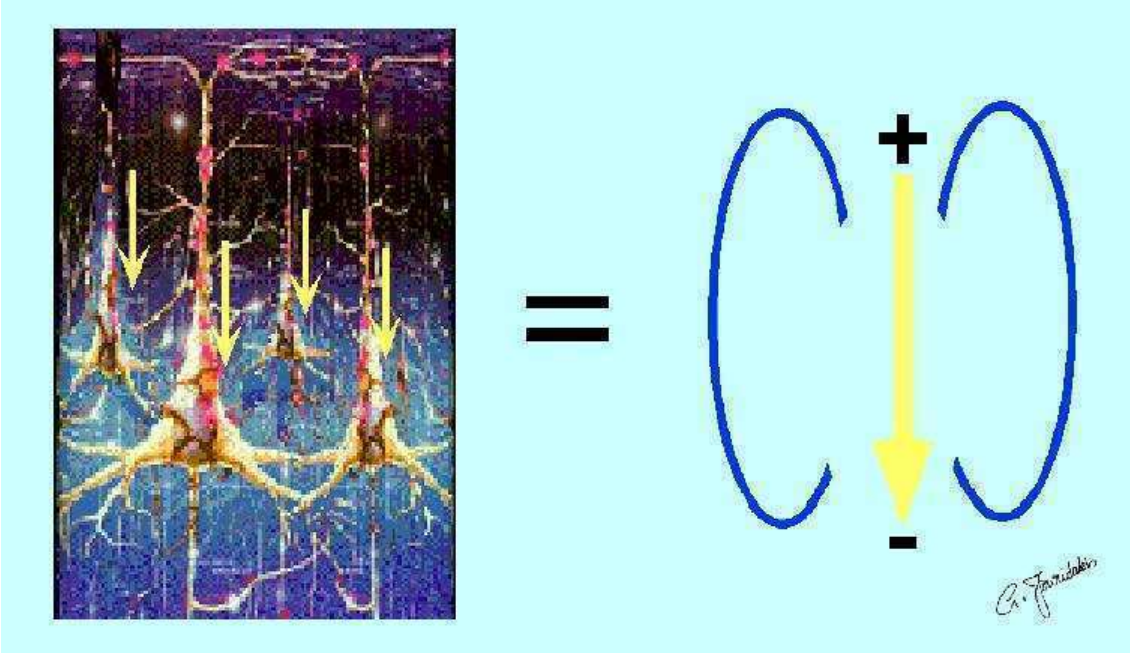



Figure 2.2: The current of a neuron can be represented with a dipole and the magnetic fields that it produces [4].

This signals, known also as primary signals, will propagate and arrive to the surface, they are known as secondary signals. On the other hand, any current that is associated with a magnetic field perpendicular to its direction is called magnetic flux [4].

The secondary signals will find an electric resistance, so when they arrive to the surface they will represent irregularly the primary signal. However, the permeability of the brain is similar to the empty space, and consequently the magnetic flux will not be distorted and will be proportional to the primary signal [4].

The secondary signals and the magnetic flux are the ones that are interesting to be recorded, processed, analyzed and studied to come up with conclusions of brain performance.

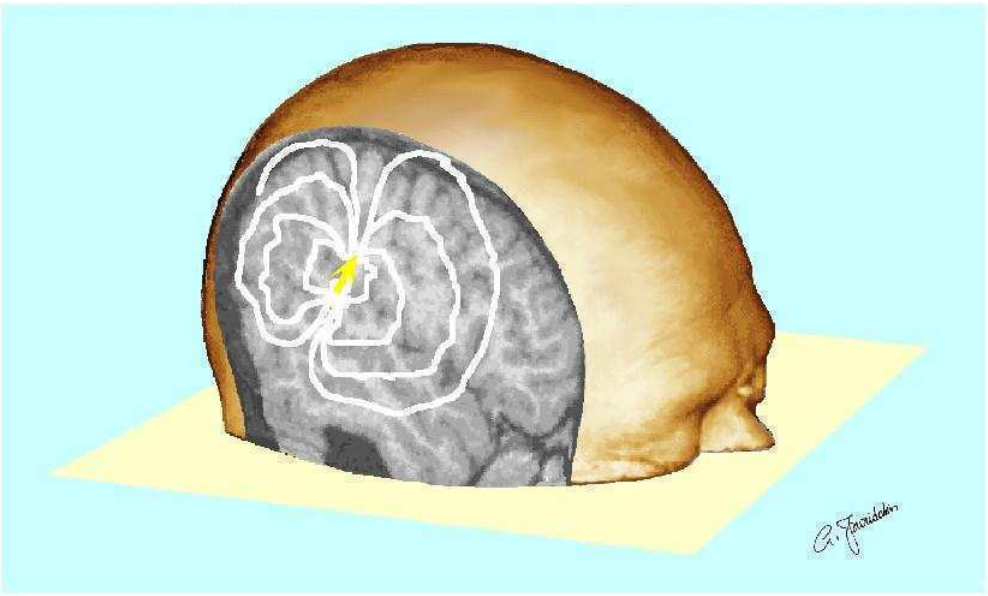


Figure 2.3: The brain activity can be modeled with a dipole and its magnetic fields[4].

## 2.2 Data recording

Nowadays there is a range of techniques to record data from the brain activity. Moreover, all these techniques can be divided into two groups, the invasive ones and the non-invasive. In this project only non-invasive procedures are used, so this section will focus only on the description of the non-invasive technique used during the project, Electroencephalography (EEG).

### 2.2.1 Electroencephalography - EEG

EEG is the technique used in this project for data recording. The EEG technique records the electrical field generated by a population of active neurons. In order to record brain signals, a set of sensors, called electrodes, is connected to the scalp using a conductive gel. This gel is used to reduce impedance, which allows for a better signal detection. The electrodes can be placed in different configurations. However,

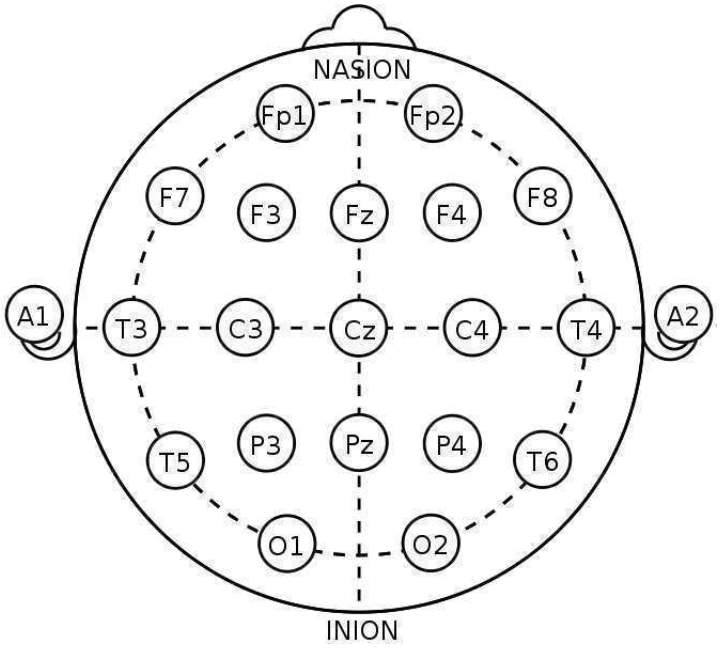



Figure 2.4: International 10 20 electrodes position [3].

the most common one is the International 10-20 system.

An EEG uses multiple electrodes. The number of used electrodes varies depending on the model, and it can go from 8 to 256 electrodes. Each channel (or electrode) is independently amplified with respect to a reference usually placed on the ear lobes. The electrodes and the amplifiers are connected by wire to a computer, so the data can be displayed on a screen, printed on paper, or saved into a file. Nowadays wireless devices for EEG can be found. Figure 2.5 shows a typical EEG recording session. The data saving file can change from one model to another, but commonly they describe the Numbers of channels, type of data, sampling frequency or total time are some typical header's parameters. The rest of the data are the recordings itself and are stored in a matrix of dimensions NxM, being N the number of channels and M the total number of points.

EEG can be used in two different procedures, continuous data recording or epoch

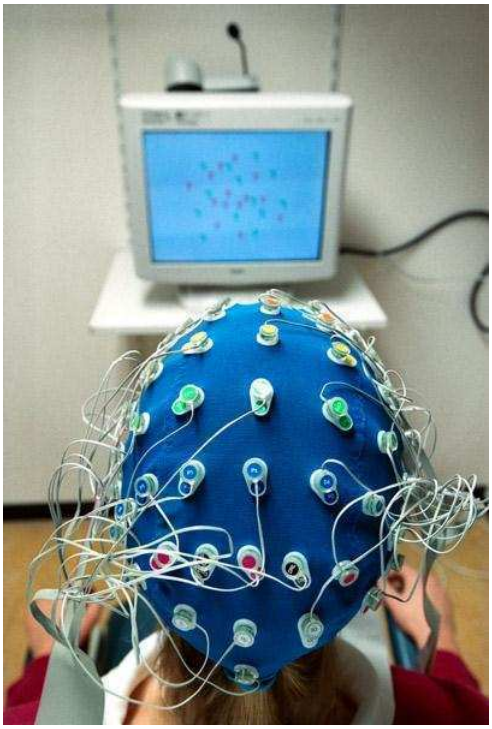


Figure 2.5: An EEG device with the electrodes connected to a screen [1].

recording. Epoch recordings are records done on repetitive experiments. Every epoch or trial is a data frame that contains a short period of time before the stimulus (pre-stimulus) and the response to the corresponding stimulus. In the case of this project, 120 trials for auditory stimuli were recorded.

## 2.3 Auditory evoked potential (AEP)

As it was explained before, the EEG shows the ongoing activity of the brain. It is not easy to pick up individual stimulus (a sound, a flash, a touch) due to the fact that they are usually hidden in the general brain activity.

Evoked potentials are electrical responses produced in the nervous system that occur owing to some kind of stimulus. Usually it is concerned to a stimulus of one of

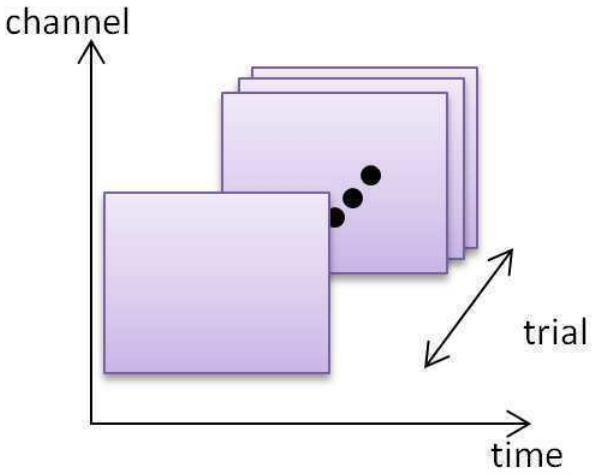



Figure 2.6: Trial based recording data structure.

the senses - hearing, vision or sensing.

Normally single evoked potentials are very small and indistinguishable from the ongoing EEG activity. To solve this problem, a number of responses are recorded and averaged. Then the signal appears clearly. All the activity that is not related with the stimulus (the ongoing EEG activity) disappears when the responses are averaged because of their randomness, allowing to see the activity related with the stimulus easily. Moreover, different processing can be applied to original trials in order to achieve a better performance.

Auditory Evoked Potentials are the evoked potentials that use sounds as stimulus. Various types of sounds can be used; however, beeps are the most usual ones for investigating purposes. Normally the response signal is the as follows:

An auditory evoked potential has several components, the most important one for this project is the N100 (in some literature also named as N1). It is a negative peak that occurs around 100ms. To collect AEP responses different experiments can be done. In the case of this project responses were measured on a paired-stimulus

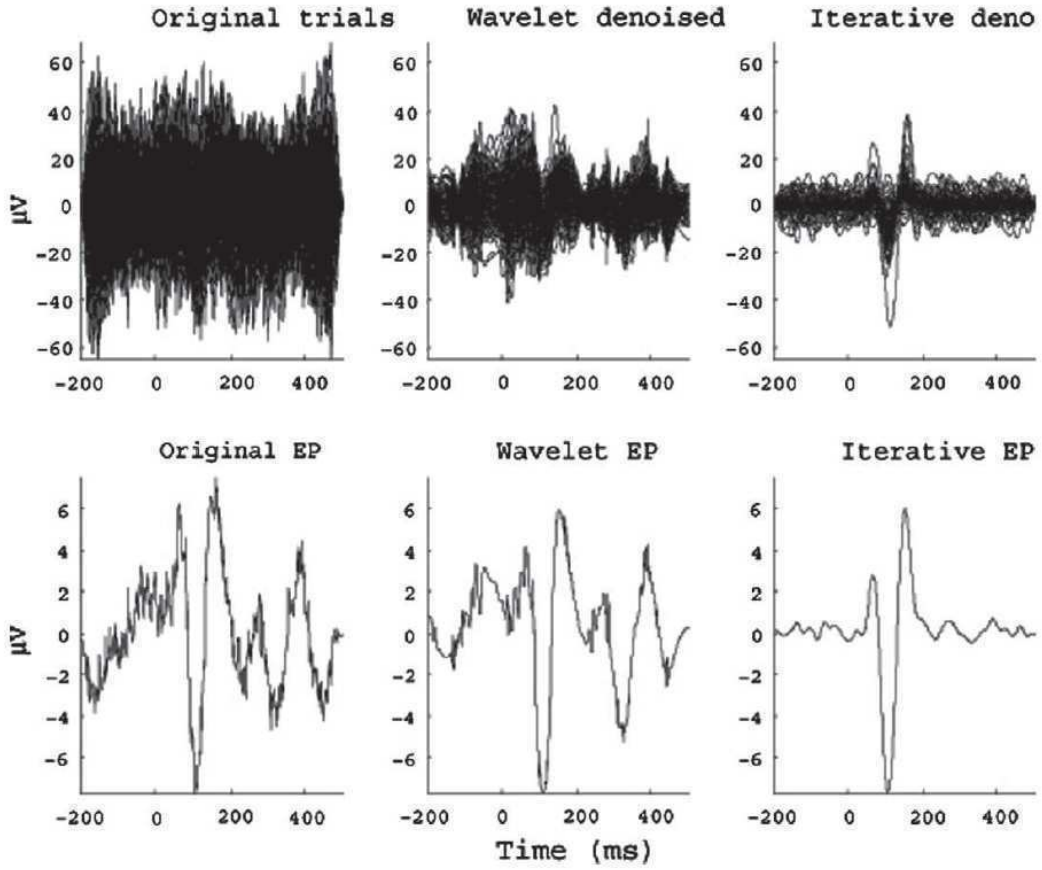



Figure 2.7: (Top row) Original single-trial responses (left) and the resulting denoised ones obtained using WT (middle) and iICA (right) processing. (Bottom row) The average EP estimates obtained by each method [14].

paradigm, in which auditory stimulus $S_1$ is followed by an identical $S_2$.

## 2.4 Previous work on schizophrenia patients and Cocaine users

### 2.4.1 Auditory evoked potential AEP

In this work responses were obtained using a paired-stimulus paradigm and the expected wave form for a normal control in this experiment is shown in Figure 2.8 . It can be appreciated that in NC the S2 response (grey), is much smaller than the

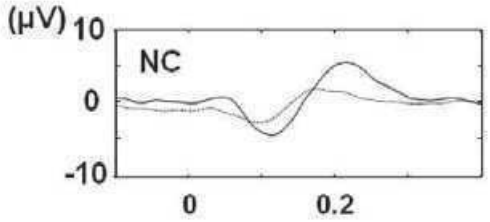



Figure 2.8: Expected response from a normal control (NC) in a paired-stimuli AEP experiment. Black line corresponds to S1 response and the grey to S2 [15].

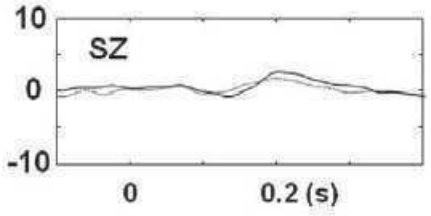



Figure 2.9: The expected response from a schizophrenic subject (SZ) in a paired-stimuli AEP experiment. Black line corresponds to S1 response and the grey to S2[15].

S1. This matches with the gating-out of the NC; the first beep is more relevant than the second one, because it has no new information. On the other hand, the expected wave form for a schizophrenia patient is shown in Figure 2.9. It can be observed that the amplitude of both S1 and S2 are much smaller than in a normal control. Also for a schizophrenia patient S1 and S2 are very similar, there is no gating-out.

### N100

In the case of N100 component also for NC the amplitude in S1 is greater than in S2. However, for SZ the difference between S1 and S2 amplitudes is smaller. In general, the schizophrenia population tended to respond later, as it can be seen in the mean

| | Amplitude | | Latency | |
|---|---|---|---|---|
| | S1 | S2 | S1 | S2 |
| NC | -8.5 ± 7.8 | -3.4 ± 5.6 | 114.6 ± 22.4 | 100.5 ± 25.9 |
| SZ | -5.2± 6.1 | -2.6 ± 4.6 | 144.6± 31.5 | 122.8 ± 31.9 |

Table 2.1: Mean and standard deviation of the amplitude and latency of the component N100 of S1 and S2 from both normal controls (NC) and schizophrenic patients (SZ). [14]

component latency, and their responses are highly variable when compared with NC, as the large variance of component latency shows.

### 2.4.2 Connectivity Networks

In 1999 Peled described schizophrenia as a disconnectivity disorder[29]. Later on, other researches found that aberrant connections between auditory area and language regions could explain the language impairments found in schizophrenic patients [21][22][32]. This is related with the concept of reduced fronto-temporal connectivity [19][23]. However, schizophrenic patients show higher connectivity between attention and language network . These findings have been supported by different studies [17][16][9].

On the other hand, Cocaine Users show an important decrease of connectivity strength in MCL seed regions. Moreover, decreased connectivity has been found between the rostral ACC and amygdala and rostral ACC and hippocampus. Same study numbered different brain areas that showed reduced connectivity strength. Nevertheless, NC and CA do not show big differences in connectivity when seeds are placed on primary motor and primary auditory cortics [10].

# Chapter 3

# Methodology

## 3.1 Subjects

The subjects that participated on this project were carefully selected to avoid influences of other health issues not related with this study. The schizophrenia patients group was formed by 3 females and 9 males, from where 4 of the subjects are white and 8 black. The average age is 42.17 years and the standard deviation 10.50 years. Nine of the subjects were classified as smokers and 3 as non-smokers. On the other hand, the healthy control group included 7 females and 5 males, from where 10 subjects are black and 2 white. The average age in this group is of 42.17 years and the standard deviation of 7.16 years. Two of the subjects were labeled as smokers and 10 as non-smokers. As for the cocaine dependent group, there were 2 females and 10 males, where 1 subjects was white, 1 oriental and the other 10 black. The average age is 42.17 years and the standard deviation 9.22 years. Two of the subjects were identified as non-smokers and the other as smokers. A more detailed description can be find in the appendix A: Subjects description.

## 3.2 Recording procedure

Auditory EPs were recorded from 12 normal controls, 12 schizophrenia patients, and 12 cocaine addicts using a paired-stimulus paradigm, where two identical 1kHz stimuli. Data were acquired from 22 electrodes with a bandpass filter set between 0.3-100Hz and digitalized with a sampling rate of 1kHz.

## 3.3 Data

The dataset of this project is composed by EEG signals recorded among 36 subjects. The data set contains three groups: schizophrenia patients (SZ), normal controls (NC) and cocaine addicts (CA) equally distributed (12 subject in each group). For each subject two DAT files are given, one for S1 and one for S2, making a total of 72 files. All subjects' EEG were recorded under the same montage conditions. (See below the list of electrodes' placements). Each DAT file (written in ascii) contains a matrix of size 84120 rows x 24 columns. Each columns match the channel labels listed below. For each subject there are 120 trials of 701 samples each one (701x120 = 84120 rows). The sampling rate of the recorded signals is 1000Hz, so each sample corresponds with 1 ms of the recording and the measured interval is [-200 500ms] (701 samples/trial). During data acquisition analog filters were set to 0.3-100Hz and in most of the cases a digital notch filter centered at 60HZ was applied. Each data file can be opened in Excel (original delimiter: tab; checking space during import wizard) and Matlab (load('file')). Data has been labeled from Subject01 S1.dat to Subject36 S2.dat.

The placement of used electrodes is listed below.

Fz Cz Pz F3 F4 C3 C4 P3 P4 T3 T4 T5 T6 M1 M2 Nz O1 O2 F7 F8 Fp1 Fp2

VEOG HEOG

## 3.4 Preprocessing: filtering and iICA

Although the stimuli used is only an auditory stimuli, the recorded data have not only the response to that stimuli beep, but also contains ongoing activity of the subject, artifacts like eye blinking, muscle movement, etc. Having a clean data is essential for successful brain signal analysis and classification. Noise can lead to a wrong feature extraction and this can end up in an incorrect description of the subject and an erroneous classification. This section describes the preprocessing applied to the signals in order to decrease the noise level and the iICA method used to enhance the N100 component on single trials.

### 3.4.1 Artifact removal

The first step of the preprocessing is the artifact suppressing. The aim of this step is to remove all signals that were not produced as a response to the auditory stimulus such as eye blinking. When one artifact is detected in any channel that epoch is removed from all the channels of both responses, S1 and S2, in order to allow trial comparison. The objective of such strict removal is to work only with acceptable repetitions of the experiment and to simplify the comparison between trials of S1 and S2.

The first step is to filter the data. A 4th order zero-phase Butterworth is used in order to filter the signal. The objective of this step is to eliminate those frequencies that do not correspond with a common brain activity. Figure 3.1 shows the characterization of the filter.

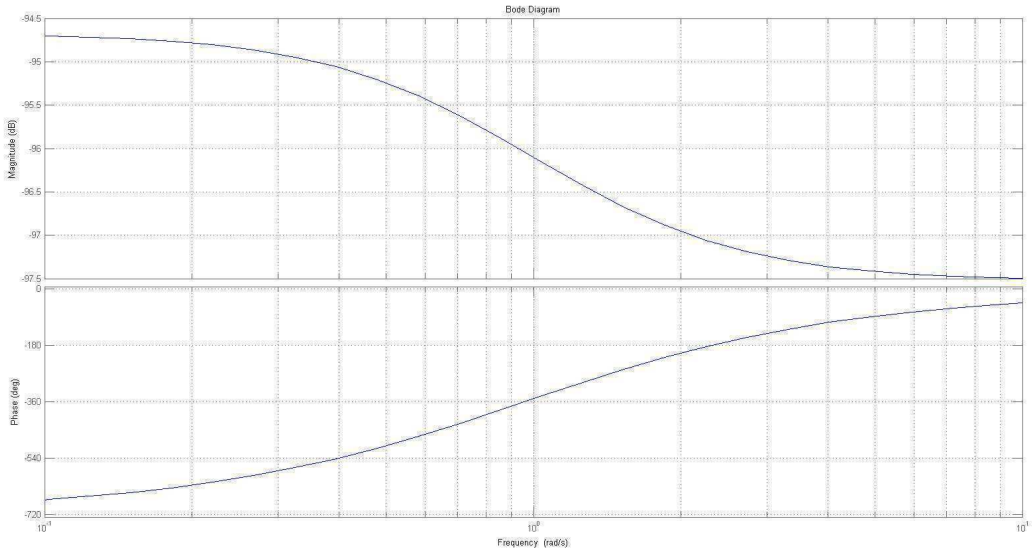



Figure 3.1: Bode diagram of the 4th order zero-phase V=Butterworth filter.

The next step is to detrend the data, with the objective beeing to remove any linear trend form it. After that, trials are tapered 5% on either side. The aim is to force all trials to start and end in zero voltage, so when trials are merged there is no sharp change.

Finally, for artifact removal three different conditions are taken into account:

1. Maximum accepted amplitude is 200uV

2. The amplitude of a trial should not be bigger (in absolute values) than the median of amplitudes of that subject + a standard deviation (in this case 4)

3. The number of zero crossings in a trial should not be bigger or smaller than the median of the zero crossings of that subjects +- a standard deviation (in this case 3)

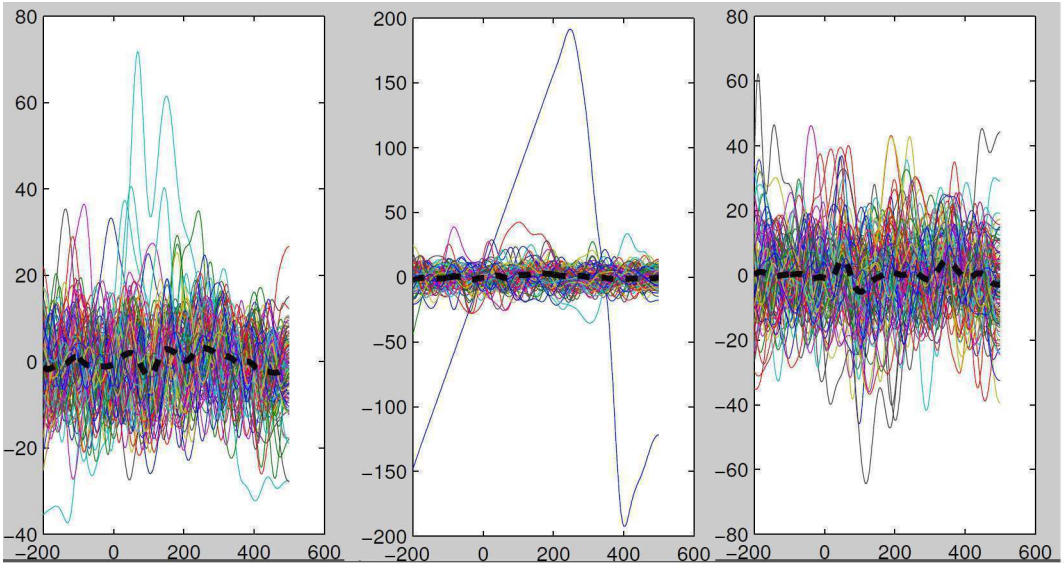



Figure 3.2: Epochs of different subjects with at least one example of a not desired trial. On the left an artifact trial. On the middle a slow trial. On the right noisy trials.

The Figure 3.2 shows examples of the three different case just discussed. The code for artifact removel can be find in the Appendix B: Artifact removal code.

### 3.4.2 Using iICA for component enhancing

After those two steps data are considered clean from artifacts and noise. However, it does not mean that all of the information of the signals is important for the study. As this signals were recorded as a response to an auditory stimuli, the study should focus on the main components of a typical response to an auditory stimulus. In this case the main component is N100 (a negative peak that occurs around 100ms after the stimulus) and for future steps, such as feature extraction, it is important that N100 is highlighted over the other components. When averaged EPs are used, there is no problem on N100 peak detection. However, on single trial analysis is really difficult to distinguish any component, including the N100. Therefore, based on previous studies

successes [14][38] and in order to enhance N100 component in each single trial, iICA algorithm was used. This method should be applied to one channel at a time, in the following steps:

1. Compute an average EP from all trials in the entire set.
2. In blocks of 10 ICA-tranform all single trials.
3. Within a predefined window Wr, compute the absolute correlation values between the current average EP and ICs in all blocks.
4. Zero thos ICs with correlation less than a predefined threshold r$base_{th}$.
5. Inverse-tranform the updated ICs back to the time domain, separately in each block.
6. Randomize the order of the updated single trials in the entire set.
7. Repeat steps 1-6 until the convergence criterion is met.

Figure 3.3 showed a more detail description of the iICA method.

## 3.5 Methods of brain characterization

### 3.5.1 Connectivity Network based on Granger Causality

Granger Causality (GC) is a method to obtain a connectivity network map. Wiener First and Granger later developed the idea that asserts "If the prediction of a time series X could be improved by knowing a second one Y, then Y is said to have a causal influence on X". This means, that by determining the influence of one signal on another, it is possible to infer which signal or brain area is driving another signal

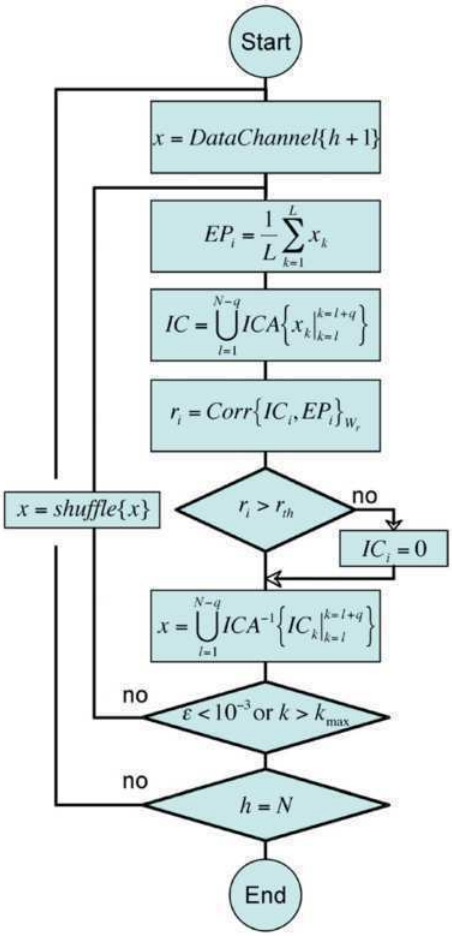



Figure 3.3: Flowchart of the iICA method [38].

or brain area. Using GC has several advantages over other connectivity network methods. The first and the most important one, is that GC is a non-invasive method whereas, other connectivity methods needed surgery. In this project GC was obtained just from the recorded EEG (also MEG can be used to obtain GC) of the subjects.

The other main advantage is that several studies show that GC is the most accurate method across all experimental conditions and that multisensory data sets can be accurately analyzed, as long as other methods only admit biosensor data sets. Moreover, GC method is not affected by network complexity, network size or SNR.

The GC is based on the relative change on the model error when series are added to improve prediction of the dependent signal. An overview to the mathematical model could help understanding the basis.

Suppose two signals represented by their autoregressive (AR) form.

$$X_1 = \sum_{j=1}^{j=\infty} A_j X_1(t-j) + e0 \tag{3.1}$$

$$X_2 = \sum_{j=1}^{j=\infty} B_j X_1(t-j) + e0 \tag{3.2}$$

As it can be infer from the equation the present value of the signal X can be determined by an amount of past values of the signal itself and an error e0.

$$X_1 = \sum_{j=1}^{j=\infty} A_j X_1(t-j) + \sum_{j=1}^{j=\infty} B_j X_1(t-j) + e1 \tag{3.3}$$

$$X_2 = \sum_{j=1}^{j=\infty} B_j X_1(t-j) + \sum_{j=1}^{j=\infty} A_j X_1(t-j) + e2 \tag{3.4}$$

These two equations show that the signal X can be also described by past values of X itself, some past values of the signal Y and an error e1. In the same way the signal Y can be described past values of Y itself, some past values of the signal X and an error e2. If the committed error predicting X using Eq. 2.3, e1, is smaller than the committed error predicting X using Eq 2.1, e0, it can be assert that X is influenced by Y, or in other words that X depends on Y. However, if the errors are the same, it can be assert that X is independent of Y. The dependency is not bilateral. X depending on Y, does not necessarily mean Y depending on Y.

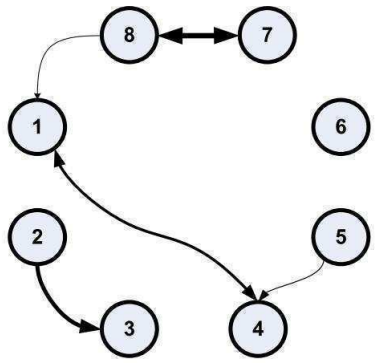



Figure 3.4: Eight nodes Granger Causality connectivity network [13].

For the moment GC has been mathematically explained only with two signals X and Y. Nevertheless, a human brain, hopefully, has more than two signals at the same time and lots of different dependencies occur. Fortunately, as it was commented some paragraphs before, GC is able to analyze accurately multisensory data sets. It has been proved that its accuracy can reach the 100 percent. This is the reason why GC was chosen to calculate the connectivity networks instead of using another method.

The simplest way to illustrate a Granger causality network is using a graph. The graph, as it can be seen in the next figure, contains nodes and arrows.

Each node of the graph represents one channel of recorded data. The arrows show the dependencies between the nodes. GC method has the next characteristics:

1. The thicker is the arrow, the bigger is the dependence. Sometimes, a set of colors is used.

2. Causal density defines the significant number of connections over all connections.

3. Causal flow defines the rate between the outgoing and ingoing dependencies of a single node.

4. Causal disequilibrium is related with the degree of reciprocity within a neural network.

According to the literature [34] [24] is crucial to have a good model order selection criteria. Nevertheless, there is not a perfect criteria that gives the optimum order for all AR models. Depending on different facts (AR model creation algorithm, time domain length, etc.) the available criteria perform better or worst. Based on that information an in an existing research [34], for this project the best criteria would be Hannan-Quinn criterion (HQC) due to the fact that our time domain data is categorized as large data (more than 120 points).

HQC was presented as an alternative to BIC and AIC criteria in 1979 and it has been seen (citation) that its performance is better in large sample data. The formula to calculate the HQC index is

$$HQ(p) = ln\left|\widetilde{\Sigma}(p)\right| + \frac{2ln(ln(\widehat{T}))}{\widehat{T}}pM^2 \tag{3.5}$$

where

$$ln\left|\widetilde{\Sigma}(p)\right| \tag{3.6}$$

is the logarithm of the determinant of the estimated noise covariance matrix (prediction error) for a VAR model of order p fit to the M-channel data, where

$$\widehat{T} = TN \tag{3.7}$$

is the total number of data points used to fit the model (T samples per trial x N trials).

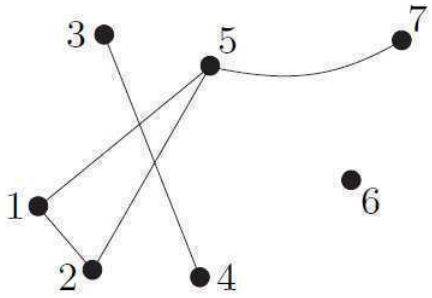



Figure 3.5: A representation of a graph with 7 nodes and 5 edges [6].

### 3.5.2 Graph Theory based Connectivity Network

In Computer Science and Mathematics the study of graphs is called Graph Theory. A graph is a collection of nodes/vertices and the edges that connect those nodes. A graph can be undirected, no difference between the two edges of two nodes, or directed, the direction of the edge of two nodes matters. The most usual way to represent a graph is using dots for nodes and lines for edges as can be seen in Figure 3.5.

The number of nodes/vertices is the order of that graph. Two different edges e and f are adjacent if they have a node in common and a set of edges is called independent if no two of its elements are adjacent. In our case, graph theory if used to extract features that will characterize the connectivity network form by the electrodes. Graph Theory will provide information such as the largest path, number of short connections and diameter of the graph.

## 3.6 Feature extraction

After signal preprocessing, features are extracted. In a brain signal features of different kind can be extracted, some of them need more processing and others less. In this

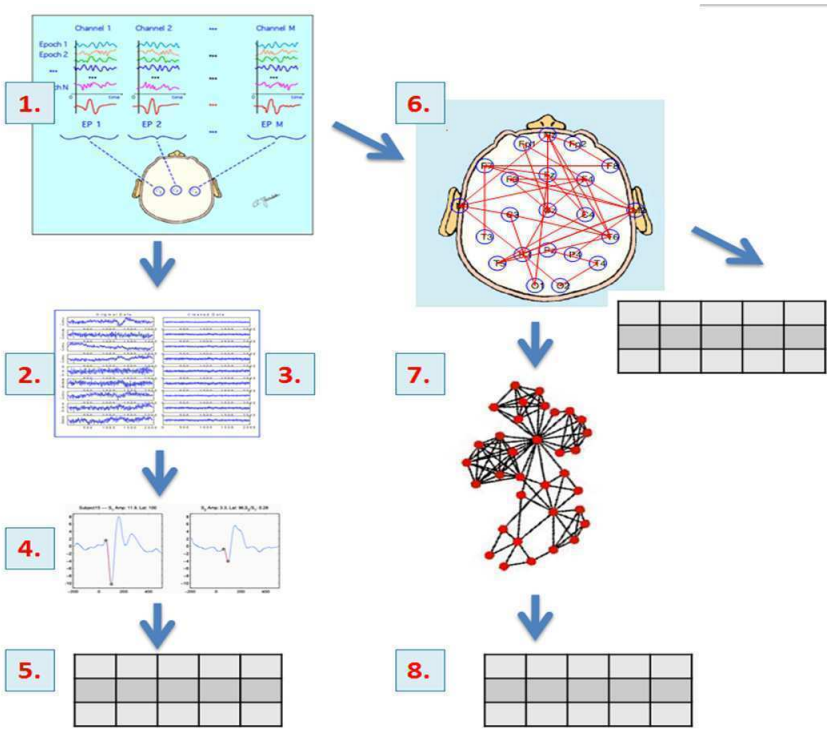



Figure 3.6: Signal preprocessing and feature extraction steps [40].

project first single trials are averaged and features of the resulting EP are extracted. On a completely different approach a Network Connectivity matrix is constructed based on Granger Causality theory. Also features are extracted from this Connectivity Network. On step more is done, and from this Connectivity Network a graph is constructed and features related with the graph theory are used for this project. In total each subject was described with 1014 features, 12 coming from the time domain, 968 coming from the Granger Causality and 34 from the graph theory. It is important to remark that all this features were extracted in an automated process. Figure 3.6 summarizes the general process from where features are extracted.

### 3.6.1 AEP time domain based features

As it was just mentioned, for almost all the features related with the time domain response to AEP the average of all trials is used. Just for the last two features each single trials is taken into account separately. In total for the time domain 12 different features are extracted.

1. **Peak to peak amplitude for S1:** Signal voltage difference between N100 and previous peak in S1.
2. **Peak to baseline amplitude for S1:** Signal voltage difference between N100 and the baseline in S1.
3. **Pre-stimuli activity variance for S1:** Signal variability in pre-stimulus time in S1.
4. **Latency for S1:** Time where averaged EP's N100 occurs in S1.
5. **Peak to peak amplitude for S2:** Signal voltage difference between N100 and previous peak in S2.
6. **Peak to baseline amplitude for S2:** Signal voltage difference between N100 and the baseline in S2.
7. **Pre-stimuli activity variance for S2:** Signal variability in pre-stimulus time in S2.
8. **Latency for S2:** Time where averaged EP's N100 occurs in S2.
9. **S2 S1 peak to peak amplitude ratio:** Peak to peak amplitude decreasing factor from S1 to S2

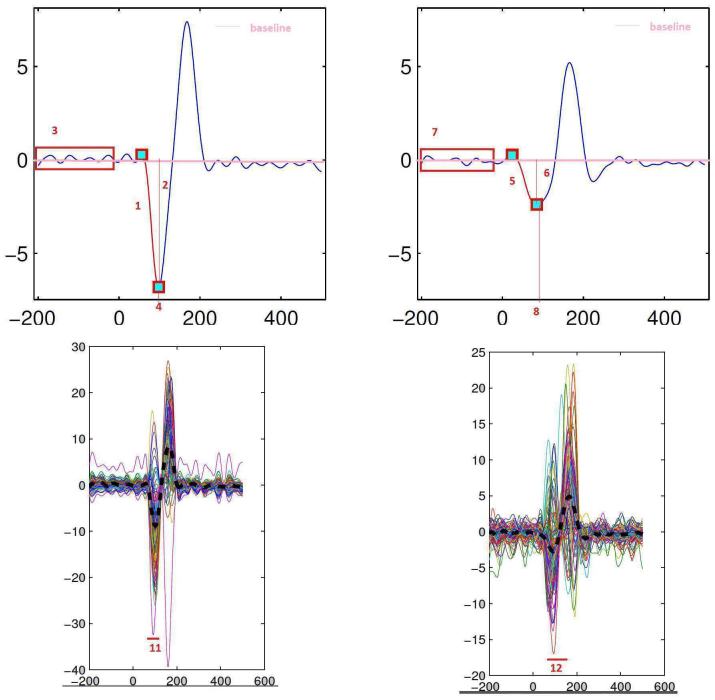



Figure 3.7: Time domain features represenetd over the original signal.

10. **S2 S1 peak to baseline amplitude ratio:** Peak to peak baseline decreasing factor from S1 to S2

11. **S1 latency variance over trials:** Variance over all trials of the time when N100 occurred in S1.

12. **S2 latency variance over trials:** Variance over all trials of the time when N100 occurred in S2.

Figure 3.7 shows all this features represented over the AEP in time domain.

The first ten features are related with the averaged signal. In order to obtain the first, second, forth, fifth, sixth and eighth features a window of 20ms width is set around 100ms. Then the algorithm searches for the highest peak (in absolute terms) and measures the amplitude respect previous peak and respect the baseline.

The baseline is considered the average of the signal in the range 100ms before the stimulus until 20ms after the stimulus. Once the main peak (N100) is detected it's location in time is saved (latency).

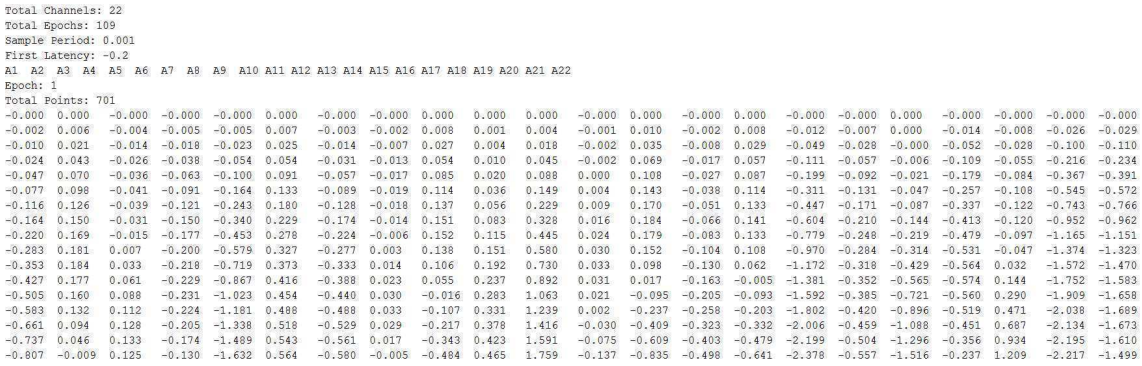
For the last two features processing of single trials is needed. For each subject the latency of the averaged response is used as a reference. Then a window of +/- 15ms is set around that reference. Then for each single trial the algorithm looks for the highest peak (in absolute terms) and measures the peak-to-peak and peak-to-baseline amplitudes and saves the latency. The variance of these latencies for each subject in S1 and S2 are used as the eleventh and twelfth features.

### 3.6.2 GC based features

For each subject two Connectivity Networks can be constructed based on the response to the first stimulus and to the second respectively. This Connectivity Network were created based on Granger Causality model. A mathematical way to represent these networks is using matrixes. Each elements of the matrix represents the strength of the confection between to electrodes. Each one of this element has been considered as descriptive feature of the subject. So in total, 22*22*2=968 features were used.

In order to obtain the GC matrixes from the signals many operations are required. Due to that, a cluster with many processors working in parallel is used. In this case the cluster is called Zeus and SSH Secure File Transfer Protocol has been used.This tool provides file access, file transfer and file management functionalities over any reliable data stream.

For this project a new directory was created in the cluster /home/abastarr. This directory is divided in two folders DATA and SCRIPTS. The first one contain all the



```
Total Channels: 22
Total Epochs: 109
Sample Period: 0.001
First Latency: -0.2
A1  A2  A3  A4  A5  A6  A7  A8  A9  A10 A11 A12 A13 A14 A15 A16 A17 A18 A19 A20 A21 A22
Epoch: 1
Total Points: 701
-0.000  0.000  -0.000  -0.000  -0.000  0.000  -0.000  -0.000  0.000  0.000  0.000  -0.000  0.000  -0.000  0.000  -0.000  -0.000  0.000  -0.000  -0.000  -0.000  -0.000
-0.002  0.006  -0.004  -0.005  -0.005  0.007  -0.003  -0.002  0.008  0.001  0.004  -0.001  0.010  -0.002  0.008  -0.012  -0.007  0.000  -0.014  -0.008  -0.026  -0.029
-0.010  0.021  -0.014  -0.018  -0.023  0.025  -0.014  -0.007  0.027  0.004  0.018  -0.002  0.035  -0.008  0.029  -0.049  -0.028  -0.000  -0.052  -0.028  -0.100  -0.110
-0.024  0.043  -0.026  -0.038  -0.054  0.054  -0.031  -0.013  0.054  0.010  0.045  -0.002  0.069  -0.017  0.057  -0.111  -0.057  -0.006  -0.109  -0.055  -0.216  -0.234
-0.047  0.070  -0.036  -0.063  -0.100  0.091  -0.057  -0.017  0.085  0.020  0.088  0.000  0.108  -0.027  0.087  -0.199  -0.092  -0.021  -0.179  -0.084  -0.367  -0.391
-0.077  0.098  -0.041  -0.091  -0.164  0.133  -0.089  -0.019  0.114  0.036  0.149  0.004  0.143  -0.038  0.114  -0.311  -0.131  -0.047  -0.257  -0.108  -0.545  -0.572
-0.116  0.126  -0.039  -0.121  -0.243  0.180  -0.128  -0.018  0.137  0.056  0.229  0.009  0.170  -0.051  0.133  -0.447  -0.171  -0.087  -0.337  -0.122  -0.743  -0.766
-0.164  0.150  -0.031  -0.150  -0.340  0.229  -0.174  -0.014  0.151  0.083  0.328  0.016  0.184  -0.066  0.141  -0.604  -0.210  -0.144  -0.413  -0.120  -0.952  -0.962
-0.220  0.169  -0.015  -0.177  -0.453  0.278  -0.224  -0.006  0.152  0.115  0.445  0.024  0.179  -0.083  0.133  -0.779  -0.248  -0.219  -0.479  -0.097  -1.165  -1.151
-0.283  0.181  0.007  -0.200  -0.579  0.327  -0.277  0.003  0.138  0.151  0.580  0.030  0.152  -0.104  0.108  -0.970  -0.284  -0.314  -0.531  -0.047  -1.374  -1.323
-0.353  0.184  0.033  -0.218  -0.719  0.373  -0.333  0.014  0.106  0.192  0.730  0.033  0.098  -0.130  0.062  -1.172  -0.318  -0.429  -0.564  0.032  -1.572  -1.470
-0.427  0.177  0.061  -0.229  -0.867  0.416  -0.388  0.023  0.055  0.237  0.892  0.031  0.017  -0.163  -0.005  -1.381  -0.352  -0.565  -0.574  0.144  -1.752  -1.583
-0.505  0.160  0.088  -0.231  -1.023  0.454  -0.440  0.030  -0.016  0.283  1.063  0.021  -0.095  -0.205  -0.093  -1.592  -0.385  -0.721  -0.560  0.290  -1.909  -1.658
-0.583  0.132  0.112  -0.224  -1.181  0.488  -0.488  0.033  -0.107  0.331  1.239  0.002  -0.237  -0.258  -0.203  -1.802  -0.420  -0.896  -0.519  0.471  -2.038  -1.689
-0.661  0.094  0.128  -0.205  -1.338  0.518  -0.529  0.029  -0.217  0.378  1.416  -0.030  -0.409  -0.323  -0.332  -2.006  -0.459  -1.088  -0.451  0.687  -2.134  -1.673
-0.737  0.046  0.133  -0.174  -1.489  0.543  -0.561  0.017  -0.343  0.423  1.591  -0.075  -0.609  -0.403  -0.479  -2.199  -0.504  -1.296  -0.356  0.934  -2.195  -1.610
-0.807  -0.009  0.125  -0.130  -1.632  0.564  -0.580  -0.005  -0.484  0.465  1.759  -0.137  -0.835  -0.498  -0.641  -2.378  -0.557  -1.516  -0.237  1.209  -2.217  -1.499
```

Figure 3.8: Beginning of a .jl file.

*.jl files with the preprocessed epochs of each subject, the second one contains the scripts that will compute all necessary operations to get GC.

*.jl files are a transformation of *.dat files that include some extra information as a header. Figure 3.8 shows the beginning of a .jl file. The header includes information such as total number of channels, total number of epochs, first latency, sample duration and name of electrodes. After the header the data is placed in form of 120-artifacts matrixes of 701x22. 701 are the points of a epoch and 22 the number of channels.

Before any process is started, the parameters of the algorithm should be chosen. These parameters can be easily modified through the file called parameters that is located in the folder SCRIPTS. This file contains parameters such as directories path, list of bad channels, order of GC, frequencies of filtering, etc. Figure 3.9 shows the file parameters.

The steps to obtain the GC matrixes are listed below.

1. Reserve a space in the cluster

2. Change to the directory of the SCRIPTS

```
Preprocessing directory where meg_prep.m and compiled executable is present
prep_dir=/home/abastarr/SCRIPTS/seth_parallel

Granger directory
granger_dir=/home/abastarr/SCRIPTS/seth_parallel

Data directory
data_dir=/home/abastarr/DATA/BOUTROSclean

Subject list filename
subject_list_file=/home/abastarr/DATA/BOUTROSclean/subject_list_AINHOA.txt

Lower and higher frequencies between which to filter
low_frequency=1
high_frequency=80

Downsampling factor
downsample_factor=1

Low and high lag
lag_low=4
lag_high=4

Range of channels taking part in this computation. Typically 1 through number of channel
low_chan=1
high_chan=22

----------------------
Number of threads used in computation, usually 2 is a good-enough compromise
NUMTHREADS=2

Wall-time required
WALLTIME=03:59:59

Physical memory, virtual memory. Don't tinker much here...
PMEM=-1
VMEM=2gb

Email to which notifications should be sent
EMAIL_HERE=abastarrika@uh.edu

----------------------
The following parameters are used to initialize the parallel computational grid

Number of columns in the computational grid (fixed at 1 for optimal performance)
npcol=1
blocking_factor=32
Number of rows is determined by the data size inside pipeline.sh

Number of items distributed to each process
items_per_proc=6000000

Number of processes is determined inside pipeline.sh
```

Figure 3.9: An example of configuration of the parameters for GC matrixes calculation.

3. Run the code to calculate GC matrixes

4. Clean unnecessary intermediate files

```
>>qsub -IV
>>cd SCRIPTS/seth parallel
>> ./run me EEG.sh
>> ./cleanup.sh
```

Then, *.txt files will be created containing the GC and FS matrixes in the specified directory. This files will be used in Matlab to extract the matrixes and obtain the desired features.

### 3.6.3 Graph theory based features

Once the Granger Causality matrixes are computed, a graph is created for each one. Then, based on the graph theory different characteristics of these graphs are chosen as features. In total, from each of the graph seventeen features are selected.

1. **lambda, characteristic path length of binary matrix:** global mean of the distance matrix

2. **ecc, mean eccentricity of binary matrix:** Eccentricity of a vertex is the maximum distance between that vertex and any other. The mean of all eccentricity of the graph is used as a feature in this project.

3. **rad, radius of graph of binary matrix:** Is the minimum eccentricity of the graph.

4. **d, diameter of graph of binary matrix:** Is the maximum eccentricity of the graph.

5. **id, mean indegree of all vertices:** Indegree of a vertex is the number or edges ending up in a node. The mean of that number is used as a feature.

6. **od, mean outdegree of all vertices:** Outdegree of a vertex is the number or edges going out of a node. The mean of that number is used as a feature.

7. **deg, mean degree of all vertices:** Degree of a vertex is the number of all incident edges loops counted twice.

8. **kden, connection density:** Number of connections present out of all possible.

9. **r, assortativity:** Indicates the degree of correlation.

10. **Eglob, Global efficiency:** It is equivalent to the mean of inverted distance matrix excluding main diagonal. matrix (excluding main diagonal) is equivalent to the global efficiency.

11. **Eloc, Local efficiency:** of a particular vertex is the inverse of the average shortest path connecting all neighbors of that vertex.

12. **C, clustering index:** Is a measure of degree to which nodes in a graph tend to cluster together.

13. **Q, maximized modularity:** The strength of division of a network into modules, also called groups, clusters or communities.

14. **lambda', characteristic path length of weighted matrix:** global mean of the distance matrix

15. **ecc', mean eccentricity of weighted matrix:** Eccentricity of a vertex is the maximum distance between that vertex and any other. The mean of all eccentricity of the graph is used as a feature in this project.

16. **rad', radius of graph of weighted matrix:** Is the minimum eccentricity of the graph.

17. **d', diameter of graph of weighted matrix:** Is the maximum eccentricity of the graph.'

### Feature reduction

For feature reduction two methods were selected among a large variety: Add one feature and subtract one feature.

**Add one feature:** At the beginning clustering is performed with one feature and the performance is calculated. Then a second feature is added and if the performance improves, the new feature is maintained if not is dropped. This is repeated until there are no more features to be added.

**Subtract one feature:** At the beginning the clustering is performed with all the features. Then a feature is subtracted and if the performance improves, that feature is maintained if not is dropped. This is repeated until there are no more features to be subtracted.

When data are unlabeled for performance evaluation, a label independent evaluation criteria must be chosen. In this project Xie S criterion function index was used

[39]. This index tries to identified well separated dense clusters and it is based on the intra-inter cluster distances ratio. The higher the index the better the performance.

$$S = \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} \mu_{ij}^{q} \left\| V_i - V_j \right\|^2}{N min_{ij} \left\| V_i - V_j \right\|^2} \tag{3.8}$$

The code for the S criterion can be found in the Appendix B : Xie S criterion.

For this feature reduction, the variable was the feature number. The number of clusters was fixed to three and the used clustering method was fcm. As fcm has not a unique belonging, an object was grouped in the cluster with higher belonging index.

On the other hand, when labels are available performance evaluation is much easier. Predicted labels are compared with real labels and accuracy is computed. The higher the accuracy the better the performance.

## 3.7 Clustering

Clustering objective is to separate a set of data into groups in such a way that objects in one cluster are more similar to each other than to objects from other clusters. Usually clustering is done when objects are not labeled and it is also called unsupervised learning. Although at the beginning most clustering methods were developed only for numerical data, now it is easy to find several methods that deal with both numerical and categorical data. Depending on the chosen method, the membership of an object to a cluster can be fuzzy (0,1) or total {0,1}. The way to define those clusters and memberships depends on the clustering method that is selected.

Sometimes labels are available and instead of using them just for external validation they can be used for supervision. These labels can be used to adjust the clustering

process (force a number of clusters instead of exploring different possibilities, forcing certain instances to belong to determined clusters, etc.). In this case it is called semi-supervised clustering. Instead, when labels are just used for validation purposes of the results, it is called unsupervised clustering. In this project, unsupervised clustering is used. The selected methods are EM, K-means and X-means.

### 3.7.1 Expectation maximization, EM

EM is a method that search for the maximum likelihood of parameters in statistical models. It is an iterative process to estimate some parameters O based on some measurements U. The tricky part is that some hidden variables Z are not given. the method consist of two steps:

Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of Z given X under the current estimate of the parameters theta:

Maximization step (M step): Find the parameter that maximizes this quantity:

EM algorithm has several applications, and one of them is data clustering. On the first step a model is fixed and the labels are estimated. On the Maximization step labels are fixed and a model that maximizes the expected value of the log-likelihood of the data is found. This iterations are repeated until some convergence criteria is fulfilled.

### 3.7.2 K-means

K-means is a clustering method that divides n objects into k clusters (fixed a priori) based on the proximity to each clusters mean. It is one of the simplest method of

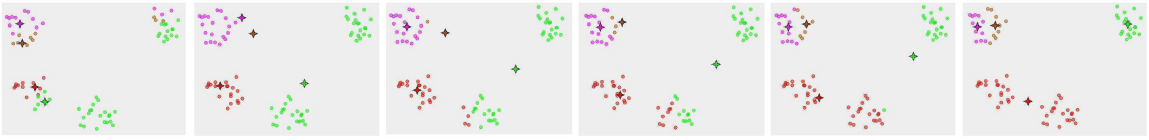


Figure 3.10: A typical example of the k-means convergence to a local minimum. In this example, the result of k-means clustering contradicts the obvious cluster structure of data set. The small circles are data points, the four ray stars are centroids (means). The illustration was prepared with the Java applet, E.M. Mirkes, K-means and K-medoids: applet. University of Leicester, 2011. This applet is published under CC Attribution 3.0 unported license [6].

unsupervised learning. There is also a Fuzzy C-means Clustering algorithm that is a soft version of K-means. In this case each observation does not simply belong or not belong to a cluster, but it has a degree of belonging.

The standard algorithm of K-means starts selecting k random objects that would be set as cluster centers. Then the rest of the objects are grouped in the nearest center cluster. Using different kind of distance measures (Euclidean, Mahalanobis, etc.) will lead in different groupings. In the next iteration the center is recalculated doing the average of all the objects in that cluster. These steps are repeated until convergence is reached. Figure 3.10 is a typical example of the standard algorithm of the K-means clustering method.

The advantage of this method is that it is simple and fast. The disadvantage is that the randomly selected first centers create a trend that could end up in an erroneous grouping of the objects. Choosing the correct distance measure, depending on the data dimensionality and nature could help avoiding this trend. However, sometimes it is not enough.

### 3.7.3 X-means

X-means appeared as a response to the K-means trending problem. Moreover, it offers solution to the problem of supplying a given number of K clusters. X-means will look among a range of possible number of cluster and select the optimum one.

The algorithm starts with K randomly selected centers, where K is the lower bound of the given range. Then it continues to add centers until the upper bound is reached. Finally the center set that achieves the best score is selected.

Each iteration is based on two phases, Improve Params and Improve Structure. Improve Params is intended to run the conventional K-means until convergence. The Improve Structure phase consist of deciding if new centers should appear. For this phase different approaches are possible, the most common one is to split each centroid into two children. The new children are moved a distance proportional to the size of the cluster in opposite directions along a randomly chosen vector. Then a local K-means(K=2) is run in each cluster. If using the children leads on no improvement, children are killed. However, if they apport improvement the father (original) center is killed and children are used as new centroid. Figure 3.11 shows an example of how X-mean algorithm works.

## 3.8 Classification

In statistics classification is called to the problem of identifying to which of the known sub-populations/groups belongs a new object. There is a training set of various observations whose memberships is known. These observations are described with several features that could be qualitative or quantitative. Based on those labels and features a model is created. This model will classify any new instance/object in the

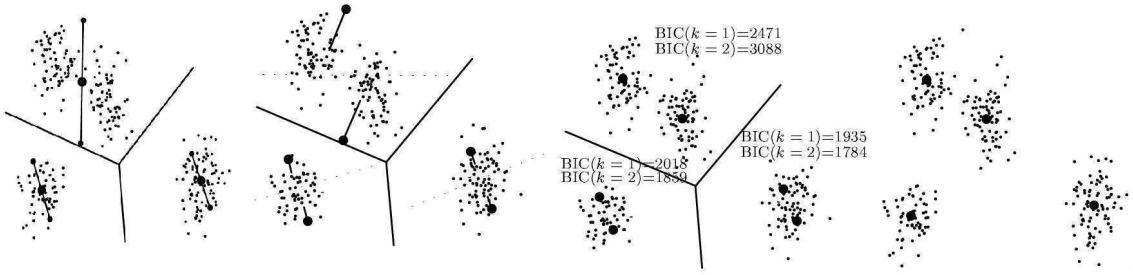



Figure 3.11: An example of X-means algorithm. From the left to the right. First step shows each originla centroid splitted into two children. Next step shows first step of parallel local 2-meeans. Lines show where centroids will move to. Next step shows result of all local parallel 2-means. Final step shows the selected centroids after best scoring selected.

corresponding group. In this project all classification problems were solved in Matlab code. For this purpose, two modules were developed: one for feature combination and the other for the classification itself. The first one creates a big matrix of dimensions subject x number of features, and the second one needs attributes of type subject x number of features. This approach makes the problem more flexible, due to the fact that to add or remove features only changes in the first module are needed. The second module won't need any change.

As it was just explained a classification problem needs first to create a model and then to use that model to classify the rest of the data. For that purpose data should be divided into two parts: the training set (for model creating) and the testing set (for classification). There are several ways to divide the data into training and testing sets. In this project the leave-one-out approach is used. This means that one subject is separated for the testing set and the other thirty-five are used for training set to create the model. This model is used to predict the first subject. Then the real and predicted labels are compared and the confusion matrix is updated. Then the second

subject is used as a testing set and the other thirty-five as a training set, and the confusion matrix is updated.This continues until all subjects have been classified.

Several methods can be found in the literature for developing these models. On this project we have just focused on two, Support Vector Machine (SVM) and a decision tree.

### 3.8.1 Support vector machine, SVM

SVM is a statistical concept of a supervised learning method to analyze and recognzie pattern, used for both regression and classification models. Often data sets are not linearly separable, so it was proposed to map the original finite-dimensional space into a much higher-dimensional space, in order to simplify the separation in that space.

Different hyperplanes can be created to separate data sets. The trick is to choose the hyperplane that has the biggest margin. Margin are the parallel imaginary hyperplanes that contain the closest objects to the constructed hyperplane. Those objects that are in the margins are called Support Vectors. The hyperplane with the highest margin distance is the chosen one for creating the model.

### 3.8.2 Decision tree

A decision tree is a tool that uses a tree-like graph of decisions and consequences. It can be used for prediction (costs in economics problems), for data regression, data classification, etc. There are different softwares to create these kind of trees. In our case the Matlab function classregtree was used. Given a matrix of predictor (features) and a vector of objects (labels), the output of this function is a decision tree. The advantages of the decision tree method is that, once the tree is created it is visually

easy to identify the important features and their boundary values.

This method contains three different phases. First of all, the selection of the splits. Then it must be decided if the node is terminal, or if it will continue splitting. And finally, a class should be assigned to a terminal node.

## 3.9 Procedure

In previous section only a few methods of all available in the literature have been described. As it has been explained, not all methods can be used in all scenarios, and the same method can be used in different ways in different occasions. For that reason, the aim of this project is not only to present the available methods but also to create a procedure of how to use them in order to obtain the best possible results.

For obvious reasons, this section is divided in two parts: A procedure for unlabeled data and a procedure for Labeled Data. However, a set of general steps can be defined for both procedures. The differences will be on how to develop those steps.

1. Preprocessing and feature extraction
2. Feature reduction
3. Object grouping

### 3.9.1 Preliminary exploratory analysis with unlabeled data

At the beginning of the project no label was available, the only information that was known was that the three groups have same number of subjects 12-12-12. This fact was used to have an idea if the performance of the procedure was going on the right direction or not.

### Preprocessing and feature extraction

The first approach was to try to separate the data using the Graph Theory and Time domain features separately. For that reason the preprocessing of each kind of features was different.

As graph theory features are calculated from the graph created based on the GC matrixes, the preprocessing for GC and GT features is the same. In this case only a bandpass filter with cutoff 1-80HZ is applied. This filtering is done in the cluster just before the algorithm of AR model is called.

For the time domain features, first artifact removal is done and then a low pass filter of cutoff frequency 20Hz is applied. Finally in order to enhance the N100 component of each trial iICA algorithm is applied.

The second approach was to combine all features to see if combination gave any improvement.

### Feature reduction

For unlabeled data both methods proposed before were used: Add-one-feature and Subtract-one-feature and for performance evaluation Xie S criterion was used.

### Object grouping

Graph Theory features

The problem can be approached directly or in two steps. Both possibilities were explored.

1. Directly. When the problem was faced directly, the clustering method that was used was fcm. It was tried using four different distance measures (Euclidean,

Minkowski, Chebychev and Correlation) and for exponent for the partition matrix default value (2) and value 1.2 were used. In order to reduce the number of features both techniques (adding and subtracting one feature) were explored.

2. Two steps.The first step is to separate all the subjects into two groups, fortunately, it will end with healthy and unhealthy subjects. Then a group from these two groups one is selected for the second stage. At this point the knowledge that groups should be 12-12-12 was used to select the group, the group with higher number of subjects was used for the second stage. So this part of the project is a semi-supervised clustering.

   This clustering was performed using different clustering methods (EM, K-means and Density based clustering) and two different distance measures, Euclidean and Manhattan. Clustering was done using all the graph theory features, only features from S1 and only features from S2.

Time domain features

Time domain features can be treated in two ways:

1. Single trials of a subject can be averaged and features from the averaged wave can be used for solving the problem.

2. All single trials can be treated as individual waves, obtain features from single trials and group all the single trials. Then the subject will belong to the group that the majority of each trials belong to.

In our case mainly features from the average wave were obtained. However, to those features there were added also features related with single trials like variance of

latency among all trials, variance of amplitude among all trials, etc. But no analysis of individual trials was performed.

For time domain features a new idea came up. As the number of features was no too high, time domain data was clustered with and without feature reduction. After seeing the results obtained using graph theory features, it was decided directly to use the two stage approach.

Combined

The 34 features coming from the graph theory were combined with the time domain features. For that purpose, time domain features must come from the averaged wave due to the fact that each subject contains a different number of trials and they could not be comparable. As it was done in other phases, in this case also different clustering methods were tried again (EM, K-means, X-means).

### 3.9.2 Labeled Data

Once the labels are known the problem can be refocused. In the case of this project labels are not used for a completely supervised learning approach, the idea is to use the labels just for a good feature selection and then using those features see if data can be naturally grouped with an unsupervised clustering approach. Once it is decided that labels will be use only for feature selection three major steps can be defined in this procedure.

1. Preprocessing and feature extraction
2. Feature reduction using labels
3. Data grouping using clustering methods

### Feature reduction using labels

In order to reduce the number of features first is important to know which feature is more important than the others. For that reason, before reducing the number of features, a ranking of the features is done. The next step is to decide how many features are necessary for a good description and classification of the subjects. For that, classification methods are used. The idea is to use a feature adding approach. First just one feature was used and all the subjects were classified. Using the labels the accuracy of that classification was computed. Then two features were used and the new accuracy was computed. After that the "cheapest" solution was selected. This means, that the highest accuracy produced with the minimum number of features was selected.

At the beginning the classification problem was faced in just one step. Subjects were directly classified in one of the three groups. Later on, it was faced in two steps. First SZ patients were separated from the rest of the subjects, and then CA and NC were separated.

One stage classification

The first approach was to try to classify all subjects just in one stage. The procedure for that has several steps that are explained in this subsection.

The first one is to load the data matrix and to normalize it. The reason for normalization is that several SVM guide books [28] show better accuracy when data is normalized. Next step is to rank all the features based on their contribution to a better classification. For that purpose the whole data set is classified using just one feature at a time. Based on the obtained accuracy features are ranked.

Once the list is created, a feature reduction must be done. The selected approach

is the add-feature, the first iteration begins with just one feature and in each iteration the next feature of the list is added and so on. At the end the number of features is selected based on the get best accuracy. The next paragraphs explain the whole procedure in more detail.

1. Select the first feature from list1
2. Create models and make predictions unseing Leave-One-Out approach
3. Compare predictions with labels and get accuracy
4. Add the next feature from the list1
5. Repeat from step 2 until features are finished
6. Select the number of features that gave the maximum accuracy

Two stage classification

The first step was to make a ranked list of all the features, from most important to less important. As the classification is made in two steps, two different list are made for the first stage and the second. To create the list t-test is used. T-test is a common technique used to evaluate if two groups are different. The value that is obtained when t-test is realized (p-value) is the probability of error we will have if the hypothesis of difference is accepted. Therefore, in order to create the first list the t-test is applied to each feature individually and assuming group one SZ and group two CA and NC.The feature with the lowest p-value is the most important feature for the first stage. For the second list, t-test is again applied to each feature and now the two groups are CA and NC.

Once both lists are created, a feature reduction must be done. The selected approach is the add-feature one, the first iteration begins with just one feature and in each iteration the next feature of the list is added and so on. At the end the number of features is selected based on the best accuracy.The next paragraphs explain the whole procedure in more detail.

Separate SZ from the rest of the subjects,

1. Select the first feature from list1
2. Create models and make predictions unseing Leave-One-Out approach
3. Compare predictions with labels and get accuracy
4. Add the next feature from the list1
5. Repeat from step 2 until features are finished
6. Select the number of features that gave the maximum accuracy

Separate those subjects that were classified as non schizophrenic.

1. Select the first feature from list2
2. Create models and make predictions unseing Leave-One-Out approach
3. Create confusion matrix using predictions and previous stage labels and get accuracy
4. Add the next feature from the list2
5. Repeat from step 2 until features are finished

6. Select the number of features that gave the maximum accuracy

The models of this procedure were created using both SVM and decision TREE.

### Data grouping using clustering methods

The aim of this phase of the project is to ensure that the accuracies obtained in previous step (classification) were due to a good selection of features and not because of the classification models. For that reason the selected features were used in a clustering problem to see if those feature create naturally separable groups. The selected clustering approach was the two stages, first SZ were separated from the others, and in the second stage CA and NC were separated.

In this case, instead of using Matlab code clustering problems were performed using Weka. Weka is a collection of machine learning algorithms for data mining. It contains tools for preprocessing, clustering, classification, regression problems. Weka has a friendly user interface and it provides several clustering methods that can be customized by the user.

These steps explain the overall procedure.

1. Subjects are described only with the most important features from list1
2. All the subjects are used in order to find 2 groups
3. Select one of the two groups for the second stage
4. Subjects are described only with the most important features from list2
5. Clustering is applied again to remaining subjects
6. Confusion matrix is created

For step 3 a selection criteria must be decided. Based on other researches [25] an easy way to see if a data set contains clusters or not is to analyze the shape of the normalize distances histogram. A data set without clusters will show a bell shape distribution, whereas a data with clusters will show a different distribution that will have small (intra-cluster) and high (inter-cluster) distances.

As different classification approaches leaded in different number of selected features, 3 different clustering problems were done in order to find the best accuracy with the least number of features.

1. Features selected from SVM for both first and second stage
2. Features selected from decision TREE for both first and second stage
3. For the first stage feature from the decision TREE was used, and for the second stage features from the SVM were used.

### 3.9.3 Parameter exploration

Firstly, the procedure above was checked with certain values of preprocessing. However, would the procedure work if a different filter is used? Would the accuracy be the same if iICA is not applied? In other words, is our procedure a robust procedure? Is there any way to simplify the procedure but maintain the same accuracy?

In order to find an answer to these questions a parameter exploration was performed. All possible combination that can be done using the next values were used,

1. Lowpass filters with cutoff frequency of 10 and 20 HZ
2. Applying/Not applying iICA

3. GC matrixes using AR models of order 2, 4 and 9

It is important to note that for the parameter exploration not all features were used. As it will be explained in the results chapter, the first parameters configuration used in this procedure showed that parameters coming from amplitude and latency were low in the ranking list. So with the purpose of time saving in parameter exploration, just features coming from Granger Causality matrixes and Graph Theory were used.

But, why only use AR model order 2, 4 and 9? Are this numbers special? These three values are randomly selected exploratory values to measure the effect of the AR model on the whole procedure performance. However, apart of the model order effect, it is interesting to know which order will be optimal. Nowadays, literature provides several criteria to select the order of AR models. Nonetheless, does it mean that the optimum order of the AR model will end up in a better performance of the whole procedure? Not necessarily, what is meant by best is controversial and it depends on the selected criteria. Hence, the optimum model order may not have relevant information regarded to this specific problem. The only way to know it is to check the performance of the method using also the optimum order.

### 3.9.4 Process automatization

The procedure explained above had good results. However, it involves different softwares: Matlab and Weka. This makes the procedure more tedious than it seems. In order to work with both of them, first data from Matlab must be saved into a file. Then this file is used in Weka, and it generates another file as an output. However, files generated by Weka are ARFF format, which are unreadable for Matlab. So

before going back to Matlab they should be converted to a Matlab readable format (CSV, for example). This involves a lot of manual work that can lead to several errors (converting the uncorrcet file, mixing files, etc).

In order to make the procedure more simple and automatic for the user, a script was created that collects results from Matlab, converts them into ARFF, obtains the results from Weka, converts them to Matlab readable, etc automatically. This way, all the procedure can be done just running one Matlab script. To create this script a Matlab module was downloaded from mathworks community and it was modified to our convenience.

# Chapter 4

# Results

## 4.1 Preprocessing: filtering and iICA

Figure 4.1 illustrates the results for artifact removing for one subject. The Figure shows three stages: all the singles trials of a subject before the preprocessing (left), all trials that were considered good (middle) and those trials that were considered artifacts. The black line that is common in the three graphs represents the average of all the trials of that graph.

In general for the 36 subjects, few trials were classified as artifacts. Nevertheless, in order to simplify further comparison between trials, when one trial of one stimulus

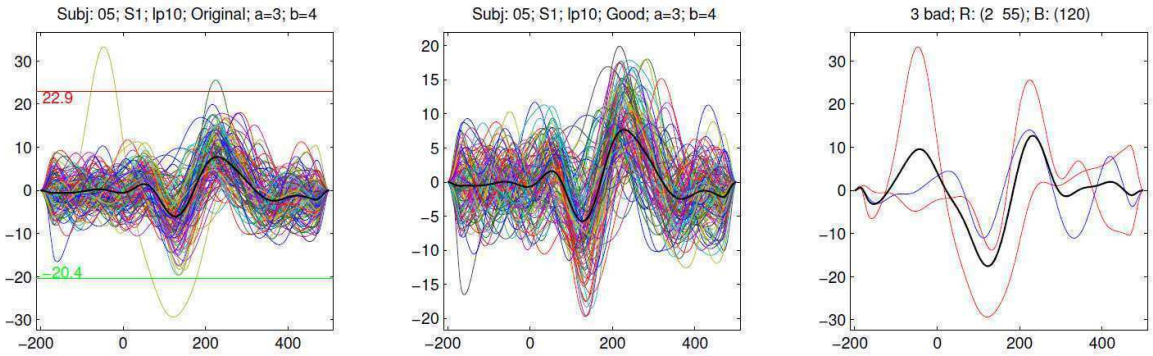



Figure 4.1: Flowchart of the iICA method.

response of one channel was considered artifact, that trial was removed from all channels, and both stimuli of that subject. Table 4.1 shows the number of good trials (out of 120 performed in the experiment) and the number of the bad trials.

Once the artifacts were removed, the iICA algorithm enhanced the N100 component. Figure 4.2 illustrates the good performance of iICA algorithm. On the left, all the trials of a subject before iICA are shown and on the right graph the resulted signals after the processing. The black line represents again the average of all the trials of that graph. It can be observed that all components out of the complex N100-P200 have reduced their amplitude drastically, allowing for a better visualization of the N100 component. As a consequence, on the left graph N100 component can be hardly seen, while on the right it is easily detectable. These cleaned trials will allow not only a better study of the averaged AEP response, but also will allow to study each single trial separately.

## 4.2 Feature extraction

After all features were extracted, they were combined to create a big file that contained all the subjects described by the 1014 features. The script that combines all the features is flexible. This means that if any new feature is found and we want to include it on the file can be easily added.

For time domain features, amplitude and latencies were extracted from both single trials and averaged AEP. As it was explained in previous chapter, these extraction were done automatically using Matlab. Figure 4.3 is an example of the automatically detected N100 component for S1 and S2 of the same subject. Figure 4.4 shows the N100 component detected in a single trial of a subject for both S1 and S2. It is

| Subject | Num. Good trials | Bad Trials |
|---|---|---|
| 1 | 115 | 16, 26, 106, 116, 117 |
| 2 | 111 | 9, 18, 39, 43, 45, 53, 65, 104, 120 |
| 3 | 113 | 4, 10, 23, 28, 41, 106, 120 |
| 4 | 109 | 92, 103, 105, 107, 109, 110, 111, 116, 117, 119, 120 |
| 5 | 117 | 2, 55, 120 |
| 6 | 113 | 1, 2, 38, 48, 84, 89, 120 |
| 7 | 116 | 14, 15, 58, 59 |
| 8 | 113 | 2, 19, 20, 39, 48, 56, 120 |
| 9 | 115 | 25, 34, 56, 66, 116 |
| 10 | 115 | 4, 46, 57, 109, 120 |
| 11 | 115 | 2, 25, 31, 78, 104 |
| 12 | 113 | 1, 20, 30, 53, 109, 111, 120 |
| 13 | 118 | 98, 99 |
| 14 | 115 | 35, 36, 42, 48, 92 |
| 15 | 113 | 55, 56, 70, 72, 76, 93, 112 |
| 16 | 114 | 29, 30, 38, 46, 53, 120 |
| 17 | 114 | 71, 76, 82, 110, 111, 120 |
| 18 | 114 | 68, 75, 104, 115, 118, 120 |
| 19 | 106 | 3, 9, 14, 19, 25, 35, 37, 39, 44, 49, 51, 57, 59, 120 |
| 20 | 112 | 10, 14, 37, 80, 87, 103, 104, 105 |
| 21 | 111 | 8, 30, 34, 37, 39, 61, 67, 103, 120 |
| 22 | 113 | 49, 66, 92, 98, 104, 108, 113 |
| 23 | 115 | 36, 45, 53, 60, 120 |
| 24 | 112 | 43, 60, 69, 91, 92, 99, 108, 120 |
| 25 | 114 | 1, 25, 34, 52, 87, 91 |
| 26 | 110 | 7, 34, 48, 52, 59, 63, 69, 70, 76, 116 |
| 27 | 112 | 4, 17, 19, 21, 26, 27, 29, 120 |
| 28 | 113 | 7, 9, 28, 49, 52, 54, 120 |
| 29 | 115 | 32, 92, 106, 107, 120 |
| 30 | 112 | 6, 23, 32, 40, 45, 84, 85, 91 |
| 31 | 112 | 45, 57, 68, 82, 91, 113, 114, 120 |
| 32 | 112 | 1, 8, 25, 28, 31, 42, 57, 97 |
| 33 | 115 | 1, 30, 47, 93, 120 |
| 34 | 113 | 9, 24, 49, 54, 55, 59, 114 |
| 35 | 116 | 2, 9, 49, 120 |
| 36 | 115 | 22, 23, 43, 58, 60 |

Table 4.1: Number of good trials and the list of bad trials for each subject.

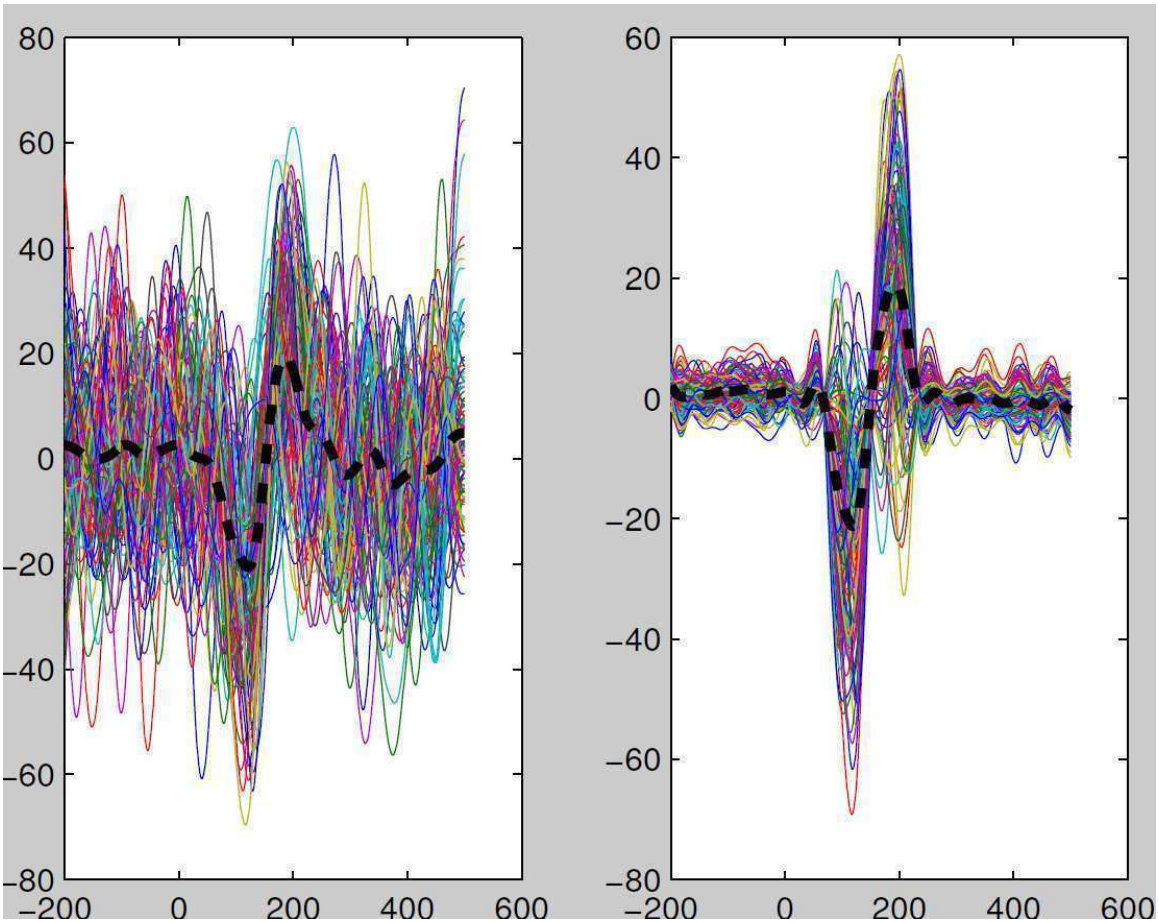



Figure 4.2: Flowchart of the iICA method.

important to remark that although the performance of the automatical detection is good, it did not work perfectly in all single trials. Few of them were not correctly detected (Figure 4.5 ).

## 4.3 Preliminary Exploratory Analysis with Unlabeled data

### 4.3.1 Graph Theory features

The results obtained from directly clustering the data into 3 groups using fcm is resumed in the tables Table 4.2 and Table 4.3.

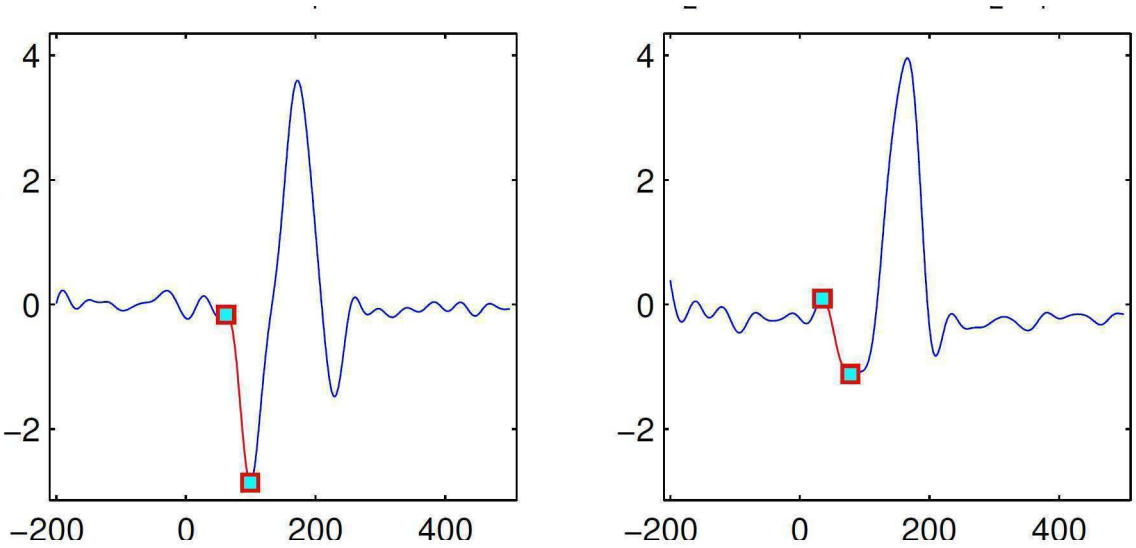



Figure 4.3: N100 detection in averaged AEP response.

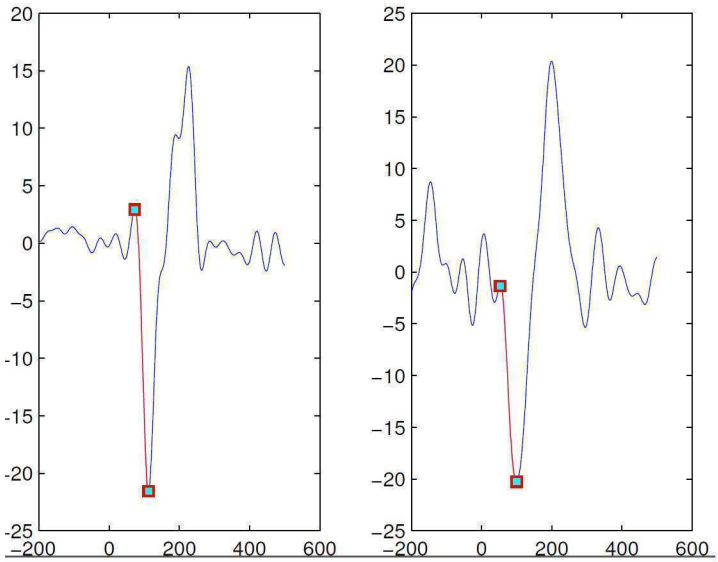



Figure 4.4: N100 detection in a single trial of AEP response.

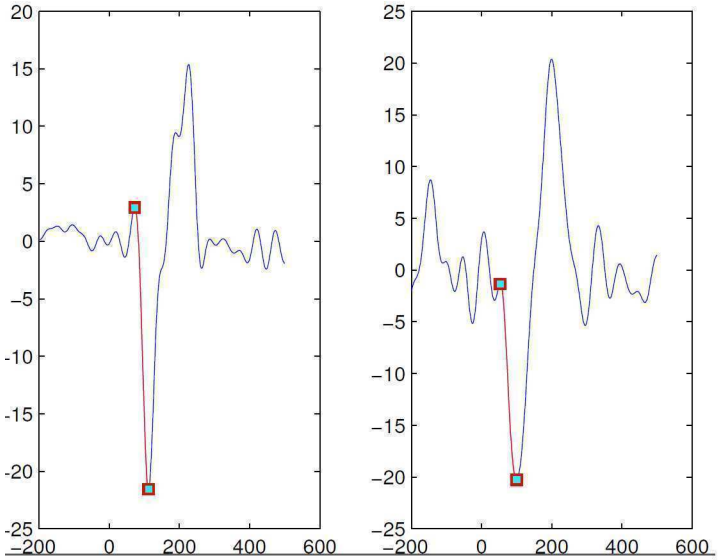



Figure 4.5: A bad N100 detection in a single trial AEP response.

| Distance method | Adding feature |
|---|---|
| Euclidean | S=0.2846 F=[1 2 12] Cl=(17,17,1) |
| Minkowski | S=0,2846 F=[1 2 12] Cl =(17,17,1) |
| Correlation | S=Nan F=[1] Cl=(21,6,8) |
| Chebychev | S=0,2831 F=[1 2 12] Cl=(17,17,1) |
| Distance method | Subtracting feature |
| Euclidean | S=0,4173 F=[34 32 28 27 25 24 23 22 18 10] Cl= (16,11,8) |
| Minkowski | S=0,4173 F=[34 32 28 27 25 24 23 22 18] Cl= (16,11,8) |
| Correlation | S=0,5312 F[34 33 32 23 22 21 19 9 8 6 5] Cl=(17,10,8) |
| Chebychev | S=0,4241 F=[34 27 26 25 23 22 18 14] Cl=(17,10,8) |

Table 4.2: S index values, selected features (F) and cluster belongings(C), for clusterings realized with fcm and using both Adding feature and Subtracting feature, feature reduction technique. q=2

| Distance method | Adding feature |
|---|---|
| Euclidean | S=0,2707 F=[4 7 1 2 6 7 10] Cl=(15,17,4) |
| Minkowski | S=0,2126 F=[9 2 16] Cl =(26,8,1) |
| Correlation | S= Nan F=[17] Cl=(14,14,7) |
| Chebychev | S=0,3400 F=[15 9 15 26] Cl=(13,18,4) |
| Distance method | Subtracting feature |
| Euclidean | S=0,3627 F=[34 32 29 17 11] Cl= (14,14,7) |
| Minkowski | S=0,3627 F=[34 32 29 17 10] Cl= (14,14,7) |
| Correlation | S=0,2694 F=[33 24 22 17 16 15 14 11 9 7 6 5 3 2] Cl=(23,11,1) |
| Chebychev | S=0,3268 F=[34 30 28 27 26 21 20 19 18 14 13 9 8 2 1] Cl=(17,10,8) |

Table 4.3: S index values, selected features (F) and cluster belongings(C), for clusterings realized with fcm and using both Adding feature and Subtracting feature, feature reduction technique. q=1.2

| Used features | Distance | Kmeans | EM |
|---|---|---|---|
| All | Euclidean | 12/24 | 12/24 |
| All | Manhattan | 12/24 | Does not depend on distance |

Table 4.4: Number of objects belongin to each cluster for stage 1 of the clustering problem for different distance measures and clustering methods.

As can be inferred from the tables, the results are not good. The clustering results should be close to 12-12-12 and obviously these results are not. For this reason, these results were not even validated with the labels, because the performance is not good.

On the other hand, the results obtained form clustering the data in two steps is summarized in the tables Table 4.4 and 4.5.

In this case some of the groupings are close to 12-12-12, so the best approach was validated once the labels were provided. So summarizing, the best approach using just graph theory features would be

1. Use k means and 34 features to obtain two groups of 12/24 2. Use k means and the 17 features from S2 to obtain create 13/11

and the performance is shown in the next confusion matrix.

| Used features | Distance | Kmeans | EM |
|---|---|---|---|
| All | Euclidean | 10/14 | 0/24 |
| All | Manhattan | 10/14 | |
| S1 | Euclidean | 9/15 | 9/15 |
| S1 | Manhattan | 10/14 | |
| S2 | Euclidean | 13/11 | 8/16 |
| S2 | Manhattan | 13/11 | |

Table 4.5: Number of objects belongin to each cluster for stage 2 of the clustering problem for different distance measures and clustering methods.

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 6 | 3 | 3 | SZ |
| 2 | 6 | 4 | NC |
| 2 | 5 | 5 | CA |

Table 4.6: Confusion Matrix of the classification results using 2 stage EM.

### 4.3.2 Time domain features

The results obtained from clustering data using time domain features with feature reduction is shown in the Table 4.7.

When data was clustered without any feature reduction technique, the clustering problem en up in a 12-12-12 grouping. So at first appearance it seemed that the clustering was OK. However, having three groups of 12 subjects each one, does not mean that the subjects inside each group are the correct ones. Once labels were provided this result was validated using the labels, and the confusion matrix obtained from this validation is:

In this case, only the results obtained without feature reduction were validated using provided latex. As in graph theory features case, for data with feature reduction it can be seen that the results are not good. So it was not worth it to validate the results.

| Distance method | Adding feature |
|---|---|
| Euclidean | S=0,2755 F=[3 2 3 6] Cl=(14,18,4) |
| Minkowski | S=0,2126 F=[9 2 16] Cl =(26,8,1) |
| Correlation | S=Nan F=[6] Cl=(12,15,9) |
| Chebychev | S=0,3118 F=[2] Cl=(17,10,9) |
| Distance method | Subtracting feature |
| Euclidean | S=0,277 F=[6 5 3] Cl= (14,18,4) |
| Minkowski | S=0,277 F=[6 5 3] Cl= (14,18,4) |
| Correlation | S= 334.9328 F=[5 3 2 1] Cl= (14,18,4) |
| Chebychev | S=0,3167 F=[6 3 2] Cl= (14,18,4) |

Table 4.7: S index values, selected features (F) and cluster belongings(C), for clusterings realized with fcm and using both Adding feature and Subtracting feature, feature reduction technique. q=1.2

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 7 | 4 | 1 | SZ |
| 2 | 6 | 4 | NC |
| 3 | 2 | 7 | CA |

Table 4.8: Confusion Matrix of the classification results using SVM.

### 4.3.3 Combined

Once both type of features were combined and clustering was done, the results were validated. However, the performance of the clustering using both types of features was worst than using only time domain features. In this case in the first stage data were separated into two group 12-24. However, in the second stage Em method did not identify two groups.When EM was forced to find two clusters it separated the data into 9 and 15. X-means separated it into two groups of 18-6. It was assumed that the best approach was EM for both first and second stage (forcing it to find 2 clusters) and it was validated using the labels.

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 6 | 3 | 3 | SZ |
| 4 | 4 | 3 | NC |
| 5 | 5 | 3 | CA |

Table 4.9: Confusion Matrix of the clustering realized with combined features.

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 7 | 2 | 3 | SZ |
| 1 | 6 | 5 | NC |
| 2 | 2 | 8 | CA |

Table 4.10: Confusion Matrix of the classification results using one stage SVM.

## 4.4 Labeled Data

### 4.4.1 Preprocessing and feature extraction

The results presented in this section are the results obtained with the data preprocessed with the next values:

1. Data is filtered with a lowpass 20Hz filter, detrended and tappered
2. Artifacts are removed
3. iICA is applied to single trials

### 4.4.2 Feature reduction using labels

One stage classification

One stage classification showed an accuracy of 0.58, the confusion matrix is shown in Table 4.10.. As it is below 80% it was decided to try the two stage approach before continuing with this one. The results were that the two stage classification resulted in a higher accuracy so the One stage approach was discarded.

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 12 | 0 | 0 | SZ |
| 0 | 12 | 0 | NC |
| 1 | 0 | 11 | CA |

Table 4.11: Confusion Matrix of the classification results using two stage SVM.

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 7 | 2 | 3 | SZ |
| 0 | 10 | 2 | NC |
| 0 | 2 | 10 | CA |

Table 4.12: Confusion Matrix of the classification results using two stage TREE.

Two stage classification Classification using SVM method resulted in a 0.97 accuracy. In order to achieve that accuracy 37 features were needed on the first step and 31 on the second one. Although reducing the features from 1014 to 68 is a good result, a bigger reduction will be welcome. The confusion matrix of this results is shown in Table A.1.

Despite the fact that a 0.97 accuracy is a good accuracy, is important to remark that the subject that was misclassified was confused between SZ and CA. This means, that no normal control was classified as unhealthy and no patient was classified as healthy.

Classification using the decision TREE method resulted in a 0.75 accuracy. In order to achieve that accuracy just one feature was needed on the first step and 2 on the second one. In this case the feature reduction was extremely good. However, an accuracy higher than 0.75 is not possible to achieve.The confusion matrix of this results is shown in Table 4.12.

An important fact of using TREE algorithm is that only one feature is needed for the first stage and 2 for the second one. Although the accuracy is not as high

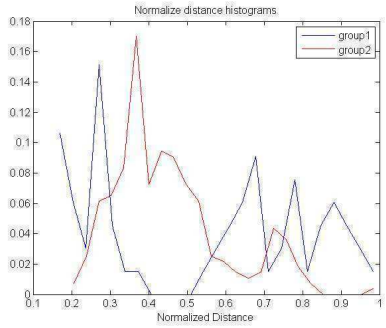



Figure 4.6: Histograms of the normalized distances among subjects of the two resulting groups from the first clustering stage.

as using SVM, the feature reduction is impressive. Now, the next step is to see how these features work on a clustering problem and see if using only three features will lead to a natural grouping of the subjects, or if the 68 features from SVM approach are needed for a good clustering result.

### 4.4.3 Data grouping using clustering methods

As it was explained previously, the idea of clustering is to check that the selected features in classifications are good by themselves and not because of the model. We are looking a natural grouping of the subjects when they are described only by these features. For this purpose three clustering methods were explored K-means, EM and X-means independently. It means if EM was used in the first stage, EM was also used in the second stage. No combination of methods was used.

As it was described in the Methods chapter. First stage leads into two groups and one of this group must be selected for the second stage. In order to select it, the distance histogram method was used. The histograms of both groups are represented in Figure 4.6.

For the first group (blue line) the majority of distances are small distances, whereas

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 12 | 0 | 0 | SZ |
| 0 | 12 | 0 | NC |
| 0 | 0 | 12 | CA |

Table 4.13: Confusion Matrix of the clustering results using X-means and Simple K-Means methods, and using features obtained from classification using SVM.

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 12 | 0 | 0 | SZ |
| 0 | 7 | 5 | NC |
| 0 | 3 | 9 | CA |

Table 4.14: Confusion Matrix of the clustering results using features obtained from classification using TREE.

for the second group (red line)distances are more spread out among medium-large distances. Due to that, group 2 has more possibilities of having clusters inside. Therefore, second group is selected for the second stage.

When features selected from SVM approach were used, using X-means and Simple K-Means resulted into a perfect clusterization. EM resulted in a 97% accuracy.

When features selected from decision TREE approach were used, maximum achievable accuracy resulted in 0.77. However, there is an important fact. The first stage has a hundred percent accuracy. This means that is possible to separate SZ from other subjects just using one feature.

In order to obtain the cheapest solution, for the first stage the feature selected from the decision TREE approach was used, and for the second stage features obtained from the SVM approach (31 features) were used. It resulted in a perfect clusterization too.

| SZ | NC | CA | Classified as |
|---|---|---|---|
| 12 | 0 | 0 | SZ |
| 0 | 12 | 0 | NC |
| 0 | 0 | 12 | CA |

Table 4.15: Confusion Matrix of the clustering results using X-means method, and using features obtained from classification using SVM.

## 4.5 Parameter exploration

The aim of this part of the project is to see the influence of each parameter in the final result.

Table 4.16 summarizes the accuracies obtained for the different parameter combination when EM clustering method was used in both stages. Different names in the first column correspond to different preprocessing parameters. There are two remarkable values in this table, first one the corresponding to preprocessing AR+lp10 Order 9. In this case subject's GC matrixes contained NaN values that did not allow creating a graph from where Graph Theory features could be extracted. So the explained procedure could not be applied. On the other case, preprocessing AR+lp20+iICA Order 9, fetures could be extracted. The problem was that an unsupervised clustering lead very little clusters instead of just 3. Better accurcay can be obtained with these features, but clustering must be guided.

BP1-80: Band pass filter between 1 and 80 Hz. AR+lp10: Filtering with 10Hz cut-off low pass filter, detrending,tappering and artifact removal. AR+lp20: Filtering with 20Hz cut-off low pass filter, detrending,tappering and artifact removal. AR+lp20 + iICA: Filtering with 20Hz cut-off low pass filter, detrending,tappering, artifact removal and N100 component enhancement using iICA method.

The best preprocessing approach when clustering is done using EM method is

| | Order 2 | Order 4 | Order 9 |
|---|---|---|---|
| BP1-80 | 0.61 | 0.63 | 0.83 |
| AR+lp10 | 0.63 | 0.5 | - |
| AR+lp20 | 0.97 | 1 | 0.55 |
| AR+lp20+iICA | 0.97 | 0.16 | 0 |

Table 4.16: Summary of obtained accuracies for different parameter configurations when clustering was done using EM Method.

using a 20HZ low pass filter to clean the data, then detrend, tapper it and remove artifacts (Artifact definition provided in Chapter 3.).

But how does each parameter affect the final accuracy? Figure 4.7 shows the effect of increasing the AR model order and Figures 4.8 and 4.9 show the number of used features to obtain those accuracies. The graph shows that in general, accuracy decreases when model order is increased. However, for Original data, the higher the model order higher the final accuracy. The reason could be that as this data is the most noise one, higher model orders are needed for obtaining same amount of information. On the other hand, when data is cleaned, higher model order end up in GC matrixes with few number of connections and therefore, less amount of information and consequently less accuracy. The graph also shows that for low AR model orders, removing artifacts (AR+lp10,AR+lp20 and AR +lp20+iICA) increases the accuracy, and that using a 20Hz low pass filter during the preprocessing increases it considerably compared to using a band pass filter between 1 and 80 Hz.

Do these accuracies remain equal when the clustering method is different? The best preprocessing approach when clustering is done using Simple K-Means method is using a 20HZ low pass filter to clean the data, then detrend, tapper it and remove artifacts (Artifact definition provided in Chapter 3.) both applying or without applying iICA.

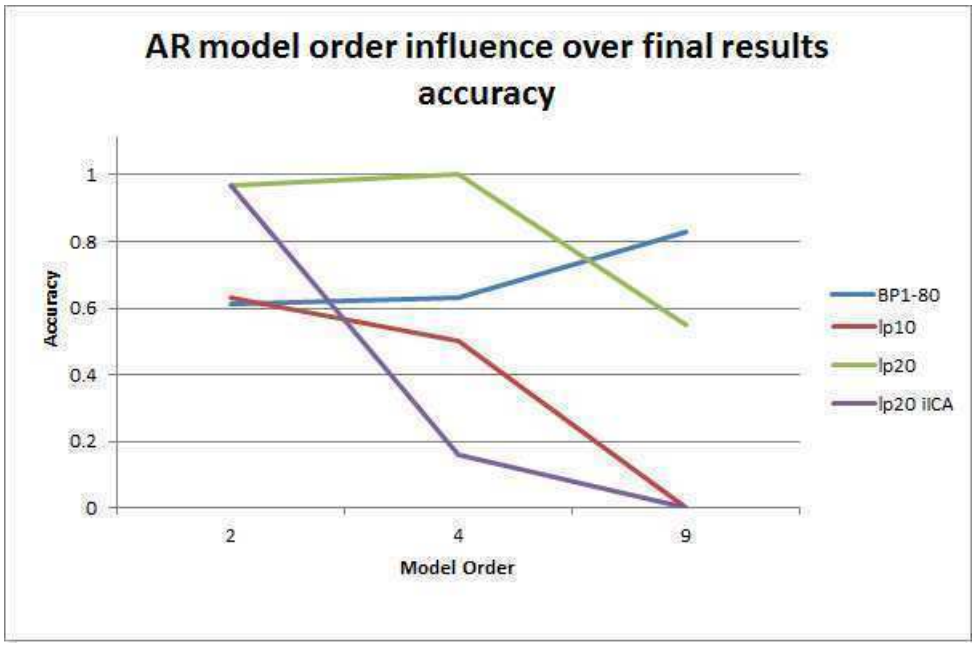



Figure 4.7: Graphical representation of AR model order effect over final result accuracy.Clustering done with EM method.

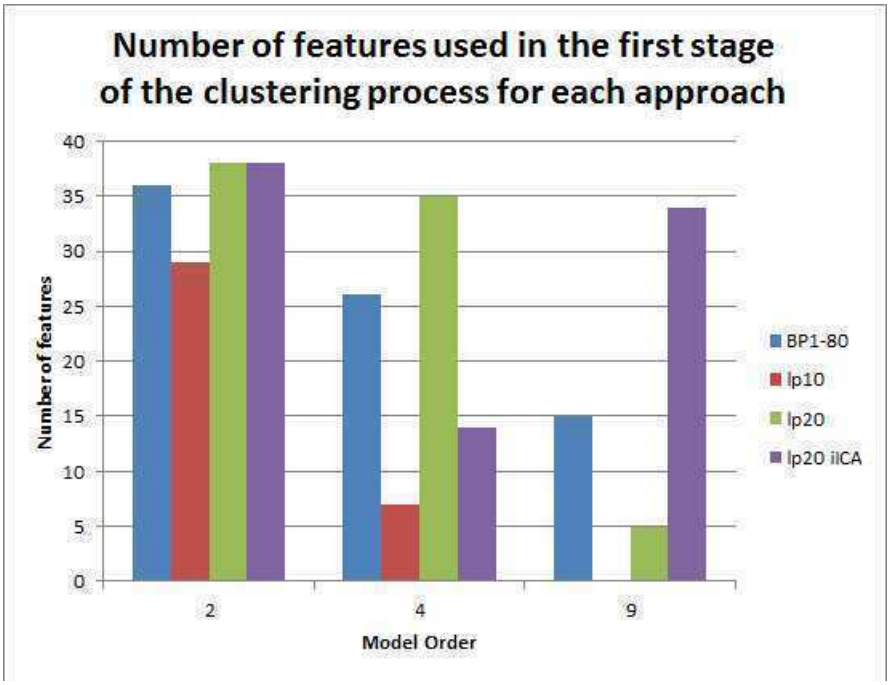



Figure 4.8: Number of used features in the first stage for accuracies shown in Figure 4.7.Clustering done with EM method.

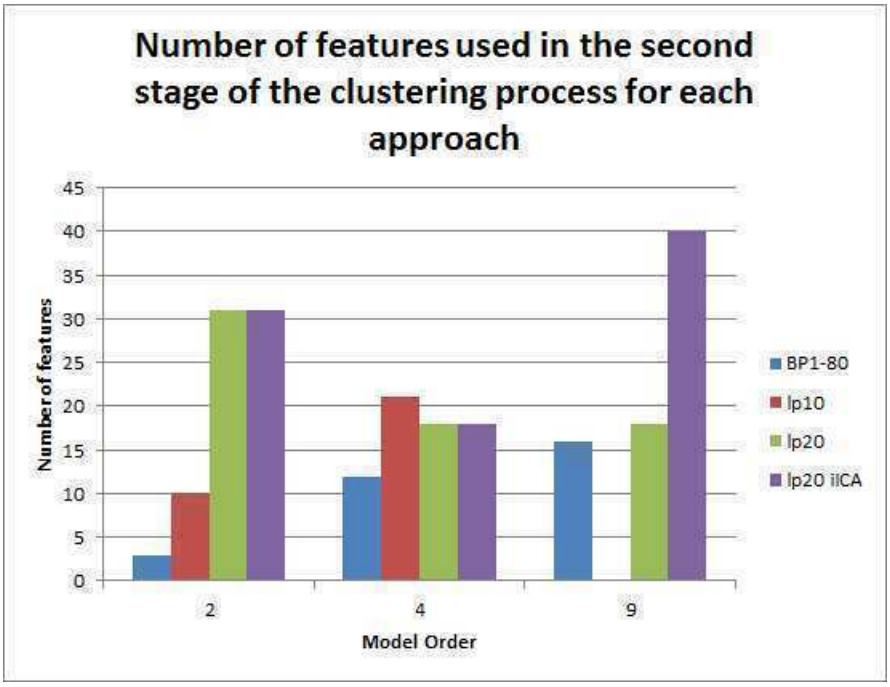



Figure 4.9: Number of used features in the second stage for accuracies shown in Figure 4.7.Clustering done with EM method.

Figure 4.10 shows the effect of increasing the AR model order and Figures 4.11 and 4.12 show the number of used features to obtain those accuracies when Simple K-Means clusterer was used. The graph shows that for all type of preprocessing using a higher AR model order will lead into a lower accuracy. Moreover, as it was expected, using a more complex preprocessing improves the final results accuracies, but only for low AR model orders.

The best preprocessing approach when clustering is done using Simple K-Means method is using a 20HZ low pass filter to clean the data, then detrend, tapper it and remove artifacts (Artifact definition provided in Chapter 3.) both applying or without applying iICA.

Figure 4.13 shows the effect of increasing the AR model order and Figures 4.14

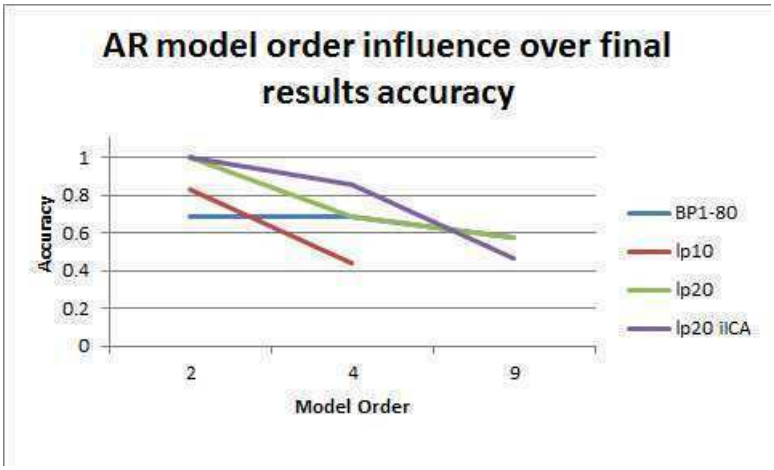



Figure 4.10: Graphical representation of AR model order effect over final result accuracy.Clustering done with Simple K-Means method.

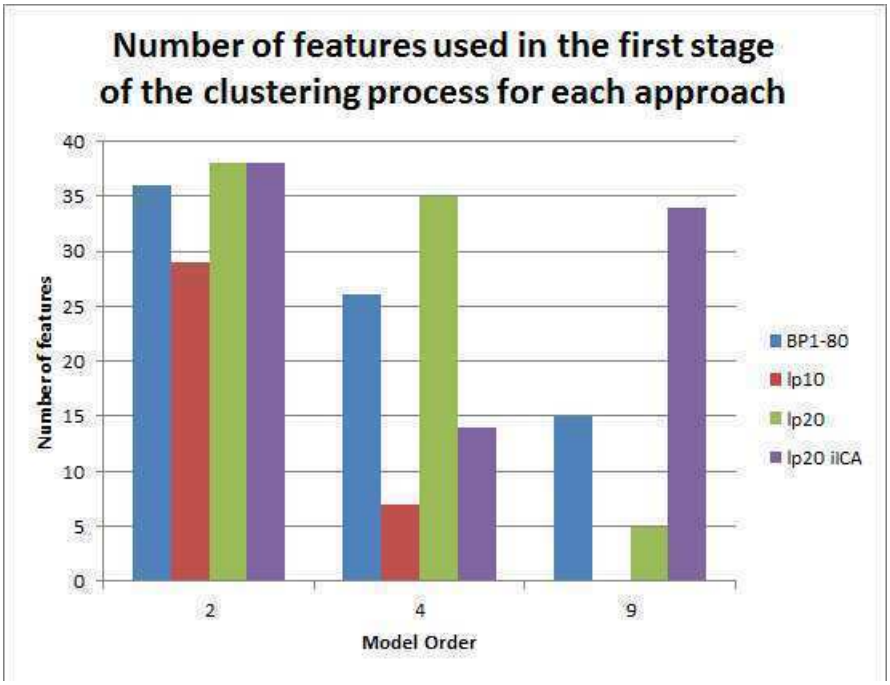



Figure 4.11: Number of used features in the first stage for accuracies shown in Figure 4.10.Clustering done with Simple K-Means method.

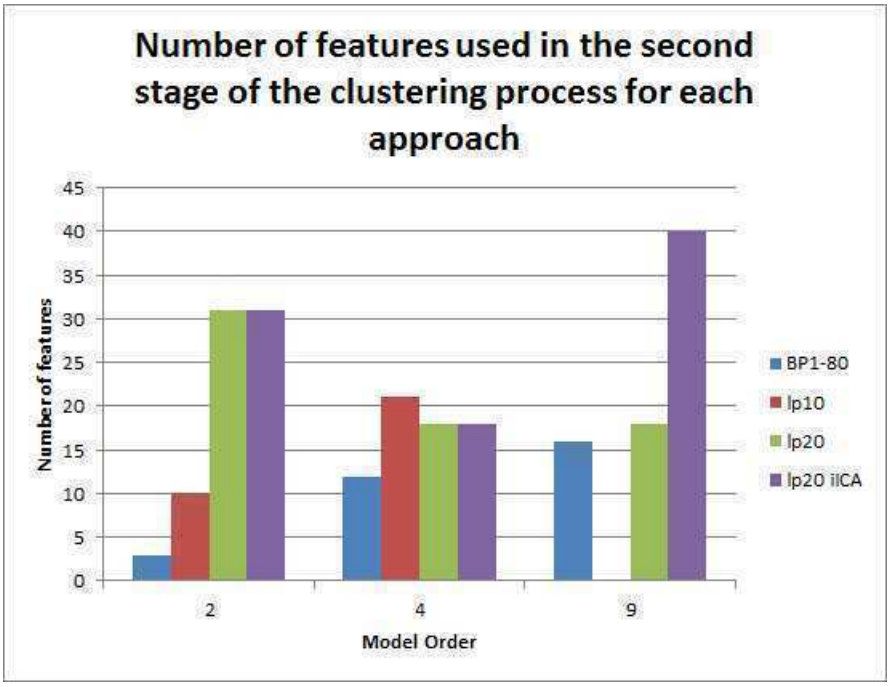



Figure 4.12: Number of used features in the second stage for accuracies shown in Figure 4.10.Clustering done with Simple K-Means method.

and 4.15 show the number of used features to obtain those accuracies. The most important information obtained from this graph is that all accuracies have improved when X-means was used comparing it with EM and Simple K-Means. The graph shows that for all type of preprocessing using a higher AR model order will lead into a lower accuracy. Moreover, as was expected, using a more complex preprocessing improves the final results accuracies, but only for low AR model orders.

Apart from this parameter exploration, HQC AR model order selecting criteria was used to see which was the best AR model order and to see if that model order would improve the accuracy. When criteria was used in data that had had AR+lp20+iICA preprocessing, the criteria suggested model orders of 16-17 depending the subject. The resulting GC matrixes were almost zero matrixes, they had really

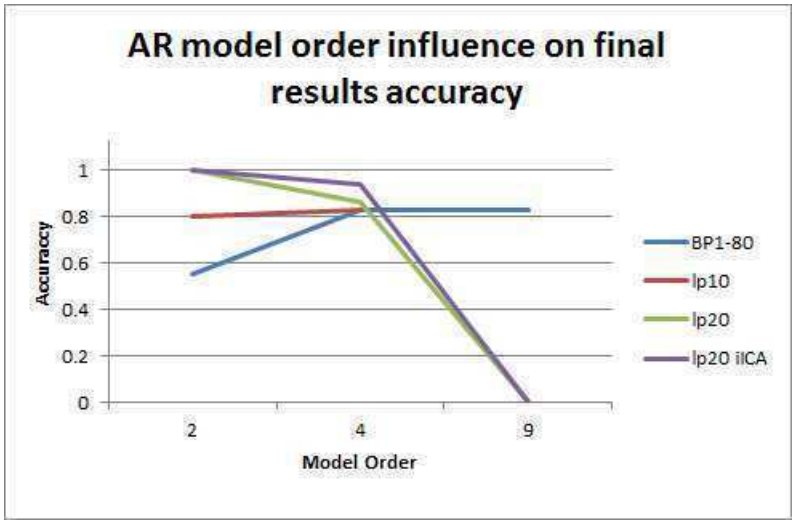



Figure 4.13: Graphical representation of AR model order effect over final result accuracy.Clustering done with X-Means method.

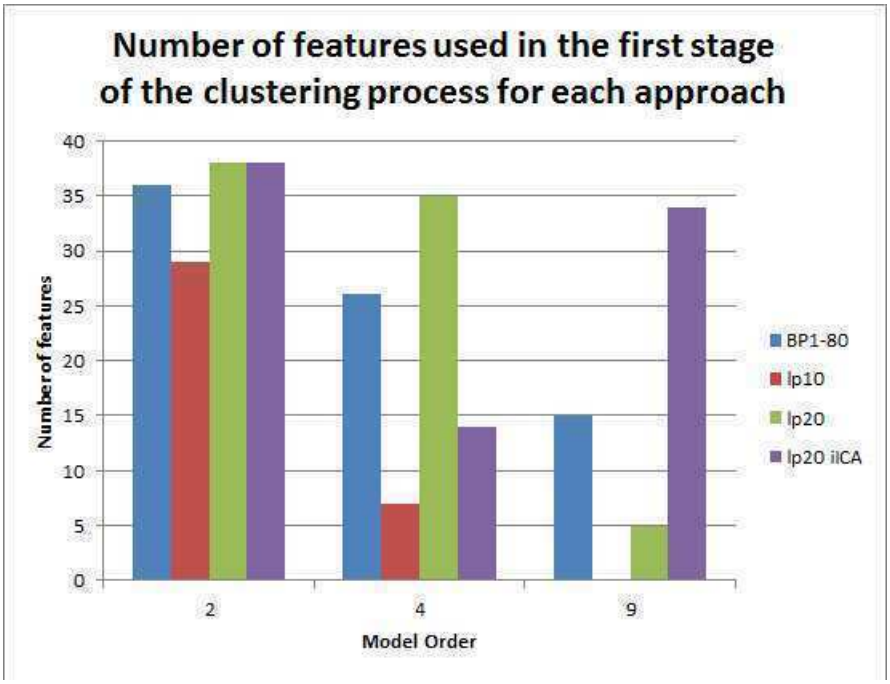



Figure 4.14: Number of used features in the first stage for accuracies shown in Figure 4.13.Clustering done with X-Means method.

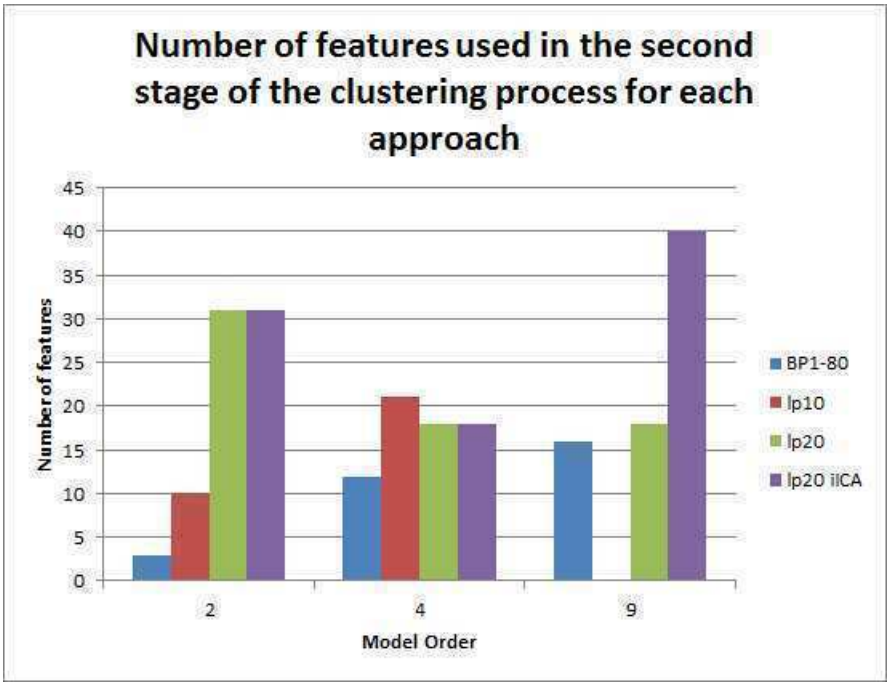



Figure 4.15: Number of used features in the second stage for accuracies shown in Figure 4.13.Clustering done with X-Means method.

few values showing connections. Then, GC and graph theory features were extracted and clustering was done using EM method. However, the result was no good. The clustering process divided the data into many little groups at the first stage. So, we can conclude that the best AR model order based on HQC is not the best one for this particular problem.

# Chapter 5

# Discussions and Conclusions

## 5.1 Accomplishments

In this Master Thesis, research efforts were presented on the process of unsupervised clustering for Schizophrenia Patients, Normal Controls and Cocaine users based on AEP. Our achievements are summarized in this chapter.

1. Provided data set was successfully clustered with 100% accuracy by different Clustering methods (EM, Simple K-Means and X-means)

2. Important features that describe and differentiate the three groups of subjects have been recognized.

3. A Methodology for correct Schizophrenia Patients, Normal Controls and Cocaine clustering has been proposed.

## 5.2 Discussion

As it has been shown in the Results chapter, Preliminary Exploratory Analysis with Unlabeled data did not give good results. For that reason the discussion section will

focus only on the procedures done with the labeled data.

### 5.2.1 Feature reduction using labels

One stage classification was not good enough to classify the data, it showed an accuracy of 58% whereas the two stage approach resulted in a 97% accuracy when SVM classification technique was used and 75% when TREE classifier was used. While SVM classifier needed 38 features for the first stage and 31 for the second, TREE classifier only needed 1 for the first and two for the second. So SVM classifier obtained a better results but it needed a higher amount of features.

#### Data grouping using clustering methods

When features selected from SVM classification were used, both Simple K-Means and X-Means techniques resulted in a perfect separation while EM resulted in a 97%. On the other hand, when features selected from TREE classification were used the maximum achievable accuracy was 77% but schizophrenia patients were perfectly separated in the first stage. The confusion was in the second stage between cocaine addicts and healthy controls. The most remarkable finding from this section is that, only one feature is needed to distinguish schizophrenia patients from the other two groups.

#### Parameter exploration

When parameter exploration was done using EM as a clustering technique, no trend was visible on the results. Apparently a parameter change resulted on an arbitrary change of the accuracy. However, when Simple K-Means and X-Means were used

a general trend appeared. In general using higher AR model orders resulted in a lower accuracy. Only in one case, rising the order was productive, in case of BP1-80 preprocessing.

On the other hand, it was observed that using a more complete preprocessing (low pass filter, detrend, tappering, iICA) increased considerably the final accuracy compared to the simple preprocessing (1-80Hz bandpass filter). Moreover, when lowpass filter was set at 20Hz resulted in better accuracy than when it was set to 10Hz.

Finally, using X-Means clustering technique gave better results than using EM or Simple K-Means, although Simple K-Means results were similar to ones obtained with X-Means.

## 5.3 Conclusions

Once all possibilities were explored the most important features have been identified and the best procedure is proposed in the next paragraphs.

### 5.3.1 Procedure

After all experiments were performed the best procedure was selected and it is described in the next paragraph.

1. Filter all trials with a 20Hz low pass filter
2. Detrend and tapper the signal
3. Remove artifacts
4. Obtain GC using AR model order 2

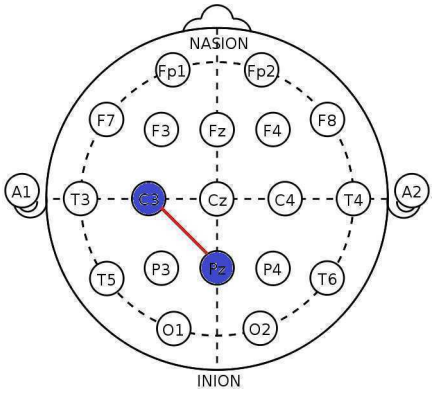



Figure 5.1: Influence of electrode C3 to Pz

5. Construct graph

6. From GC and graph obtained the important features (Listed in the next section)

7. Use Simple K-Means or X-Means to cluster the data in two steps

### 5.3.2 Important features

Schizophrenic, cocaine addicts and healthy controls can be separated with 100 percent accuracy if the correct features are selected. The most important feature is the influence of C3 to PZ in the response to the second stimuli.

This feature leads into a perfect separation of Schizophrenic patient from the rest of the subjects. Is the feature used in the first stage.

For a perfect prediction, 31 features are used in the second stage. This is the list of those features ranked from more importance to less.

1. S2GC PzT6

2. S2GC F4F7

3. S1GC Fp2M1

4. S1GC FzP3
5. S1GC PzT6
6. S1GC T5Pz
7. S1GC P3C3
8. S2GC O1Pz
9. S1GC F3M2
10. S1GC Fp2P4
11. S2GC P3F3
12. S2GC T5P4
13. S1GC F8M2
14. S1GC Fp2C4
15. S1GC O1M1
16. S2GC F3M2
17. S2GC T5Pz
18. S1GC T3T4
19. S1GC O2Pz
20. S2GC P3C3
21. S2GC Fp1O1

22. S2GC C3O1
23. S1GC M1T4
24. S1GC F7T4
25. S1GC F7Fp1
26. S1GC T4T3
27. S1GC FzFp2
28. S2GC A17A14
29. S2GC FzC4
30. S2GC O2Pz
31. S1GC T6F4

Although these 31 features come from different brain areas, lot of them involve C3,Pz,P3 F3 (location of these electrodes available in Figure 5.1) electrodes. So, it can be concluded that parietal lobe plays an important roll for clustering for Schizophrenia Patients, Normal Controls and Cocaine users.

## 5.4 Future Work

Obtaining a 100% accuracy for the unsupervised clustering is a good achievement. However, the number of used features is still high. A future research can focus on finding other kind of features (not based on AEP) that lead in the same accuracy but using a smaller amount of features. Another branch pf research could focus on seeing if these features differentiate also subjects with other mental disorders.

# Bibliography

[1] Attention and cognitive control laboratory:what is eeg? "http://www.lsa.umich.edu/psych/danielweissmanlab/whatiseeg.htm".

[2] Biomedcentral, crepalde et al. bmc genomics. http://www.biomedcentral.com/1471-2164/12/S4/S14/figure/F1.

[3] Electrode locations of international 10-20 system for eeg (electroencephalography) recording. "http://en.wikipedia.org/wiki/10-20_system_%28EEG%29", 2010.

[4] G. Zouridakis A. C. Papanicolaou. *FUNDAMENTALS OF FUNCTIONAL BRAIN IMAGING: A guide to the methods and their applications to psychology and behavioral neuroscience*. 2000.

[5] Oguz Demirci, Michael C Stevens, Nancy C Andreasen, Andrew Michael, Jingyu Liu, Tonya White, Godfrey D Pearlson, Vincent P Clark, and Vince D Calhoun. Investigation of relationships between fMRI brain networks in the spectral domain using ICA and Granger causality reveals distinct differences between schizophrenia patients and healthy controls. *NeuroImage*, 46(2):419–31, June 2009.

[6] Reinhard Diestel. *Graph Theory*. 2010.

[7] Gert J. Ter Horst Branislava Curcic-Blake Henderikus Knegtering Andr Aleman Edith J. Liemburg, Ans Vercammen. Abnormal connectivity between attentional, language and auditory networks in schizophrenia. *Schizophr. Res. 135*, pages 15–22, 2012.

[8] Richard E Frye and Meng-Hung Wu. Multichannel least-squares linear regression provides a fast, accurate, unbiased and robust estimation of Granger causality for neurophysiological data. *Computers in biology and medicine*, 41(12):1118–31, December 2011.

[9] Lee K. Preus A.P. McCarley R.W. Wible C.G.. Hashimoto, R. An fmri study of functional abnormalities in the evrbal working memory system and the relationship to clinicalysmptoms on chronic schizophrenia. *Cereb. Cortex 20*, pages 46–60, 2010.

[10] Thomas J. Ross Xiujuan Geng Wang Zhan Elliot A. Stein Yiohang Yang Hong Gu, Betty Jo Salmeron. Mesocorticolimbic circuits are impaired in chronic cocaine users as demonstrated by resting-state functional connectvity. *NeuroImage 53*, pages 593–601, 2010.

[11] Min-Wei Huang, Frank Huang-Chih Chou, Pei-Yu Lo, and Kuo-Sheng Cheng. A comparative study on long-term evoked auditory and visual potential responses between Schizophrenic patients and normal subjects. *BMC psychiatry*, 11:74, January 2011.

[12] a Hyvärinen and E Oja. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13(4-5):411–30, 2000.

[13] D J Iglesias. A Matlab toolbox for real time EEG functional mapping . *Unpublished master thesis, University of Houston, Houston, United States.*

[14] Darshan Iyer, Nash N Boutros, and George Zouridakis. Single-trial analysis of auditory evoked potentials improves separation of normal and schizophrenia subjects. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, pages 1–11, February 2012.

[15] L. H. N. N. B. Ben H. Jansen. auditory evoked potential variability in healthy and schizophrenia subjects. *Clinical Neurophisiology*, pages 1233–1239, 2010.

[16] Wible C.F. Hashimoto R. Kubicki M. Jeong, B. Functional and anatomical connectivity abnormalities in left inferior frontal gyrus in schizophrenia. *Hum. Brain Mapp. 30 159*, pages 4138–4151, 2009.

[17] Van Erp T. Poldrack R.A. Bearden C.E. Nuechterlein K.H. Cannon T.D. Karlsgodt, K.H. Diffusion tensor imaging of the superior longitudinal fascicculus and working memory in recent-onset schizophrenia. *Biol. Psychiatry 63*, pages 512–518, 2008.

[18] C C Kerr, C J Rennie, and P a Robinson. Physiology-based modeling of cortical auditory evoked potentials. *Biological cybernetics*, 98(2):171–84, February 2008.

[19] Westin C.F. Maier S.E. Frumin M. Nestor P.G. Salisburry D.F. Kikinis D.F. Jolesz F.A. McCarley R.W. Shenton M.E. Kubicki, M. Uncinate fasciculus findings in schizophrenia: a magnetic resonance diffusion tensor imagig study. *Am. J. Psychiatry 159*, pages 810–820, 2002.

[20] Victoria M Leavitt, Sophie Molholm, Walter Ritter, Marina Shpaner, and John J Foxe. Auditory processing in schizophrenia during the middle latency period (10-50 ms): high-density electrical mapping and source analysis reveal subcortical antecedents to early cortical deficits. *Journal of psychiatry & neuroscience : JPN*, 32(5):339–53, September 2007.

[21] Branch C.A. DeLisi LE. Li, X. Language pathway abnormalities in schizophrenia: a review of fmri and other imaging studies. *Curr. Opin. Psychiatry 22*, pages 131–139, 2009.

[22] Branch C.A. Nierenberg J.DeLisi LE. Li, X. Disturbed functional connectivity of cortical activation during semantic discrimination in patients with schizophrenia and subjects at genetic high-risk. *Brain Imaging Behav. 4*, pages 109–120, 2010.

[23] Edith J Liemburg, Ans Vercammen, Gert J Ter Horst, Branislava Curcic-Blake, Henderikus Knegtering, and André Aleman. Abnormal connectivity between attentional, language and auditory networks in schizophrenia. *Schizophrenia research*, 135(1-3):15–22, March 2012.

[24] Venus Khim-Sen Liew. Which lag length selection criteria should we employ? *Universiti Putra Malaysia*.

[25] Manoranjan Dash Kiseok Choi Peter Scheuermann Huan Liu. Feature selection for clustering  a filter solution. *Northwestern University, Arizona State University.*

[26] Elaine N. Marieb. *Essentials of Human Anatomy and Physiology with Essentials of InterActive Physiology CD-ROM, 9th Edition.* 2008.

[27] Scott McGlashan, Thomas H.Woods. *Psychiatric Times; Mar2011, Vol. 28 Issue 3,*, 2011.

[28] Andrew Ng. *CS229 Lecture notes.*

[29] A. Peled. *Multiple constraint organization in the brain: a theory for schizophrenia.* 1999.

[30] Luca Pollonini, Swaroop Pophale, Ning Situ, Meng-Hung Wu, Richard E Frye, Jose Leon-Carrion, and George Zouridakis. Information communication networks in severe traumatic brain injury. *Brain topography*, 23(2):221–6, June 2010.

[31] Oxford University Press. *Concise Medical Dictionary.* 2010.

[32] Van Erp T.G.M. Hardt M.E. Dapretto M. Caplan R. Cannon T.D. Bearden C.E. Sabb, F.W. Language network dysfunction as a predictor of outcome in youth clinical high risk for psychosis. *Schizophr. Res. 116*, pages 173–183, 2010.

[33] Corkery John2 Schifano, Fabrizio. *Journal of Psychopharmacology; Jan2008, Vol. 22 Issue 1,*, pages p71–79,.

[34] M. Karimi F. Babazadeh Sh. Khorshidi, M. Towhidi. Optimal predictive order selection criterion for the autoregressive (ar) model. *International Review of Automatic Control (I.RE.A.CO.), Vol. 2, N. 6*, 2009.

[35] Rodriguez Carmen Moreno Valverde, Pilar Ramos. *Health and Addictions / Salud y Drogas; 2010, Vol. 10 Issue 2,*, pages p13–36, 24p.

[36] Jeffrey Weihing, Shannon Daniels, and Frank E. Musiek. The Effect of Visual and Audiovisual Competition on the Auditory N1-P2 Evoked Potential. *Journal of the American Academy of Audiology*, 20(9):569–581, October 2009.

[37] Meng-Hung Wu, Richard E Frye, and George Zouridakis. A comparison of multivariate causality based measures of effective connectivity. *Computers in biology and medicine*, July 2011.

[38] G Zouridakis, D Iyer, J Diaz, and U Patidar. Estimation of individual evoked potential components using iterative independent component analysis. *Physics in medicine and biology*, 52(17):5353–68, September 2007.

[39] G Zouridakis, B H Jansen, and N N Boutros. A fuzzy clustering approach to EP estimation. *IEEE transactions on bio-medical engineering*, 44(8):673–80, August 1997.

[40] George Zouridakis. Unsupervised clustering of schizophrenia patients, cocaine users, and normal controls based on auditory evoked potentials. *Presentation on Graduate Research Day at College of technology, University of Houston.*

# Appendix A

# Subject Description

The following table gives a detailed description of each subject and also provides the labels for identification purpose.

| Schizophrenia | Subject | Sex | Race | Age | Smoker |
|---|---|---|---|---|---|
| CLA | 24 | F | B | 31 | smoker |
| MIE | 17 | M | W | 38 | smoker |
| ROB | 18 | M | B | 54 | smoker |
| PR 1042 | 23 | M | B | 46 | smoker |
| PR 1049 | 28 | F | W | 51 | smoker |
| PR 1052 | 29 | M | B | 53 | smoker |
| PR 1053 | 10 | M | B | 55 | smoker |
| PR 1054 | 11 | M | W | 42 | Non-smoker |
| PR 1055 | 26 | M | B | 22 | Non-smoker |
| PR II subj3 | 35 | M | W | 38 | smoker |
| PR II subj4 | 15 | F | B | 31 | smoker |
| PR II subj14 | 27 | M | B | 45 | Non-smoker |
| Age mean | 42.17 | | | Standard deviation | 10.50 |
| Healthy | Subject | Sex | Race | Age | Smoker |
| PR 2050 | 22 | F | B | 52 | Non-smoker |
| PR 2051 | 3 | M | B | 46 | Non-smoker |
| PR 2052 | 21 | F | B | 45 | Non-smoker |
| PR 2053 | 8 | M | B | 41 | Non-smoker |
| PR 2054 | 13 | F | B | 48 | Non-smoker |
| PR 2055 | 30 | M | B | 38 | Non-smoker |
| PR 2056 | 2 | F | W | 50 | Non-smoker |
| PR 2057 | 9 | F | B | 36 | smoker |
| PR 2058 | 25 | F | W | 26 | Non-smoker |
| PR 2060 | 19 | M | B | 37 | smoker |
| cTMS 160 | 12 | F | B | 44 | Non-smoker |
| cTMS 576 | 32 | M | B | 43 | Non-smoker |
| Age mean | 42.17 | | | Standard deviation | 7.16 |
| Cocaine | Subject | Sex | Race | Age | Smoker |
| cTMS 116 | 33 | M | B | 43 | Non-smoker |
| cTMS 317 | 31 | F | B | 41 | smoker |
| cTMS 447 | 20 | M | B | 44 | smoker |
| cTMS 476 | 36 | M | B | 40 | smoker |
| cTMS 482 | 16 | M | W | 37 | smoker |
| cTMS 551 | 7 | M | B | 57 | smoker |
| cTMS 577 | 4 | M | B | 48 | smoker |
| cTMS 707 | 5 | M | B | 47 | smoker |
| cTMS 745 | 34 | M | O | 43 | smoker |
| cTMS 825 | 1 | M | B | 24 | smoker |
| cTMS 976 | 14 | F | B | 29 | Non-smoker |
| cTMS 1001 | 6 | M | B | 53 | smoker |
| Age mean | 42.17 | | | Standard deviation | 9.22 |

Table A.1: Description of all the subjects divided in three categories: Schizophrenia patients, Healthy Controls and Cocaine Dependents.

# Appendix B

# Xie S criterion

frame

```
1 function [S U]=Sfcm(data,k)
2 %----------%----------%----------%----------%----------%----------
3 % This program evaluates the performance of the fcm clustering for a
      given
4 % number of clusters k, and features data
5 %----------%----------%----------%----------%----------%----------
6 % Ainhoa Bastarrika: December, 2011
7
8
9 nsubjects=size(data,1);
10
11
12 [center,U] = fcm(data,k,[1.2 100 1e-5 0]);
13
14
```

```
15 d=zeros(nchoosek(k,2),1);
16 h=1;
17 for i=1:k
18     m=k-i;
19     d(h+(0:m-1))= pdist2(center(i,:),center(i+1:k,:),'Euclidean');
20     h=h+m;
21
22 end
23 dmin=min(d);
24 VX=zeros(k,nsubjects);
25
26
27
28 for i=1:nsubjects
29
30         VX(1:k,i) = pdist2(center(1:k,:),data(i,:),'Euclidean');
31 end
32 s=0;
33 for i=1:k
34 s=s + U(i,:).*VX(i,:);
35 end
36 S=sum(s)/(nsubjects*dmin);
37
38 end
```