Experimental Setup to Test Neurological Artifact Elimination Techniques

Study of ICA for Artifact Elimination in Neuroscience Signals

Magdalena Tobar

Supervisor: Ramele Rodrigo

Final Project

Data Science Specialization

Abstract

Reading the electrical activity of the brain is a widely spread methodology used for diagnosis purposes. Specifically, it is applied for the diagnosis of brain pathologies. This methodology, also has applications in Brain Computer Interfaces, in which computers can read the electrical activity of the brain and convert it into an artificial output, such as motion or activity. This technology is particularly useful in medicine, for patients with motor disabilities, which could be solved through the use of technology.

However, such measurements do not come without a challenge. The electrical brain signals are usually accompanied by unwanted signals, which may affect the resulting data by contaminating measurements of interest. These signals are known as artifacts.

Independent Component Analysis (ICA) is a blind source separation technique, which requires manual intervention to reject independent components with visually detected artifacts after decomposition.

This work proposes a methodological approach to the insertion of a particular artifact produced by eyes blinking, in EEG data, in order to create a pseudo-real dataset, which in turn provides an experimental setup to test artifact removal algorithms. The topic of ocular waves' propagation throughout EEG channels is also covered. The optimal amount of ICA components is discovered through an iterative approach, which includes an observation step.

Keywords – Artifacts, EEG, Biosignals, ICA, Electrophysiology.

Resumen

La lectura de la actividad eléctrica del cerebro es una metodología muy utilizada, especialmente con fines de diagnóstico. En concreto, se aplica para el diagnóstico de patologías cerebrales. Esta metodología, también tiene aplicaciones en las interfaces cerebro-computadora, en las que las computadoras pueden leer la actividad eléctrica del cerebro y convertirla en una respuesta artificial, como realizar un movimiento o una actividad. Esta tecnología es especialmente útil en medicina, para pacientes con discapacidades motoras, que podrían solucionarse mediante el uso de ésta tecnología.

Sin embargo, estas mediciones no están exentas de dificultades. Las señales eléctricas del cerebro suelen ir acompañadas de señales no deseadas, que pueden afectar a los datos resultantes, contaminando las mediciones de interés. Estas señales se conocen como artefactos.

El Análisis de Componentes Independientes (ICA) es una técnica de separación de fuentes, que requiere la intervención manual para rechazar los componentes independientes que contienen artefactos, detectados visualmente después de la descomposición.

Este trabajo propone un enfoque metodológico para la inserción de un artefacto particular producido por el parpadeo de los ojos, en los datos de EEG, con el fin de crear un conjunto de datos pseudo-real, que a su vez proporciona un ambiente experimental para probar algoritmos de eliminación de artefactos. También se trata el tema de la propagación de las ondas oculares a través de los canales de EEG. La cantidad óptima de componentes ICA se descubre mediante un enfoque iterativo, que incluye un paso de observación.

Palabras Clave – Artefactos, Electroencefalograma, Bioseñales, ICA, Electrofisiología.

Contents

List of Figures

1 Introduction

Reading the electrical activity of the brain is a widely spread methodology used for diagnosis purposes. Specifically, it is applied for the diagnosis of brain pathologies. This methodology, also has applications in Brain Computer Interfaces, in which computers can read the electrical activity of the brain and convert it into an artificial output, such as motion or activity. This technology is particularly useful in medicine, for patients with motor disabilities, which could be solved through the use of technology.

However, such measurements do not come without a challenge. The electrical brain signals are usually accompanied by unwanted signals, which may affect the resulting data by contaminating measurements of interest. This signals are known as artifacts.

As part of this work, the electrophysiology of the brain will be reviewed in chapter 2, presenting the Brain Computer Interfaces pipeline and deep diving into the different types of unwanted signals which can interfere when reading the electrical activity of the brain.

In chapter 3, the existing methods to remove such unwanted signals from electrical brain data will be reviewed.

In chapter 4, a methodology will be proposed to generate a pseudo-real dataset, which will give us an experimental setup to enable us to test the performance of using ICA on the generated pesudo-real dataset.

In chapter 5, the results of the experimentation will be presented and chapter 6 will include the conclusions of this work.

The source code for this work, can be found in a public repository (Tobar, 2021).

2 Electrophysiology

2.1 Brain Computer Interfaces

Brain Computer Interfaces have enabled non-biological communication between the central nervous system and computer devices by measuring Central Nervous System (CNS) activity and converting into an artificial output that replaces, restores, enhances, supplements, or improves natural CNS output and thereby changes the ongoing interactions between the CNS and its external and internal environment (Wolpaw and Wolpaw, 2012). There are two ways in which this communication may take place: through invasive procedures or non-invasive procedures. Invasive procedures involve complicated and risky surgical procedures, while non-invasive procedures involve no risk for the patient. At this point, electroencephalography has proven the most widespread non-invasive method for obtaining information from the central nervous system (Ramele et al., 2019).

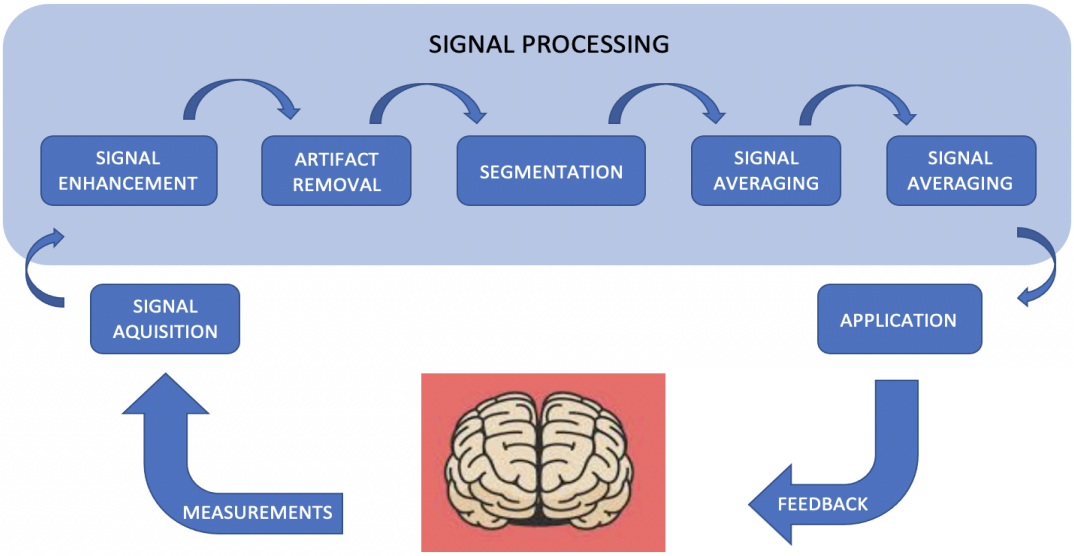


Figure 2.1: Brain Computer Interfaces

2.1.1 Signal Acquisition

2.1.1.1 Electroencephalography

Electroencephalography (EEG) is a diagnosis method which involves recording the electrical activities of the brain, by placing two or more electrodes on the scalp. It is typically

described in therms of rhythms and transients, divided into bands of frequency. The most common EEG rhythms are delta, theta, alpha and betta waves, and in certain cases gamma waves as well (Islam et al., 2016). It is commonly used for diagnosing brain pathologies: detection of type and location of epileptic activity, analysis of sleep disorders, encephalopathies, neurological infections, dementia, etc (Urigüen and Garcia-Zapirain, 2015).

EEG signals are composed of different oscillations, also called "rhythms", which have particular properties in terms of spatial and spectral localization. The frequency of EEG signals range from 0.01 Hz to around 100 Hz (Jiang et al., 2019). There are 6 main brain rhythms (Lotte, 2008):

- Alpha: These are oscillations (8-12 Hz frequency band), which appear mainly in the posterior regions of the head (occipital lobe) when the subject has closed eyes or is in a relaxed state.
- Beta: This is a relatively fast rhythm (13-30 Hz frequency band) observed in awake and conscious persons. It is affected by the performance of movements, in the motor areas.
- Delta: This is a slow rhythm (1-4 Hz frequency band), with a relatively large amplitude, which is mainly found in adults during deep sleep.
- Gamma: This rhythm concerns mainly frequencies above 30 Hz, having a maximal frequency around 80 Hz or 100 Hz. It is associated to various cognitive and motor functions.
- Mu: These are oscillations in the 8-13 Hz frequency band, located in the motor and sensorimotor cortex. The amplitude of this rhythm varies when the subject performs movements.
- Theta: This a slightly faster rhythm (4-7 Hz frequency band), observed mainly in young children and in adults during drowsiness.

In EEG, the studied signals are usually accompanied by unwanted signals, which may affect the results of studies by contaminating measurements of interest in both temporal and spectral domains with wide frequency band (Islam et al., 2016). These signals

are known as artifacts. Some artifacts can be avoided by being careful at the time of measurement, these are the ones associated with external electromagnetic sources. There is also physiological contamination to be taken into account. These signals come from biological sources, but are external to brain activity. Such is the case of cardiac activity, eye movements, blinking and muscular activity (Urigüen and Garcia-Zapirain, 2015). There will always be some artifacts present in the recording and those should be handled in the digital signal processing domain (Islam et al., 2016).

2.1.2 Signal Processing

At the moment of extraction, signals are obtained in its raw form. However, before being able to classify them, they must be cleaned and denoised, in order to enhance the relevant information embedded in the signals (Lotte, 2008). This consists of the following steps (Ramele et al., 2019):

- Signal Enhancement
- Artifact Removal
- Segmentation
- Signal Averaging
- Feature Extraction

Throughout this work we will be focusing in different algorithms used for artifact removal.

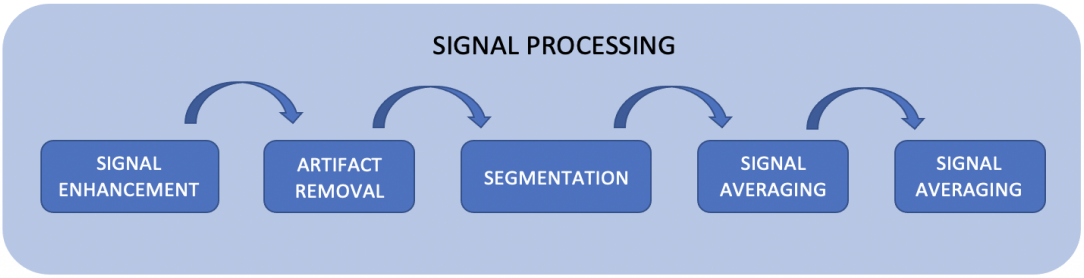


Figure 2.2: Signal Processing

2.1.2.1 Signal Enhancement

In order to enhance the relevant information within a signal, two kinds of filters are applied.

- Temporal Filters: Used to restrict the analysis to frequency bands in which we know the neurophysiological signals are (Lotte, 2008). Within the Temporal Filters, Spectral Filters are those designed for their effects on the spectrum of the signal, which help isolate particular oscillations of interest.
- Spatial Filters: Used in BCIs to remap channel signals to approximate source signals. They are designed to recover motor-cortex source activity.

2.1.2.2 Artifact Removal

The following artifact removal methods will be reviewed:

- Regression Methods
- EOG Correction Method
- Filtering Methods
- Blind Source Separation Methods
- Source Decomposition Method
- Hybrid Methods

2.1.2.3 Segmentation

Segmentation refers to subdividing the EEG data into different segments, also known as epochs, based on a number of similar properties such as average frequency components (Procházka et al., 2010).

2.1.2.4 Signal Averaging

Signal Averaging is a processing technique applied in the spatial domain, which allows the estimation of small amplitude signals that are buried in noise. It enhances time-locked signals components in noisy measurements (Van Drongelen, 2018).

2.1.2.5 Feature Extraction

EEG measurements lead to huge amounts of data. In order to improve prediction performance, it is necessary to find a smaller number of features, which describe relevant

properties of the signals (Lotte, 2008). Feature Extraction is the process through which one or several signals are transformed into a feature vector.

2.1.3 Signal Classification

A key part of BCI is being able to translate the obtained brain signals into commands for the computer. This step can be achieved through both, regression and classification algorithms. Classification algorithms are used the most in these kind of interfaces. The goal of this step is to assign a class to the extracted vector and each class represents a kind of mental task, performed by the BCI user (Lotte, 2008). Classification algorithms use labeled training sets in order to learn to identify unlabeled vectors.

2.2 Description of Artifacts

As mentioned earlier, artifacts often overlap with EEG signals in both spectral and temporal domains such that it becomes difficult to use simple filtering or straight forward signal processing technique (Islam et al., 2016). Detecting these anomalies which affect the results of studies by contaminating measurements of interest, requires being able to identify different types of artifacts. In this section, we will deep dive into the different types of artifacts.

2.2.1 Technical Artifacts

Technical artifacts are associated with external electromagnetic sources which affect the outcome of the EEG. These can be avoided by being careful at the time of measurement, attaching electrodes properly to the scalp and experimenting in a controlled environment. Otherwise, outside sources superimpose their energy to the observed EEG because of faults in the setting or conditions (Kanoga and Mitsukura, 2017).

2.2.2 Biological Artifacts

Biological artifacts come from activity within the human body but external to the brain. Hence, they can rarely be avoided, which is why they are considered the most prevalent contaminants in the literature on EEG artifact removal (Urigüen and Garcia-Zapirain, 2015).

2.2.2.1 Ocular Artifacts

Eye movement and blinking are mainly detected by frontal electrodes and both can cause the EEG to become contaminated. Blinking is usually associated with higher peaks in the EEG and higher frequency interference, while eye movement produce smaller interference. Electrooculogram (EOG) can be measured together with EEG, making it easier to detect and isolate these types of artifacts (Urigüen and Garcia-Zapirain, 2015).

2.2.2.2 Muscle Artifacts

Contracting muscles produce electrical activity on the body surface. These type of artifacts are harder to identify because they may vary on their intensity, depending on the location of the muscle and the degree of muscle concentration (Urigüen and Garcia-Zapirain, 2015).

2.2.2.3 Cardiac Artifacts

The electrical activity of the heart is another factor which may interfere with the proper measurement of brain electrical activity. If electrodes are positioned in a pulse area, this can also affect EEG, but it is easily avoidable by proper electrode placement. However, having an Electrocardiogram measures together with the EEG can help identify and isolate the waves produced by the hearts electrical activity.

2.2.2.4 Other Biological Artifacts

Other possible factors which may arise, that interfere with proper EEG are sweat and perspiration, tongue movement, metallic dental fillings and breath movement.

3 Review of Artifact Removal Techniques

In this section, we will review the existing methods for artifact identification and removal.

3.1 Regression Methods

Regression algorithms are a simple method to reduce artifacts and have low computational demands. This technique may be applied to the time or frequency domains, by estimating the influence of the reference waveforms on the signal of interest (Urigüen and Garcia-Zapirain, 2015).

The artifact would be corrected by calculating propagation factors to estimate the relationship between a reference signal and the observed EEG signal and subtracting the regressed portion. These cancels the cerebral information from each observed EEG signal upon linear subtraction which in turn means that important nontime-locked components will be lost by the averaging operation (Kanoga and Mitsukura, 2017).

In late years, regression methods have been replaced by more sophisticated algorithms which do not require a reference channel.

3.2 EOG Correction Method

The electroculogram subtraction method assumes that EEG is a linear combination of a brain signal and ocular movement signal, based on a linear regression. This regression calculates the portion of EOG present in each of the measured channels. It is corrected by subtracting the EOG from the EEG in each of the channels (Urigüen and Garcia-Zapirain, 2015).

However, this method may not take bidirectional contamination into account, which may in turn discard relevant information. More sophisticated versions of this method solve this issue in different ways, such as lowpass filtering the EOG channels.

3.3 Filtering Methods

Filtering methods are one of the classical and simple separation attempts to remove artifacts from an observed EEG signal. These methods try to adapt the filter parameters to minimize the mean square error between the estimated EEG and the desired original signal (Urigüen and Garcia-Zapirain, 2015).

3.3.1 Fixed Gain Filtering

This method is only effective if the spectral distribution of the EEG and artifact do not overlap, but is is not adequate for biological artifacts because it affects the original EEG signals (Kanoga and Mitsukura, 2017).

3.3.2 Adaptive Filtering

This method assumes that the EEG signal and the artifact are uncorrelated, and therefore the artifact is considered to be additive noise within the observed signal (Kanoga and Mitsukura, 2017). The filter generates a signal correlated with the artifact using a reference signal and then the estimate is subtracted from the acquired EEG (Urigüen and Garcia-Zapirain, 2015). The filter weights can adapt based on the feedback from output of the system and a reference input is required to compare the desired output with the observed output (Islam et al., 2016).

Adaptive filtering approach has a potential to recover "pure" EEG signal more rapidly and accurately than linear regression for ocular and cardiac artifacts (Kanoga and Mitsukura, 2017).

3.4 Blind Source Separation Methods

It is one of the most popular artifact removal methods, which aims to extract individual unknown source signals from their mixtures and to estimate unknown mixing channels using only information from the mixtures observed at the output of each channel, with limited knowledge on the source signal and mixing channel (Islam et al., 2016). This method has no need for a reference channel, but in order to separate the components, the different channels must be either independent or maximally uncorrelated.

3.4.1 Independent Component Analysis

Independent Component Analysis (ICA) is the most widespread used algorithm to decompose multi-channel data into independent components (Kanoga and Mitsukura, 2017). This method imposes statistical independence of the sources (Urigüen and Garcia-Zapirain, 2015). It requires manual intervention to reject independent components (ICs) with visually detected artifacts after decomposition. However, artifact detection and removal can be made automatic by labeling the ICs through some features that can quantify the possibility of being artifactual (Islam et al., 2016).

3.4.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a blind source separation method which assumes that the different channels are maximally uncorrelated. This method is capable of finding uncorrelated components, therefore taking temporal correlations into account (Islam et al., 2016). CCA measures the linear relation between two multi-dimensional random variables and can be applied to solve the BSS problem by taking the source vector as the first multi-dimensional random variable and a temporally delayed version of the source vector, as the second multi-dimensional random variable (Urigüen and Garcia-Zapirain, 2015).

3.5 Source Decomposition Method

This method consists of decomposing each individual channel into basic waveforms that represent separately signal and artifact, in order to remove the latter. Source decomposition is based in the idea that some source can be represented by a single decomposition unit (Urigüen and Garcia-Zapirain, 2015).

3.5.1 Wavelet Transform

It is a time-scale representation method that decomposes a signal into basis functions of time and scale, which are dilated (multiplying by a scaling factor) and translated (considering all possible integer translations) versions of a basis function called mother wavelet (Islam et al., 2016). Artifact removal based on the WT relies on the sources of

interest being decomposable on a wavelet basis, whereas artifacts cannot (depending on the type of signal, it may be the artifacts that have a better defined wavelet decomposition, for instance consider background EEG and blinks) (Urigüen and Garcia-Zapirain, 2015).

3.5.2 Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) is a computationally complex method, which performs non-stationary, non-linear stochastic processes. The algorithm decomposes signals into a sum of band-limited components called Intrinsic Mode Functions (IMF) (Islam et al., 2016), which are amplitude and frequency modulated zero mean components, plus a non-zero mean low-degree polynomial remainder (Urigüen and Garcia-Zapirain, 2015).

3.6 Hybrid Methods

Recently, the advantages of the different methods have been reinforced by combining them together into a single method.

A commonly used example of a hybrid method is WICA, which results of the combination of Wavelet and ICA methods. The idea behind both procedures is to filter out cerebral activity that may be leaked into the artifact components and would be lost by simply removing the unfiltered components. Their WICA algorithm first partitions the EEG recording into four EEG subbands, then selects the artifact-linked wavelet components and passes them through ICA. To end, the independent components related to the artifacts are found and cancelled out (Urigüen and Garcia-Zapirain, 2015). There are similar methods which can be applied to single-channel EEG data.

4 Materials and Methods

4.1 Artificial Dataset Generation

Having a dataset where artifacts are identifiable is a complex procedure which presents a problem when testing different artifact identification algorithms, since the underlying ground-truth is unknown.Which is why, in order to test the performance of some of the studied Artifact Elimination Methods, a semi-simulated EEG dataset was used, containing pre-contamination EEG signals (Ramele et al., 2018a). These semi-simulated EEG signals were manually contaminated with ocular artifacts.

The peculiarity of the base EEG data used, corresponding to subject 21 from the dataset "EEG waveform analysis of P300 ERP with applications to brain computer interfaces" (Ramele et al., 2018b) is that this particular individual produced very low contamination from blinks and ocular moves on the generated EEG stream. This dataset based on an EEG stream, was generated under a passive modality in which real P300 ERP templates obtained from a public dataset were superimposed into the generated EEG stream of four subjects, who were instructed to passively watch a flashing screen while not focusing on any particular letter (Ramele et al., 2018a). This data was collected in a single recording session. Electrodes were used on locations Fz, Cz, Pz, Oz, P3,P4, PO7 and PO8 according to the 10–20 international system (Ramele et al., 2018a). Sampling frequency was set to 250 Hz. Figure 4.1 shows 5 seconds of a sample 8-channel EEG signal.

The blinks information was obtained taking a single measure from a single subject using the Mindwave Mobile device manufactured by NeuroSky Inc. The device contains a single sensor on frontal lobe and neutral points on ears by ear clips (Katona et al., 2016). The measuring period was fifty seconds and sampling frequency was 128 Hz. Four different types of blinks were extracted from the aforementioned recording session. Figure 4.2 shows a plot of the signal measured. Figure 4.3 shows the four different kinds of blinks identified in the obtained signals.

In order for the blinks injection into the EEG to be a smooth one, the following steps were taken:

1. Weighing: Blinks were given a higher weight in channels located closer to the eyes.

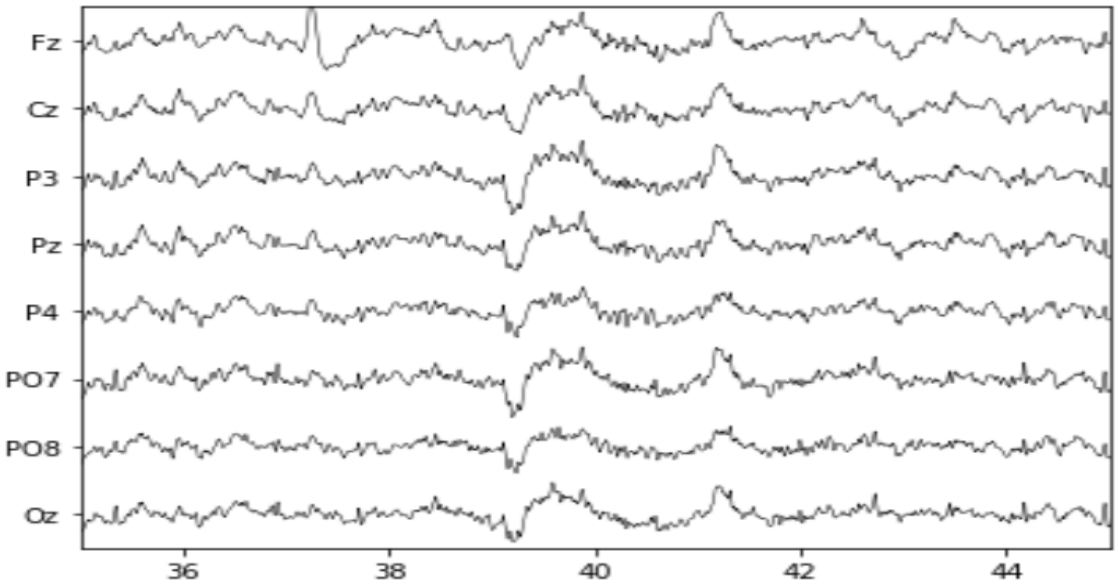


Figure 4.1: Sample 8-channel EEG signal

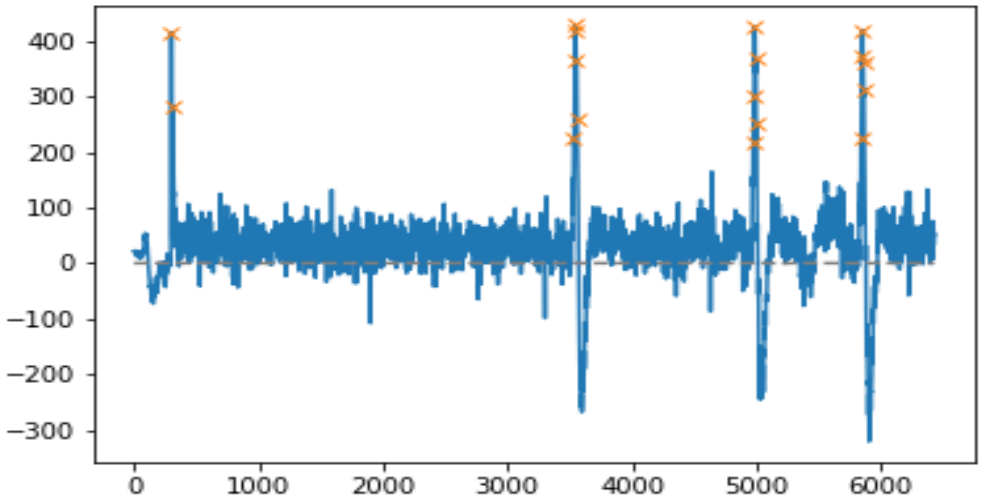
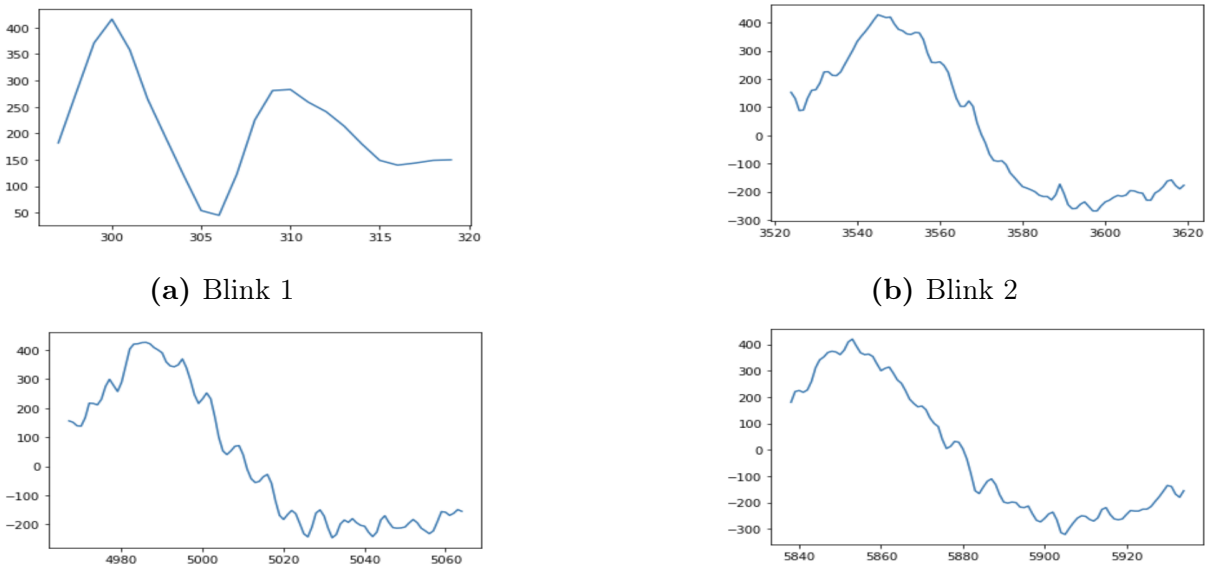


Figure 4.2: Blinks Structure



(a) Blink 1 **(b)** Blink 2

(c) Blink 3 **(d)** Blink 4

Figure 4.3: Independent Blinks Graph

```
def weighs(event):
    full_array = []
    new_array = []
    for i in event:
        for j in [1, 1, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8]:
            new_array.append(i*j)
        full_array.append(new_array)
        new_array = []
    return full_array
```

2. Upsample: The EEG dataset was measured in 250GHz, whereas the blinks were measured in 180 GHz. Hence, the blinks were upsampled prior to injection, to match GHz with EEG data.

```
def resample(event):
    resampled_event =
        scipy.signal.resample(event, round(len(event)/128*250))
    return resampled_event
```

3. Zero Mean: The mean value of the blink for each channel was deducted from the blinks value.

```
def zero_media(event):
    full_array = []
    new_array = []
    for i in event:
        media = i.mean()
        for j in i:
            new_array.append(j - media)
        full_array.append(new_array)
        new_array = []
    return full_array
```

A random function was used to decide how many seconds would elapse between blinks. For a normal eye, there are about fourteen blinks in a minute (Ernie Bowling et al., 2020), that implies a four to five seconds pause between blinks. A random function was also used to determine which of the four types of blinks would be inserted after any given pause. With these information, the function "blink_creator" was used to create a vector of the same length of the EEG, determining for each observation, weather it would be a pause between blinks (0), or one of the four blink types (1, 2, 3, 4), with its corresponding length. Inserting a blink, implies affecting as many rows as the length of the blink requires

```
def blink_creator():
    injection = pd.DataFrame(columns = ['blink_type'])
    while len(injection) <= len(signals):
        pause = pd.DataFrame(np.random.randint(low=0, high=1,
            size=(np.random.randint(low=1000,
            high = 1250))), columns=['blink_type'])
        injection = injection.append(pause, ignore_index = True)
        blink_number = np.random.randint(low=1, high=5)
        blink = pd.DataFrame(np.random.randint(low=blink_number,
            high = blink_number + 1,
            size = (len(globals()[(f"final_event_{blink_number}")]))),
            columns = ['blink_type'])
        injection = injection.append(blink, ignore_index = True)
```

```
    return (injection[0:len(signals)])
```

Finally, the pseudo-real dataset was generated according to the following steps: Iterating through each row of a modified version of the data, with an extra column detailing the blink which should be injected in each row. For each iteration we:

1. Generate an empty new array.
2. Assign the last columns value to the "blink_type" variable.
3. Define the array variable, containing the EEG data for this iteration (row).

If the corresponding blink type for the row is zero, the empty new array is assigned the EEG data for this iteration, contained in the "array" variable. If the blink type is greater than zero, the variable "size_blink" increases its value. If the value of the variable "size_blink" is smaller than the total size of the particular blink that is being inserted, the new array is assigned the value of the original EEG data plus the size of the blink at that corresponding point of the dataset. This procedure is repeated until the whole dataset has been iterated through.

```
def blink_injection(EEG_data_with_blinktype):
    new_data = []
    size_blink = -1
    for iteration in range(0, len(EEG_data_with_blinktype)):
        new_array = []
        blink_type = EEG_data_with_blinktype[iteration][-1]
        array = EEG_data_with_blinktype[iteration][0:8]
        if blink_type == 0:
            new_array = list(array)
            size_blink = -1
        else:
            size_blink += 1
            if size_blink >= len(globals()[(f"final_event_{blink_type}")]):
                size_blink = 0
            channel = 0
            for j in array:
```

```
                new_array.append(j +
                    globals()[(f"final_event_{blink_type}")][size_blink][channel])
                channel += 1
        new_data.append(new_array)
        new_array = []

    return np.array(new_data)
```

The following figure shows 5 seconds of a sample 8-channel EEG pseudo-real signal. Notice how on the 44th second, the graph differs from the original graph presented in Figure 4.1, because a blink has been inserted.

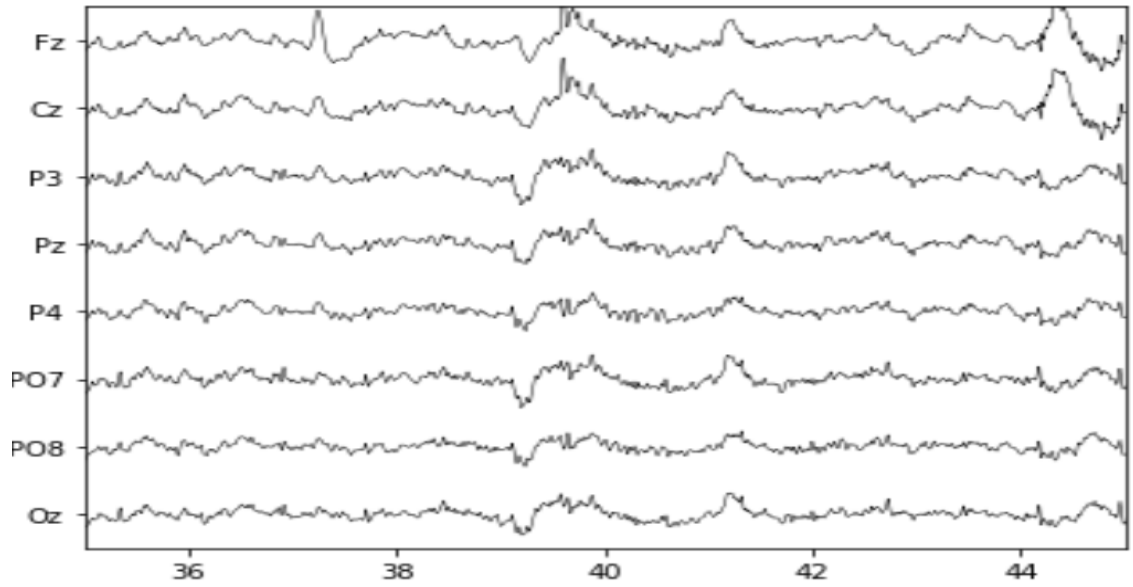


Figure 4.4: Sample 8-channel pseudo-real EEG signal

4.2 Applying ICA to Pseudo-real Dataset

Independent Component Analysis (ICA) is the most widespread used algorithm to decompose multi-channel data into independent components (Kanoga and Mitsukura, 2017). ICA was originally proposed to solve the blind source separation problem, to recover independent source signals, which have been mixed by an unknown matrix. (Jung et al., 2000).

The ICA algorithm is highly effective at performing source separation in domains where 1) the mixing medium is linear and propagation delays are negligible, 2) the time courses

of the sources are independent, and 3) the number of sources is the same as the number of sensors (Jung et al., 2000). Recent findings indicate that if ICA procedures are fine-tuned, ocular artifacts can be almost fully suppressed with relatively little distortion of genuine brain activity (Dimigen and Ehinger, 2021).

In the case of EEG signals, we assume that the multichannel EEG recordings are mixtures of underlying brain signals and artifacs. Because volume conduction is thought to be linear and instantaneous, assumption 1) is satisfied. Assumption 2) is also reasonable because the sources of eye and muscle activity, line noise, and cardiac signals are not generally time locked to the sources of EEG activity which is thought to reflect synaptic activity of cortical neurons. Assumption 3) is questionable, because we do not know the effective number of statistically independent signals contributing to the scalp EEG. However, numerical simulations have confirmed that the ICA algorithm can accurately identify the time courses of activation and the scalp topographies of relatively large and temporally independent sources from simulated scalp recordings, even in the presence of a large number of low-level and temporally independent source activities (Jung et al., 2000).

The input of ICA algorithm are different channels of EEG recordings. The outputs of ICA are the temporal independent components u and the estimated unmixing matrix W. The corrected EEG data can then be computed as

$$x^{'} = W^{-1} * u^{'}$$

where u' is the matrix u with rows containing artefact components set to zero, removing the contribution of the artefact component in the EEG data (Djuwari et al., 2006).

4.2.1 Experimentation

In our experiments, the input we give ICA is the pseudo-real dataset and the output we want to extract is the original EEG data, as it was prior to the injection of the blinks. In order to be able to gain easy visual identification of the ICA channel containing the blinks, an additional dataset was created for comparison, containing only the blinks. Figure 4.5 shows the first 5000 observations of the mono-channel blinks. We will be trying to identify this blinks within one of the channels of the pseudo-real dataset after applying

ICA, shown in Figure 4.6.

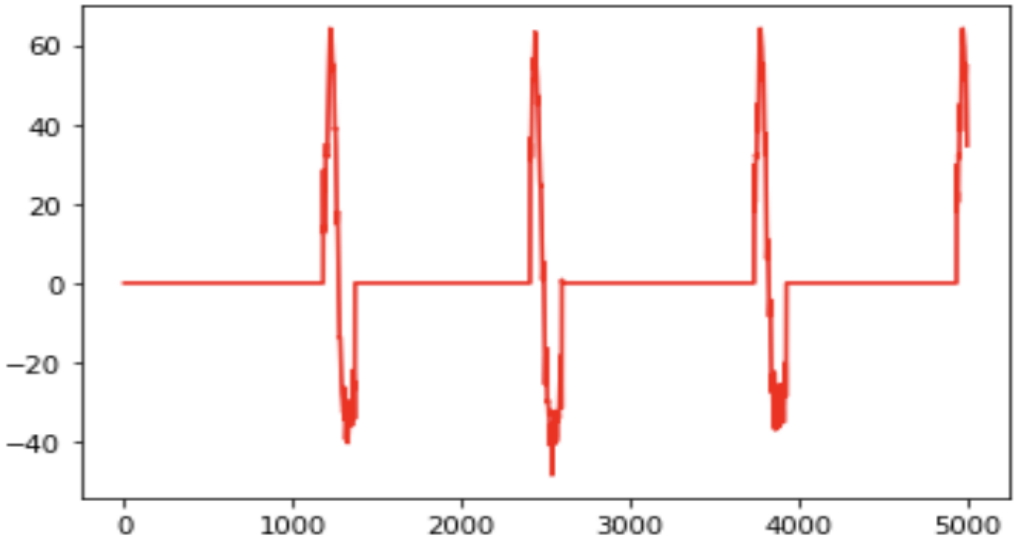


Figure 4.5: Sample of Mono-channel blinks

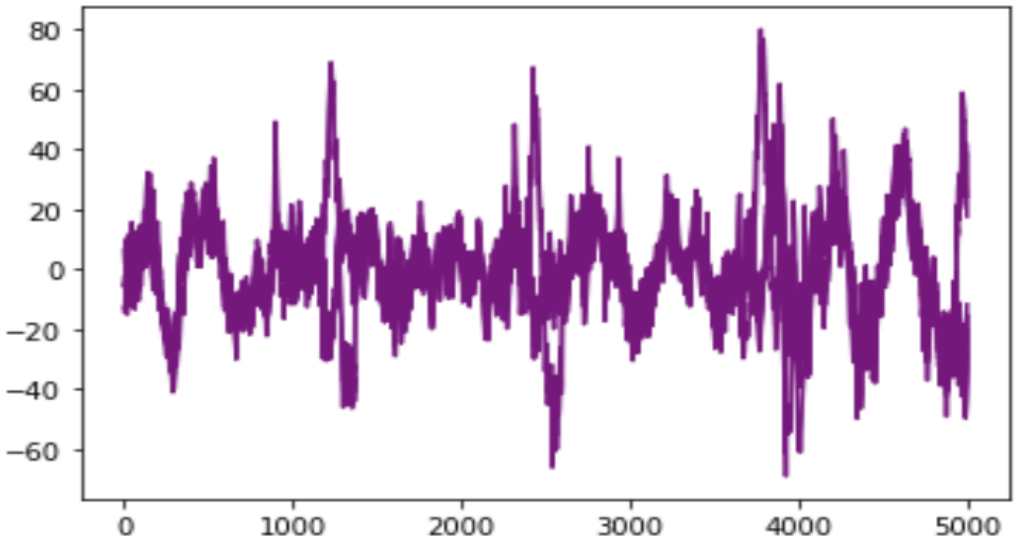


Figure 4.6: Sample of Pseudo-real EEG in single graph

5 Results

Several iterations of the experiment where made in order to find the right number of components for ICA to perform a correct separation of the artifacts. After iterating with 7, 6 and 5 components, through visual interpretation, one of the five components clearly contained the aforementioned blinks. After applying FastICA for 5 components, we plotted the first 5000 observations of each of the obtained components.

```
ica = FastICA(n_components = 5 , random_state = 19)
S_ = ica.fit_transform(modified_EEG)
A_ = ica.mixing_
output_transform = S_.T
print('ICA with 5 components')
for channel in range(0, 5):
    plt.plot(output_transform[channel,:][0:5000])
    plt.show()
```

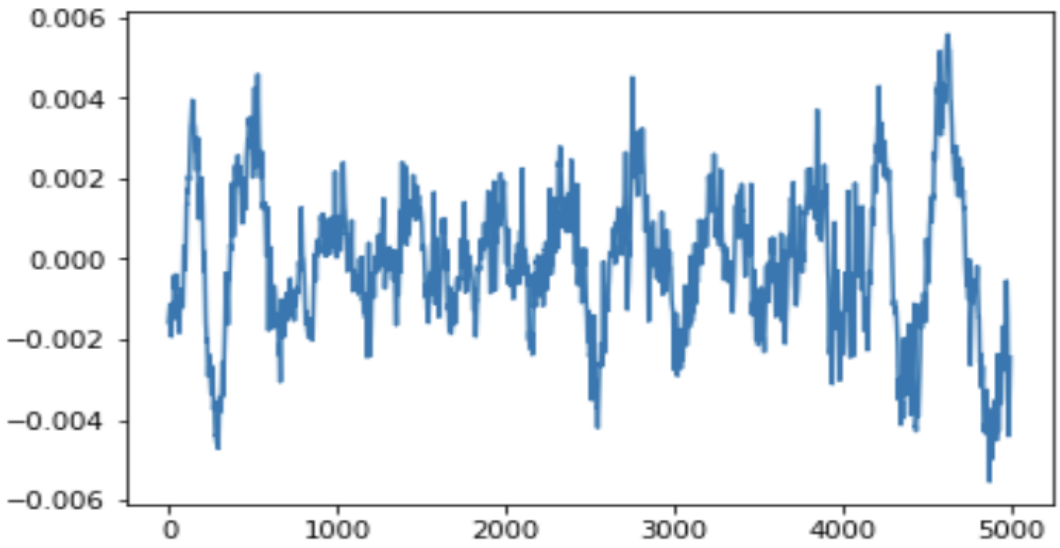


Figure 5.1: Sample of ICA Component 0

It is clearly identifiable by ocular inspection that ICA's component number 3 contains the isolated blinks.

After identifying the component which contains the artifacts, the procedure to suppress the blinks from the pseudo-real dataset is to nullify this component and rebuild the dataset

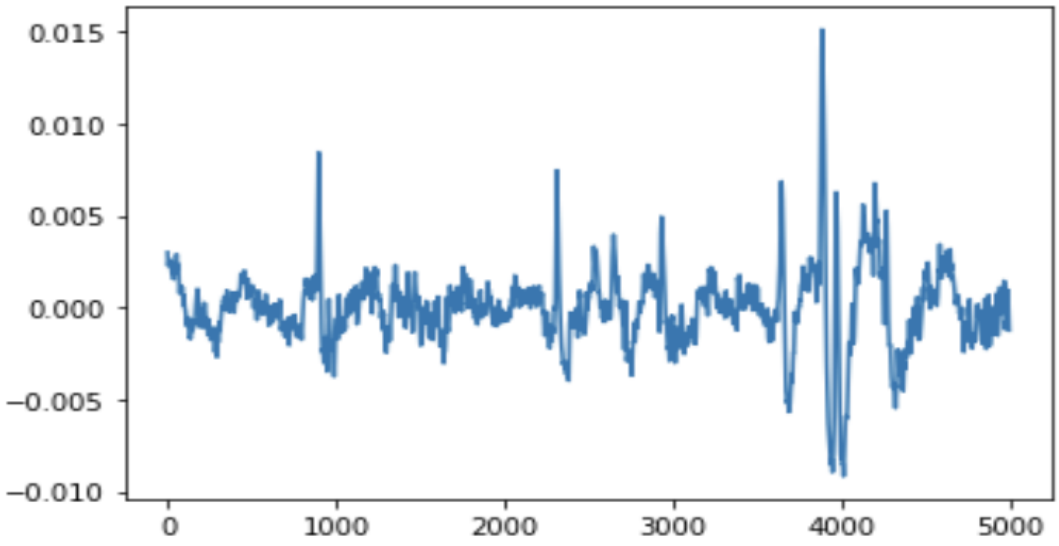


Figure 5.2: Sample of ICA Component 1

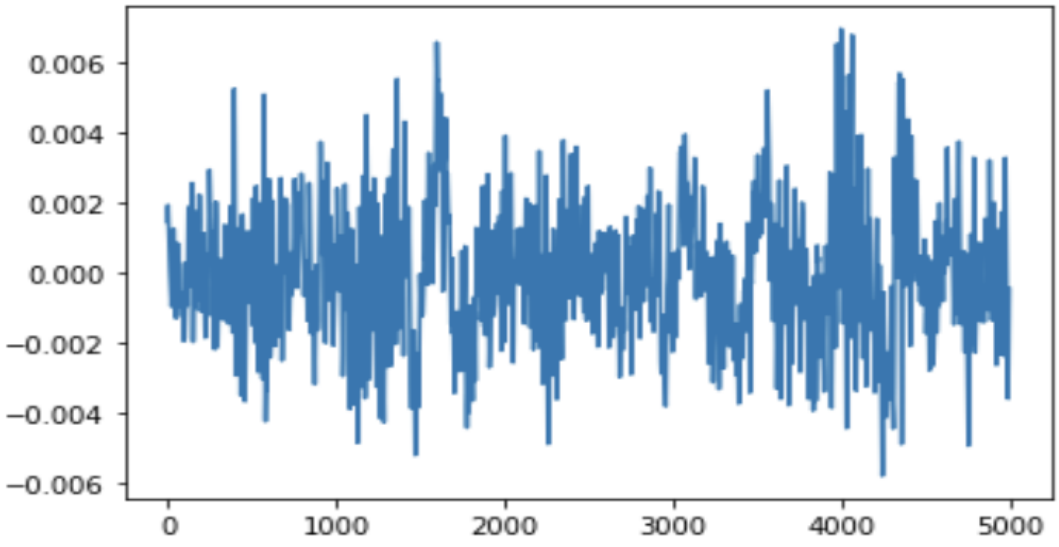


Figure 5.3: Sample of ICA Component 2

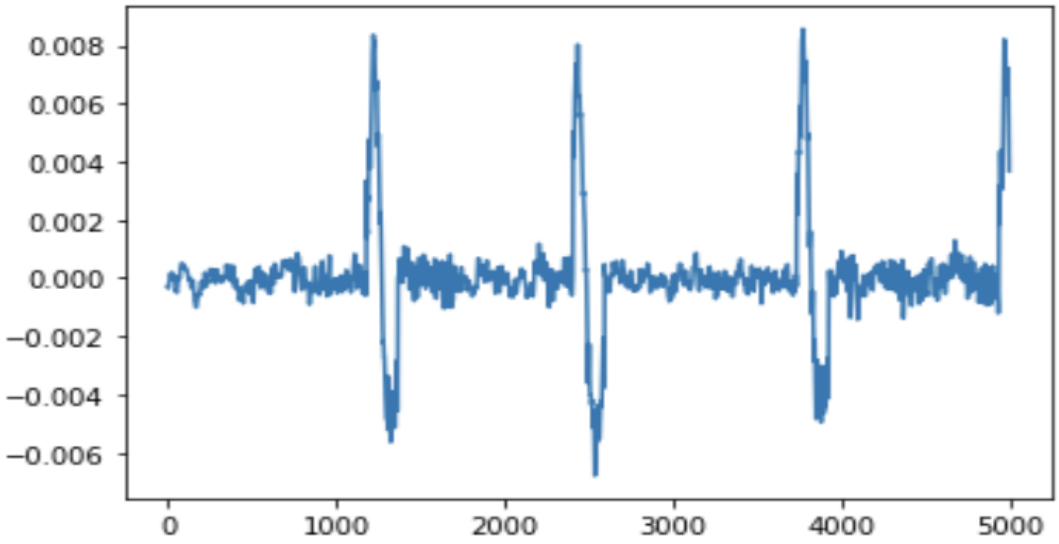


Figure 5.4: Sample of ICA Component 3

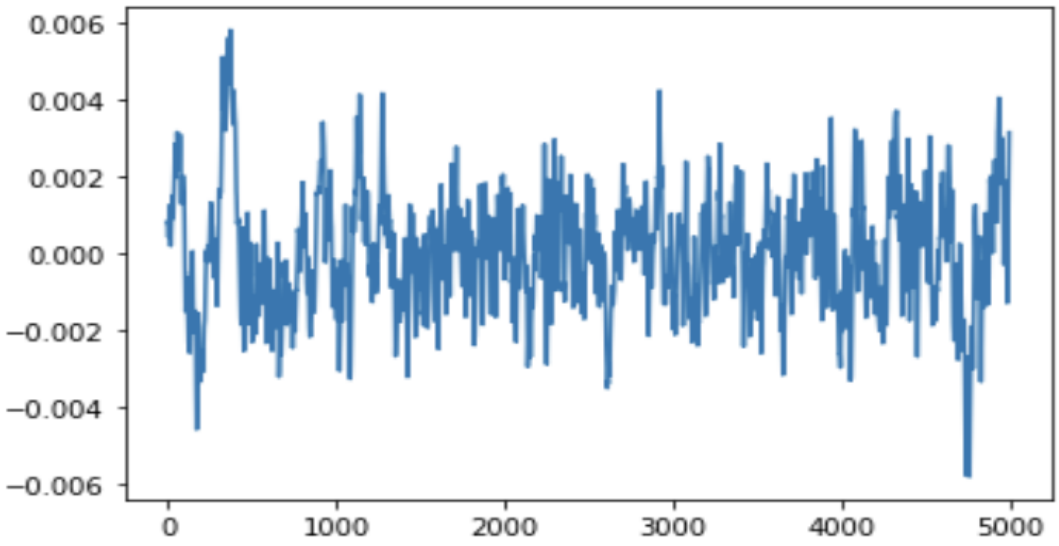


Figure 5.5: Sample of ICA Component 4

using the mixing matrix, which is an output of ICA.

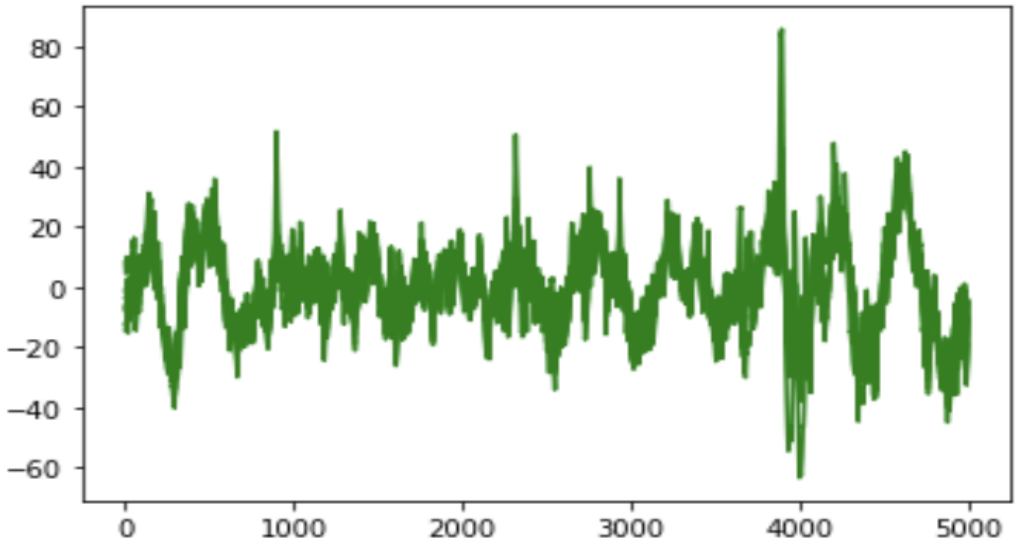


Figure 5.6: Sample of Pseudo Real EEG post ICA

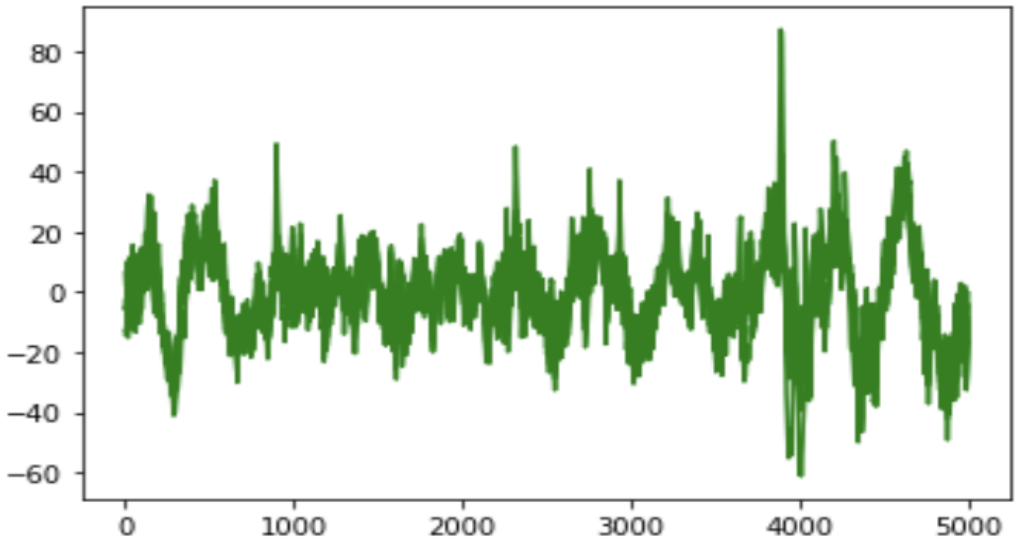


Figure 5.7: Sample of Original EEG

As proprocessed by several authors, ICA is not only enabled the separation of the artifacts, but also provided with the corresponding mixing matrix in order to rebuild the original dataset.

However, the original EEG data and the resulting EEG after blinks injection and ICA application data is not exactly identical, which is why using an assert to compare them both, will fail. In order to measure the quality of the estimated EEG, the Frobenoius

Distance between both matrices was calculated.

The result of this analysis allows us to prove that the distance between both matrices reduces to one third of its former value, after applying ICA.

Furthermore, we tested a separate dataset in which, at the time of injection, the blinks were multiplied by constant 1.000. This third distance between matrices was 1.000 times bigger than the distance between the original_EEg data and the modified_EEG data, proving that the greater the noise we inject, the bigger the distance to the original matrix. This also proves that applying ICA we obtain a much more reduce distance between matrices.

6 Conclusion

These work's main contribution was a methodological approach to the insertion of blinks in EEG data, in order to create a pseudo-real dataset, which in turn provided an experimental setup to test artifact removal algorithms. The topic of ocular waves' propagation throughout EEG channels was also covered. The optimal amount of ICA components was discovered through an iterative approach which included an observation step.

ICA method performs very well, but source identification is a very complex step which would require further investigation. An automatization of the amount of ICA components and the detection of which channel contains the artefact, would advance research a great deal.

This analysis still leaves plenty of questions unanswered and many aspects of electrophysiology which could outline a path for future work. A natural next step would be to test the proposed methodology in a real dataset, in order to eliminate the artifacts. There is also a very promising GPU implementation of the FastICA algorithm which allows for real-time artifact removal to be executed during signal measurement. (Benko and Juhasz, 2019) Some other questions this work has triggered are, how can we control if a EEG register has been manipulated? Would deep neural networks work for artifact identification? This are only few possibilities of whart future work would lead to.

7 Acknowledgements

En primer lugar, quiero agradecer al Dr. Rodrigo Ramele por su apoyo, por la idea para este trabajo y por el intercambio de opiniones continuo que me empujó avanzar y desafiarme para poder concluir el trabajo.

En segundo lugar, quiero agradecer al ITBA, por abrirme las puertas al apasionante mundo de la Ciencia de Datos, brindarme las herramientas para poder hacer de esto mi profesión y permitirme conocer a un grupo de profesionales afines de ambientes muy diversos, algunos de los cuales hoy considero amigos.

Por último, quiero agradecer a mi familia, sostén fundamental a lo largo de todos los logros de mi vida: Toto, mi marido; María y Diego, mis padres; Mer, Manu y Pe, mis hermanos; Teresa y Elvira, mis abuelas. Ustedes me impulsan a ser mejor persona y a dar siempre lo mejor de mi en todo lo que hago. Les estoy eternamente agradecida!

References

Benko, G. and Juhasz, Z. (2019). Gpu implementation of the fastica algorithm. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 196–199. IEEE.

Dimigen, O. and Ehinger, B. V. (2021). Regression-based analysis of combined eeg and eye-tracking data: Theory and applications. *Journal of Vision*, 21(1):3–3.

Djuwari, D., Kumar, D. K., and Palaniswami, M. (2006). Limitations of ica for artefact removal. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 4685–4688. IEEE.

Ernie Bowling, O. et al. (2020). Dry eye in the digital age. *Ocular Surface*, 12(4).

Islam, M. K., Rastegarnia, A., and Yang, Z. (2016). Methods for artifact detection and removal from scalp eeg: A review. *Neurophysiologie Clinique/Clinical Neurophysiology*, 46(4-5):287–305.

Jiang, X., Bian, G.-B., and Tian, Z. (2019). Removal of artifacts from eeg signals: a review. *Sensors*, 19(5):987.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., and Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178.

Kanoga, S. and Mitsukura, Y. (2017). Review of artifact rejection methods for electroencephalographic systems. *Electroencephalography*, page 69.

Katona, J., Ujbanyi, T., Sziladi, G., and Kovari, A. (2016). Speed control of festo robotino mobile robot using neurosky mindwave eeg headset based brain-computer interface. In *2016 7th IEEE international conference on cognitive infocommunications (CogInfoCom)*, pages 000251–000256. IEEE.

Lotte, F. (2008). *Study of electroencephalographic signal processing and classification techniques towards the use of brain-computer interfaces in virtual reality applications*. PhD thesis, INSA de Rennes.

Procházka, A., Mudrova, M., Vyšata, O., Hava, R., and Araujo, C. P. S. (2010). Multi-channel eeg signal segmentation and feature extraction. In *2010 IEEE 14th International Conference on Intelligent Engineering Systems*, pages 317–320. IEEE.

Ramele, R., Villar, A. J., and Santos, J. M. (2018a). Eeg waveform analysis of p300 erp with applications to brain computer interfaces. *Brain sciences*, 8(11):199.

Ramele, R., Villar, A. J., and Santos, J. M. (2018b). Eeg waveform analysis of p300 erp with applications to brain computer interfaces [source code]. https://doi.org/10.24433/CO.2231632.v2.

Ramele, R., Villar, A. J., and Santos, J. M. (2019). Histogram of gradient orientations of signal plots applied to p300 detection. *Frontiers in computational neuroscience*, 13.

Tobar, M. (2021). Experimental setup to test neurological artifact elimination techniques [source code]. https://github.com/maggietobar/artifact-elimination-ICA.

Urigüen, J. A. and Garcia-Zapirain, B. (2015). Eeg artifact removal—state-of-the-art and guidelines. *Journal of neural engineering*, 12(3):031001.

Van Drongelen, W. (2018). *Signal processing for neuroscientists*. Academic press.

Wolpaw, J. and Wolpaw, E. W. (2012). *Brain-computer interfaces: principles and practice*. OUP USA.