Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica, Economia e Finanza

RELAZIONE FINALE

OUTLIERS DETECTION IN CHILDREN'S ELECTROENCEPHALOGRAPHY SIGNAL USING SINGULAR VALUE DECOMPOSITION

Relatore: prof. Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Correlatore: prof. Simone Cutini
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Laureando: Silvia Palmieri

Matricola n. 1059197

Anno Accademico 2015/2016

Contents

Introduction

This essay is the response to a practical problem. A neuroscience research team from the Psychology Department for Development and Socialization of the University of Padua, structuring its experiment regarding electroencephalography (EEG) signals on 6-months-old children stated that each session had to be video recorded to be later revised in order to manually classify if the recorded trial had to be considered reliable or not, depending on whether the subject performed blinks or saccades or both during the trial. This procedure is later resulted to be time wasting and that it could potentially introduce some errors. This work uses the recorded signal to classify each trial, basing the deductions on Principal Components Analysis, the Receiving Operating Characteristics curve and the Cross-Validation method.

Acknowledgements

The research and coding work was supported by the Psychology Department for Development and Socialization of the University of Padua Professors Livio Finos and Simone Cutini.
I also wish to thank Professor Elena Natale who provided the dataset (twice).
I furthermore want to thank Professor Gianmarco Altoé for fruitful discussions on the present topic each wednesday morning for several month.
I would like to thank infants who participated in this study and their parents.

Chapter 1

Brief Overview on Electroencephalography

Electroencephalography (EEG) is the recording of intrinsic electrical activity in the brain, based on the propagation of electric impulses along a nerve fibre when the neuron fires. EEG is typically analyzed in frequency bands that correspond to different mental states, e.g. is the alpha-frequency (8-13 Hz) associated with a relaxed mental state. By recording small potential changes in the EEG signal immediately after the presentation of a sensory stimulus it is possible to record specific brain responses to specific sensory, cognitive and other mental events. This method is called Event-Related Potentials (ERPs) and is one of the classic methods for investigation of psychophysiological states and information processing.(**Bergen fMRI Group: ERP vs EEG**)

This first chapter is to be considered an introduction to this dissertation. Indeed in the first paragraph are enlisted the advantages and disadvantages of using Event-Related Potential techniques, related to the comparison between ERPs and other behavioral and physiological measures. In this last comparison are considered a few aspects separately: invasiveness of the measures, spatial and temporal resolution of the observed data and last but not least the costs related to each kind of

measure. The second paragraph is where we consider the electrical side of this technique: in here we explain how the electrodes register cortical signal through the braincase. The third and last parafraph makes a brief excursus through the most common recently used techniques of automatic removal of artifacts from EEG data.

1.1 Advantages and disadvantages of the ERP technique

1.1.1 Comparison with Behavioral Measures

When ERPs were first used to study issues in the domain of cognitive neuroscience, they were primarily used as an alternative to measurements of the speed and accuracy of motor responses in paradigms with discrete stimuli and responses. In this context, ERPs have two distinct advantages. First, an overt response reflects the output of a large number of individual cognitive processes, and variations in reaction time (RT) and accuracy are difficult to attribute to variations in a specific cognitive process. ERPs, in contrast, provide a continuous measure of processing between a stimulus and a response, making it possible to determine which stage or stages of processing are affected by a specific experimental manipulation. Thus, ERPs are very useful for determining which stage or stages of processing are influenced by a given experimental manipulation (for a detailed set of examples, see Luck, Woodman, and Vogel 2000). A second advantage of ERPs over behavioral measures is that they can provide an online measure of the processing of stimuli even when there is no behavioral response.

ERP recordings also have some disadvantages compared to behavioral measures. The most obvious disadvantage is that the functional significance of an ERP component is virtually never as clear as the functional significance of a behavioral response. In most cases, we do not

know the specific biophysical events that underlie the production of a given ERP response or the consequences of those events for information processing. In contrast, when a computer records a button-press response, we have a much clearer understanding of what that signal means.

A second disadvantage of the ERP technique is that ERPs are so small that it usually requires a large number of trials to measure them accurately. In most behavioral experiments, a reaction time difference can be observed with only about twenty to thirty trials per subject in each condition, whereas ERP effects often require fifty, a hundred, or even a thousand trials per subject in each condition. This places significant limitations on the types of questions that ERP recordings can realistically answer.

1.1.2 Comparison with Other Physiological Measures

The ERP technique (along with its magnetic counterpart, the event-related magnetic field, or ERMF, technique) can be compared with several other physiological recording techniques along four major dimensions: invasiveness, spatial resolution, temporal resolution, and cost. The other classes of techniques considered are microelectrode measures (single-unit, multi-unit, and local field potential recordings) and hemodynamic measures (PET and fMRI).

Invasiveness Microelectrode measures require inserting an electrode into the brain and are therefore limited to nonhuman species (or, in rare cases, human neurosurgery patients). The obvious disadvantage of primate recordings is that human brains are different from primate brains. The less obvious disadvantage is that a monkey typically requires months of training to be able to perform a task that

a human can learn in five minutes, and once a monkey is trained, it usually spends months performing the tasks while recordings are made. Thus, monkeys are often highly overtrained and probably perform tasks in a manner that is different from the prototypical naïve college sophomore. This can make it difficult to relate monkey results to the large corpus of human cognitive experiments. PET experiments are also somewhat problematic in terms of invasiveness. To avoid exposing subjects to excessive levels of radiation, each subject can be tested in only a small number of conditions. In contrast, there is no fundamental restriction on the amount of ERP or fMRI data that can be collected from a single subject.

Spatial and temporal resolution Many authors have noted that electromagnetic measures and hemodynamic measures have complementary patterns of spatial and temporal resolution, with high temporal resolution and poor spatial resolution for electromagnetic measures and poor temporal resolution and high spatial resolution for hemodynamic measures. ERPs have a temporal resolution of 1 ms or better under optimal conditions, whereas hemodynamic measures are limited to a resolution of several seconds by the sluggish nature of the hemodynamic response. This is over a thousandfold difference, and it means that ERPs can easily address some questions that PET and fMRI cannot hope to address. However, hemodynamic measures have a spatial resolution in the millimeter range, which electromagnetic measures cannot match (except, perhaps, under certain unusual conditions). In fact, the spatial resolution of the ERP technique is fundamentally undefined, because there are infinitely many internal ERP generator configurations that can explain a given pattern of ERP data. Unlike PET and fMRI, it is not currently possible to specify a margin of error for an ERP localization claim (for the typical case, in which several sources are simul-

taneously active). That is, with current techniques, it is impossible to know whether a given localization estimate is within some specific number of millimeters from the actual generator source. It may someday be possible to definitively localize ERPs, but at present the spatial resolution of the ERP technique is simply undefined. The fact that ERPs are not easily localized has a consequence that is not often noted. Specifically, the voltage recorded at any given moment from a single electrode reflects the summed contributions from many different ERP generator sources, each of which reflects a different neurocognitive process. This makes it extremely difficult to isolate a single ERP component from the overall ERP waveform. This is probably the single greatest shortcoming of the ERP technique, because if you can't isolate an ERP component with confidence, it is usually difficult to draw strong conclusions.

Cost ERPs are much less expensive than the other techniques listed so far. It is possible to equip a good ERP lab for less than US $50,000 , and the disposable supplies required to test a single subject are very inexpensive (US $1–3). A graduate student or an advanced undergraduate can easily carry out the actual recordings, and the costs related to storing and analyzing the data are minimal. These costs have dropped a great deal over the past twenty years, largely due to the decreased cost of computing equipment. FMRI is fairly expensive, the major costs being personnel and amortization of the machine. One session typically costs US $300–800. PET is exorbitantly expensive, primarily due to the need for radioactive isotopes with short half-lives and medical personnel. Single-unit recordings are also fairly expensive due to the per diem costs of maintaining the monkeys, the cost of the surgical and animal care facilities, an physiological data from awake, behaving monkeys.

1.2 Electricity

1.2.1 Volume Conduction

When a dipole is present in a conductive medium such as the brain, current is conducted throughout that medium until it reaches the surface. This is called volume conduction and is illustrated in part C in figure 1.1. The voltage that will be present at any given point on the surface of the scalp will depend on the position and orientation of the generator dipole and also on the resistance and shape of the various components of the head (most notably the brain, the skull, and the scalp; the eye holes also have an influence, especially for ERP activity generated in prefrontal cortex). Electricity does not just run directly between the two poles of a dipole in a conductive medium, but instead spreads out through the conductor. Consequently, ERPs spread out as they travel through the brain. In addition, because electricity tends to follow the path of least resistance, ERPs tend to spread laterally when they encounter the high resistance of the skull. Together, these two factors greatly blur the surface distribution of voltage, and an ERP generated in one part of the brain can lead to substantial voltages at quite distant parts of the scalp. There are algorithms that can reduce this blurring, either by estimating the flow of current or by deblurring the voltage distribution to estimate the voltage distribution that is present on the brain's surface (Gevins et al. 1999; Pernier, Perrin, and Bertrand 1988). These algorithms can be very useful, of although you should remember that they only eliminate one source of blurring (the skull) and do not indicate the actual generator location of the ERPs. Another important point is that electricity travels at nearly the speed of light. For all practical purposes, the voltages recorded at the scalp reflect what is happening in the brain at the same moment in time.

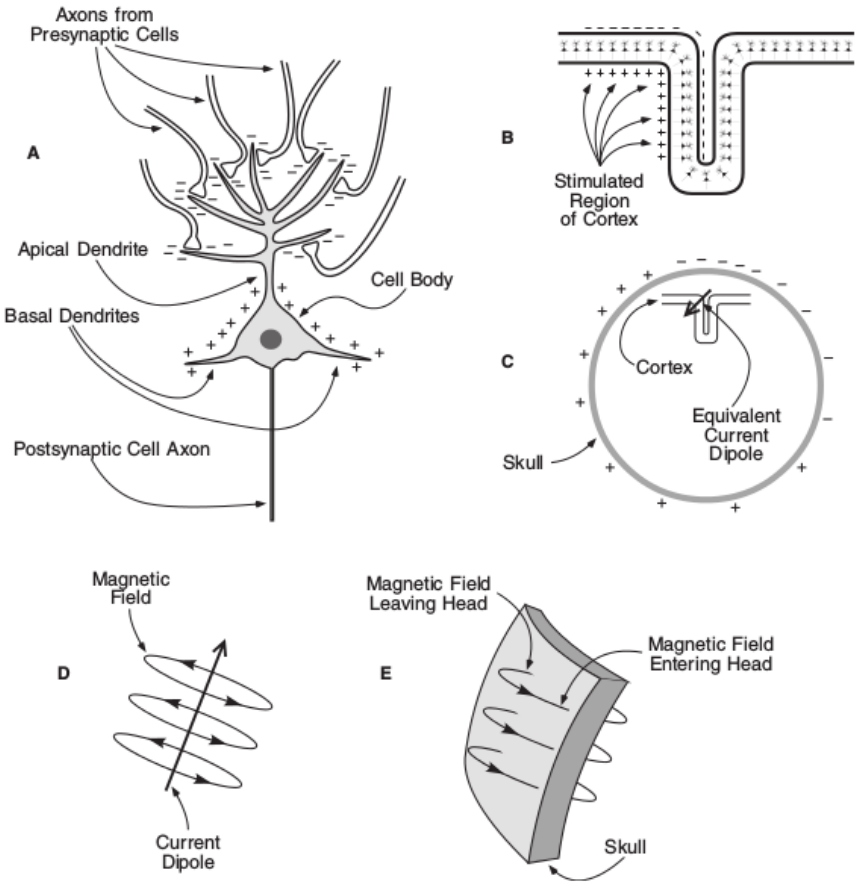

Figure 1.1: (C) The summated dipoles from the individual neurons can be approximated by a single equivalent current dipole, shown here as an arrow. The position and orientation of this dipole determine the distribution of positive and negative voltages recorded at the surface of the head. (D) Example of a current dipole with a magnetic field traveling around it. (E) Example of the magnetic field generated by a dipole that lies just inside the surface of the skull. If the dipole is roughly parallel to the surface, the magnetic field can be recorded as it leaves and enters the head; no field can be recorded if the dipole is oriented radially Luck and Girelli 1998.

1.2.2 Magnetic Fields

The blurring of voltage caused by the high resistance of the skull can be largely circumvented by recording magnetic fields instead of electrical potentials. As part D in figure 1.1 illustrates, an electrical dipole is always surrounded by a magnetic field, and these fields summate in the same manner as voltages. Thus, whenever an ERP is generated, a magnetic field is also generated, running around the ERP dipole. Moreover, the skull is transparent to magnetism, and the magnetic fields are not blurred by the skull, leading to much greater spatial resolution than is possible with electrical potentials. The magnetic equivalent of the EEG is called the magnetoencephalogram (MEG), and the magnetic equivalent of an ERP is an event-related magnetic field (ERMF). As in part E in figure 1.1 illustrates, a dipole that is perpendicular to the surface of the scalp will be accompanied by a magnetic field that leaves the head on one side of the dipole and enters back again on the other side. If you place a highly sensitive probe called a SQUID (Super-Conducting Quantum Interference Device) next to the head, it is possible to measure the magnetic field as it leaves and reenters the head. Because magnetic fields are not as smeared out as electrical potentials, they can provide more precise localization. However, the combination of ERP and ERMF recordings provides even better localization than ERMF recordings alone. Unfortunately, magnetic recordings are very expensive because supercooling is expensive and because an expensive magnetically shielded recording chamber is necessary to attenuate the Earth's relatively large magnetic field.

1.3 Automatic removal of artifacts from EEG data: methods recently adopted

Eye movements and blink contamination are pervasive problems in event-related potential (ERP) research. The electric potentials created during saccades and blinks can be orders of magnitude larger than the electroencephalogram (EEG) and can propagate across much of the scalp, masking and distorting brain signals. Automated methods are preferable because they eliminate the subjectivity associated with nonautomated correction, are significantly more time and resource efficient, and can make it practical to use such applications during online EEG monitoring for clinical and other uses. Blink artifacts are attributed to alterations in conductance arising from contact of the eyelid with the cornea (Overton and Shagass 1969). An eyeblink can last from 200 to 400ms and can have an electrical magnitude more than 10 times that of cortical signals. Eye movements called saccades generate another type of electric signal. The cornea of the eye is positively charged relative to the retina, which amounts to having a steady retino-corneal charge of between 0.4 and 1.0 mV that approximates a dipole in both eyes. As the retino-corneal axis rotates during eye movements, the orientation of this dipole in three-dimensional space also rotates, resulting in changes in electric potential. The signals due to eye movement propagate mainly through the shunt pathway provided by the eye sockets. Decomposition methods identify individual signal components in EEG data without reference to head or source propagation models, so they are not subject to the general nonuniqueness of source localization solutions and the poor spatial resolution afforded by EEG data, particularly if sources are closely spaced (e.g., Achim, Richer, and Saint-Hilaire 1991), or to the fact that the distribution of the tissues in the head must be known precisely to model this propagation accurately throughout the head. Decomposition methods identify individual signal components in

EEG data without reference to head or source propagation models, so they are not subject to the above constraints. EEG component separation procedures using principal components analysis (PCA) and its counterpart, singular value decomposition (SVD) were proposed by Berg and Scherg 1991 and Sadasivan and Dutt 1996, among others.

By definition, PCA and SVD assume that the data components are algebraically orthogonal, a condition that, in general, is hard to satisfy. The actual algebraic relationship between source vectors is a function of each source location, orientation, and to some degree the head conductance parameters.

Orientation of certain ocular generators (e.g., blinks) may even be nearly aligned with orientation of frontal EEG generators. A more advanced method by Berg and Scherg 1994 that combines source modeling, PCA, and artifact averaging provides an improvement on the individual techniques above but requires a substantial amount of calibration data and prior modeling of artifact production and event-related activities. More recently, Vigário 1997, Jung et al. 2000, and a number of other researchers have turned to indipendent component analysis (ICA) to find components of EEG/EOG[1] data. ICA aims to project (decompose) data onto statistically independent components utilizing higher-order statistical measures, beyond the second-order statistics used by PCA. These methods represent a subclass of the general group of blind source separation (BSS) algorithms.[2]

In this essay the data are processed through a new combination of those methods, in order to avoid the manual identification of the artifacts, as the current procedure applied so far to all EEG data suggests.

[1] Electrooculogram: A measurement of the electrical activity associated with eye movements as recorded with the placement of small metal discs called electrodes applied to the skin near the eyes. It is useful for monitoring eyeball movement in REM and non-REM sleep.

[2] Joyce, Gorodnitsky, and Kutas 2004

Chapter 2

The Experiment

2.1 The Data

2.1.1 Participants

Twenty-four healthy, full-term 6-months-old infants (10 females, mean age= 6 months and 7 days, range from 183 to 224 days) participated in the study. All of them were considered for behavioral analysis, but only 13 out of the 24 infants (6 females, mean age= 6 months and 7 days, range from 186 to 213 days) could be also considered for ERP analysis (see inclusion criteria below). Three infants were additionally tested, but not included in the final sample of participants because of fussiness or excessive movement artifacts, resulting in no reliable performance. Infants were tested if awake and in an alert state, and after parents gave their informed consent. The experimental protocol was approved by the Ethical Committee of the University of Padova.

2.1.2 Stimuli, apparatus and procedure

Testing took place in a dimly illuminated room. Infants were seated on a parent's lap approximately 60 *cm* from a 24 *inch* screen used for stimulus presentation. As shown in figure 2.1, each trial began with an

animated fixation point displayed at the center of the screen. As soon as the infant looked at it, this was replaced by the visual-spatial cue, i.e. a walker PLD randomly facing either to the right or to the left.[1] The cue was shown for $1,200\ ms$ and, after a variable delay (range from 300 to 500 ms), the target stimulus was displayed for 200 ms. The latter consisted of the static image of a colored ball ($1,75\ cm$ in radius) and could be randomly presented at a peripheral location (10 degrees of visual angle from the center of the screen) either congruent (i.e., valid coded by "VAL") or incongruent (i.e. invalid coded by "INV") with the walking direction. Structuring trials this way, we can get four kind of conditions to encode infants' gaze behavior:

- INV - (L/R): the cue and the target aren't congruent, if the cue is facing to the right, the target is shown in the left side of the screen and viceversa.

- VAL - (L/R): the cue and the target are congruent, so that the target is shown in the same side the cue was facing to in the previous phase of the trial.

Stimuli were presented in blocks of 16 trials, eight valid (four with left- and four with right-sided targets) and eight invalid (four with left- and four with right-sided targets). The animated fixation point varied on each trial. Also the target stimulus varied, being randomly selected among four possible types. In order to obtain as many trials as possible from each infant, there was no restriction in number of blocks or trials shown, i.e., they were played as long as the infant was not fussy. Specifically, the experimental session was terminated when infants looked away from the screen during five consecutive trials. On average, about 40 trials (range from 19 to 57) were presented to each infant, with no difference between the number of valid ($N = 19.8$) and invalid ($N = 20.2$) trials,

[1] Aaen-Stockdale et al. 2008, Troje 2008

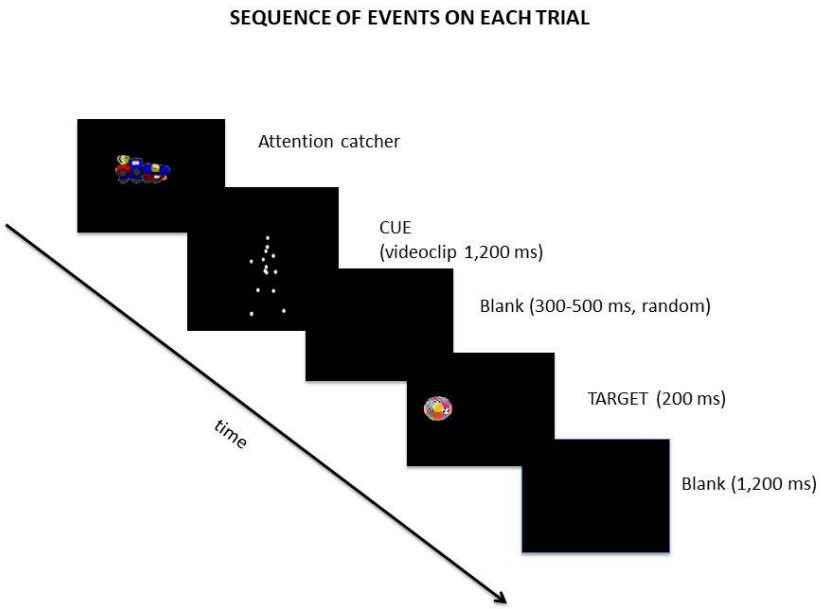

Figure 2.1: Here is shown the stimuli structure

$t(23) = -1.3$, $p = .21$. This also applied to the subgroup of 13 infants included in the ERP analysis, who on average saw about 42 trials (range from 26 to 53), with no difference between number of valid ($N = 20.5$) and invalid ($N = 21.4$) trials, $t(12) = -2$, $p = .07$. The sequence and timing of the stimuli were controlled by the computer, using E-Prime 2.0.

Chapter 3

Theoretical Bases

This chapter introduces the arguments that are the bases of this work. In paragraph 1 are presented eigenvalues and eigenvectors with a base explanation. These elements are useful to apply Singular Value Decomposition and Principal Components Analysis, presented in paragraph 2: these techniques are usually used to manipulate data matrices. Paragraph 3 provides the Receiving Operating Characteristics (ROC) curve definition with the reason for we consider it in this work.

3.1 Eigenvalues and Eigenvectors

For each squared matrix A, might be found a scalar λ and a vector x such as

$$Ax = \lambda x \tag{3.1}$$

where λ is called the eigenvalue of A and x is the eigenvector of A associated to λ. To solve the equation (3.1) we write

$$(A - \lambda I)x = 0 \tag{3.2}$$

If $|A-\lambda I| \neq 0$, $(A-\lambda I)$ is invertible and $x = 0$ is the unique solution. To exclude trivial solutions, then we impose $|A-\lambda I| = 0$ in order to obtain λ values so that they can be replaced in (3.1) and in (3.2) to obtain the

correspondent values of x. To have a solution for (3.2), it is necessary that the $A - \lambda I$ columns are linearly indipendent. In conclusion, $A - \lambda I$ has to be singular (the determinant is null) so that exist $x \neq 0$ that is the solution to (3.1). The equation $|A - \lambda I| = 0$ is called the characteristic equation. If A is a $p \times p$ matrix, the characteristic equation has p solutions; i.e. A has p eigenvalues $\lambda_1, \ldots, \lambda_p$. The λ_i values are not necessarily different from 0. The eigenvectors $x_1, \ldots, x_p$ associated to $\lambda_1, \ldots, \lambda_p$ are defined by (3.2). It is important to note that if we multiply both members of (3.1) for a k scalar (and using the commutative property), we obtain

$$(A - \lambda I)kx = k0 = 0$$

So, if x is an eigenvector of A, also kx is an eigenvector. Generally, each eigenvector is unique up to scalar values $k \neq 0$. Note that in each case that the direction indicated by the x vector is unique (multiplying for k, the coordinates ratio remain constant) and so the solution is essentially unique. For this reason, it's common to define x a unit norm vector, i.e. $x'x = 1$.

3.1.1 Some interesting results

- Despite the eigenvalues of a matrix A are not necessarily being all real, this is guaranted if A is symmetric and real.
- Let A be a $p \times p$ matrix and G a $p \times p$ matrix with rank $r(G) = p$. A and $G^{-1}AG$ have the same eigenvalues.
- An A matrix with $r(A) < p$ has at least a null eigenvalue.
- An idempotent matrix has only eigenvalues equal to 1 or 0
- – If M is semi positive definite, then $\lambda_j \geq 0 \ \forall j = 1, \ldots, p$
  – If M is positive definite, then $\lambda_j > 0 \ \forall j = 1, \ldots, p$

- Eigenvectors associated to distinct eigenvalues are linearly indipendent (if $\lambda_i \neq \lambda_j$, $x_i' x_j = 0$)

3.2 Singular Value Decomposition and Principal Component Analysis

While processing the data, we used two different techniques:

- Singular Value Decomposition (SVD): this is a useful instrument to find the principal components from the data matrix. In order to analize in the best way the supplied data, also loadings and scores are saved as results from this calculation. Thus, it is possible to analize the loadings following the main purpose of this essay.

- Principal Components Analysis (PCA): this method is similar to the previous one but it should be inserted in this context because the results from this analysis can be a very helpful instrument to provide some graphical representation of the data using the first and the second principal components to define an appropriate biplot.

3.2.1 Singular Value Decomposition (SVD)

Singular Value Decomposition Theorem. Let M be a real or complex $n \times p$ matrix. Then

$$M = UD\Gamma'$$

where U and Γ are real or complex orthonormal $n \times n$ and $p \times p$ matrices respectively so that $U'U = UU' = I_n$ and $\Gamma'\Gamma = \Gamma\Gamma' = I_p$, D is a rectangular diagonal matrix with $k = \min(n, p)$ non-negative values on the diagonal, also calld singular values. It is also known that this

decomposition can be written as a sum between matrices:

$$M = UD\Gamma' = \sum_{j=1}^{k} d_j u_j \gamma_j'$$

with u_j and γ_j vectors respectively from the U and Γ matrices and d_j diagonal elements from the D matrix.

Lemma. Let C, a $m \times n$ matrix, be the maximum of

$$\max_{u,v} \frac{u'Cv}{\sqrt{(u'u)(v'v)}} = d_1$$

where d_1 is the first singular value from the SVD and u and v are the first vector from the U and Γ matrices respectively. Moreover $d_2 \geq \cdots \geq d_r$ (the following singular values from C) are the are the following maximum obtained by pairs of (u, v) under the constraint of being orthogonal vectors with the previous ones.

3.2.2 Some interesting elements

Theorem (of spectral decomposition). Let M be a real symmetric $p \times p$ matrix. So

$$M = \Gamma\Lambda\Gamma' = [\gamma_1, \ldots, \gamma_p] \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ddots & \ldots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix} \begin{bmatrix} \gamma_1' \\ \vdots \\ \gamma_p' \end{bmatrix} = \sum_{j=1}^{p} \lambda_j \gamma_j \gamma_j'$$

where $\Lambda = diag\{\lambda_j\}$; ($\lambda_j$, $j = 1, \ldots, p$ are the eigenvalues of M) and $\Gamma = [\gamma_1, \ldots, \gamma_p]$ is an orthogonal matrix whose columns γ_j are the unit norm eigenvectors related to the eigenvalues of the M matrix.

Here are presented some connection between the SVD and the previous theorem:

- Those two theorems provide the same results *if and only if* M is a squared, symmetric and positive semi-definite matrix. In particular, this implies $U = D$.

- Lets compare the SVD results applied to the M matrix to those obtained through the spectral decomposition theorem applied to $A = M'M$:

  –

$$\begin{aligned} M'M &= \Gamma D' U' U D \Gamma \\ &= \Gamma D' I D \Gamma \\ &= \Gamma \Lambda \Gamma. \end{aligned} \tag{3.3}$$

  This result has already been used to proof the previous Lemma consequently:

  – $DD = D^2 = \Lambda$ i.e. the diagonal elements of D (the singular values of the M matrix) are equal to $\lambda_j^{1/2}$, $\forall j = 1, \ldots, p$ (the eigenvalues of the M matrix)

  – Γ is the same unit norm eigenvectors matrix in the spectral decomposition theorem.

- The SVD function, using R, receives as input X as the matrix whose SVD decomposition is to be computed (X can be a numeric or complex matrix, logical matrices are coerced to numeric); nu is the number of left singular vectors to be computed(this must between 0 and $n = nrow(X)$) and nv is the number of right singular vectors to be computed (that must be between 0 and $p = ncol(X)$). The returned value is a list with the following components:

  – d: a vector containing the singular values of X (i.e. the eigenvalues of X), of length $min(n, p)$

  – u: present if $nu > 0$. It has dimensions $n \times nu$ and is UD, i.e. the scores of the principal components

  – v: a matrix whose columns contain the eigenvectors or loadings of X, present if $nv > 0$. It has dimensions $p \times nv$.

3.2.3 Principal Components Analysis (PCA)

[1] The usual procedure is as follows. Suppose X is a p-dimensional vector with covariance matrix Σ or, as in this case, with its sample counterpart S. The first principal component is a linear combination $g^T X$, for some vector g which satisfies $g^T g = 1$, which is chosen to maximize the variance among all linear combinations. In other words, we find g to maximize $g^T S g$ subject to $g^T g = 1$. If the solution is $g = g_1$ then $g_1^T X$ is called the *first principal component of* X. If we want to go beyond the the first PC, we then repeat the optimisation in the space orthogonal to g_1: find g_2 to maximize $g_2^T S g_2$ such that $g_2^T g_2 = 1$, $g_2^T g_1 = 0$. Then $g_2^T X$ is the *second principal component.* The process continues iteratively: suppose $g_1, \ldots, g_{k-1}$ are given, for some $k \leq p$, then find g_k to maximize $g_k^T S g_k$ subject to $g_k^T g_k = 1$, $g_k^T g_j = 0$ for $j = 1, \ldots, k-1$. Then $g_k^T X$ is the *k'th principal component.* In the origin we could go on to find all p PCs, though in practice it is unusual to stop after selecting enough PCs to capture most of the variability in the data.

Now we show how to calculate $g_1, g_2, \ldots$. Let G be an orthogonal matrix such that $G^T S G = D$, where D is a diagonal matrix with diagonal entries ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. We can always find such representation, because S is symmetric and non-negative definite. Let g_k be the k'th column of G. So g_k is a norm-one eigenvector of S with λ_k eigenvalue.

Claim: These g_k's are the solutions of the optimization problem described above. Moreover, the principal components $g_1^T X, g_2^T X, \ldots, g_p^T X$ are incorrelated, and the sum of their variances is the sum of the variances of the individual components of X.

Proof of Claim. The eigenvectors $\{g_k, 1 \leq k \leq p\}$ form a complete orthonormal basis in $\mathbb{R}^p$, so far any $g \in \mathbb{R}^p$, there exist consistants

[1] Smith 1999

$c_1, \ldots, c_p$ such that $g = S_k c_k g_k$. Then

$$g^T g = \sum_{k=1}^{p} \sum_{l=1}^{p} c_k c_l g_k^T g_l = \sum_{k=1}^{p} c_k^2$$

$$g^T S g = \sum_{k=1}^{p} \sum_{l=1}^{p} c_k c_l g_k^T S g_l$$

$$= \sum_{k=1}^{p} \sum_{l=1}^{p} c_k c_l g_k^T g_l \lambda_l = \sum_{k=1}^{p} \lambda_k c_k^2$$

Since the $\{\lambda_k\}$ are ordered, $g^T S g \leq \lambda_1 S_k c_k^2 = \lambda_1$, with equality if $c_2 = \cdots = c_p = 0$. This proves that g_1 has the property of maximizing $g^T S g$ subject to $g^T g = 1$. It is the unique solution, up to changes of sign, if $\lambda_1 > \lambda_2$, but we have not excluded the possibility that $\lambda_1 = \lambda_2$ in which case the solution is not unique.

To get the second PC, we focus on $g = S_k c_k g_k$ which are orthogonal to g_1, in other words, for which $c_1 = 0$. But then, an extension of the same reasoning shows that $g^T S g \leq \lambda_2 S_k c_k^2 = \lambda_2$, with equality if $c_3 = \ldots = c_p = 0$. This proves that g_2 solves the equation for the second PC, and is unique if $\lambda_1 > \lambda_2 > \lambda_3$. We proceed in similar fashion to derive the third, fourth and subsequent PCs.

The PCs are orthogonal, because for $k \neq l$, $g_k^T S g_l = \lambda_l g_k^T g_l = 0$ by orthogonality of the g_k's, and the k'th PC has variance $g_k^T S g_k = \lambda_l g_k^T g_k = \lambda_k$. Finally the sum of the variances of the PCs is $S_k \lambda_k = tr(D) = tr(S)$, which is the sum of the variances of the individual components of X. □

One difficulty associated with this is that the problem is not scale invariant – if the original data were lengths measured in inches and weights measured in pounds, and if we then changed the scales of measurement to centimetres and kilograms, the PCs and the corresponding λ_k's would change. A way to avoid this difficulty is to rescale the problem prior to computing the PCs, so that each component of X had either population

variance or sample variance equal to 1. This is, of course, equivalent to replacing S by the corresponding correlation matrix. Thus, an alternative way of proceding is via the correlation matrix instead of the covariance matrix.

3.3 Receiving Operating Characteristics curve (ROC curve)

Classification is a statistical method used to build predicative models to separate and classify new data points. The populations we want to distinguish between are invalid trials manually defined and trials defined invalid after the analysis. Considering the variable `u`, called feature, we want to measure to assess whether a trial is effectively valid or not. Finally, classification will make a decision function, as in this case, using a logit regression based analysis, or similarly using a linear or quadratic discriminant analysis or classification and regression trees. Last but not least, it is necessary an evaluation of the goodness and to do so, in this case it is dued to consider the ROC curve and the cross-validation procedure.

The ROC curve is a valid instrument to measure, by a graphic representation, the accuracy of the classification procedure. It's based on the relative error of classification related to a set of observations that can or can not be the same set used to form the procedure. The ROC curve is constitute in order to classify binary response.

The focus is on binary classification, we consider to classify in G_2 if $P(G_2|\mathbf{x}) > c$. In the beginning, we consider the following confusing matrix, where c is fixed in the $(0, 1)$ interval, and the *threshold*,that in R represents each value we can use to recalculate different confusing matrices related to different values of sensitivity and specificity.

Note that, the growth of c causes the decrease of $n_{12}(c)$, $n_{22}(c)$ and

		Expected Class		
		$\hat{G}_1$	$\hat{G}_2$	
Real Class	G_1	$n_{11}(c)$	$n_{12}(c)$	$n_{1.}$
	G_2	$n_{21}(c)$	$n_{22}(c)$	$n_{2.}$
		$n_{.1}(c)$	$n_{.2}(c)$	$n_{..}$

$n_{.2}(c)$, but the arise of $n_{11}(c)$, $n_{21}(c)$ and $n_{.1}(c)$. The row totals don't change according to c because they don't depend on the classification rule adopted but exclusively on the sample composition.

From the previous cofusing matrix, it's calculated the probability to assign to G_2 a subject belonging to G_1, $P(\hat{G}_2|G_1)$, case that's also called false positive (FP) and defined by the following equation

$$\alpha(c) = \frac{n_{12}(c)}{n_{1.}}$$

and viceversa, the estimation of the probability to assign to G_1 a subject belonging to G_2, $P(\hat{G}_1|G_2)$, event also called false negative (FN) and defined by the following equation

$$\beta(c) = \frac{n_{21}(c)}{n_{2.}}$$

At this point, we can define two ways to measure the dependence between dichotomous variables:

- Specificity: it's defined as the probability to correctly classify a subject belonging to the G_1 class and one's complement of the probability of false positive.

  $$P(\hat{G}_1|G_1) = 1 - \alpha(c) = \frac{n_{11}(c)}{n_{12}(c) + n_{11}(c)}$$

  This value increases following the growth of c ($n_{12}(c)$ decreases and $n_{11}(c)$ increases)

- Sensitivity: it's defined as the probability to correctly classify a subject belonging to the G_2 class and equals the complementing form of $\beta(c)$.

$$P(\hat{G}_2|G_2) = 1 - \beta(c) = \frac{n_{22}(c)}{n_{21}(c) + n_{22}(c)}$$

  This value decreases following the growth of c.

The ROC curve then is the locus of points having as coordinates $\alpha(c), 1-\beta(c)$, contained in the square $[0,1] \times [0,1]$. In other words, the curve is the set of points $P_c(\hat{G}_2|G_2), P_c(\hat{G}_2|G_1)$ and puts together the probability to correctly predict G_2 with the probability to uncorrectly predict G_2, false positive.

Now we take advantage of this instrument to find the threshold related to the maximum value of the sum between sensitivity and specificity. We consider that threshold our cutting value to classify our data as reliable or not in order to find outliers in our analysis without the support of video-recorded trials or unreliability manually noted. Similarly it is possible to consider as cutting value that point where the probability to classify fake outliers is low, in order to keep the highest number possible of reliable trials.

3.4 Cross-validation

As previously announced, the usual method to measure a classification function quality consists in using part of the sample to estimate the model and a restriction to evaluate the error. Operationally, given the sample $[\mathbf{y}, X]$ composed by n units, some $n_1 < n$ units are chosen to build the classification rules and the rest $n_2 = n - n_1$ are set aside to evaluate the effectiveness of the rules, i.e. to measure the generalizability of the procedure. The first units set is called *training set* and the second *test set*. There is not a fixed rule to choose the dimension of each set,

neither there is a unique way to choose the n_1 units but it is common to define n_1 equal to the $90 - 95\%$ of the sample extracting random units. Refining this idea, it is also possible to repeat the procedure (divide the sample, define the classification rule, count of the errors) obtaining that the mean number of committed errors will be used to estimate generalization errors made by the classification rule applied. Alternatively, as in this case, it is applicable the *leave-one-out cross-validation* method, which consists in isolating one i random unit, building the classification rule on the remaining $n - 1$ units and comparing the criteria results to the effective i starting class; this procedure is repeated for each unit in the sample. Increasing the rule complexity indefinitly decreases the classification error on the initial class, but the generalizing error, the one measured on the test set, decreases to a certain level and then increases if the rule becomes more complex. The point is to try to identify the right complexity level.

Chapter 4

EEG Data Analysis

the first paragraph entirely explains the purpose of this essay i.e. the procedure applied to detect outliers from data, the classification methods to mark the trials as "unreliable"; throughout the paragraph an example, conducted on a restriction of the dataset, is used to better clarify the procedure adopted. In the second paragraph is described in detail each step of the data processing, considering the data's initial conditions, describing the filtering procedure and parameters and how the principal components analysis has been conducted. The third paragraph reports some graphics and the modeling results.

4.1 Methods

In order to discuss the whole procedure in the best possible way, an example using only a single subject partition of the dataset is going to be necessary, but in the real case this procedure has been applied to all the subjects included in the experiment.

Now, just the "INV-R" combination is chosen between the four kind of trials and according to the figure 4.1 only the enlighten channels are included: those channels were picked to better explain through signal graphics where is to detect the unreliable trials due to an extraordinary

line in the trace.[1]

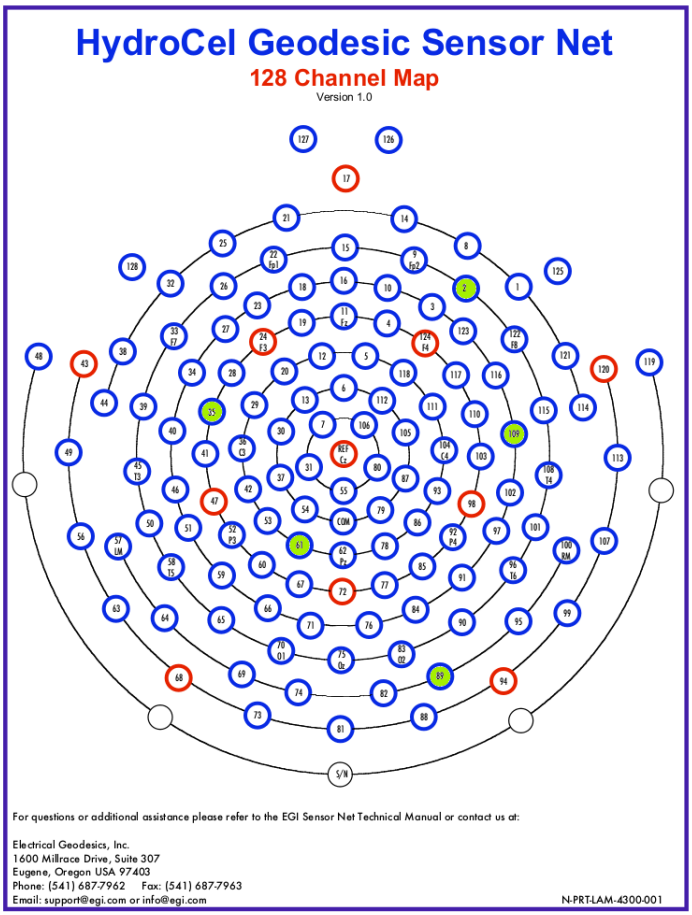

Figure 4.1: This is the channels map. Enlighten in green we can locate the channels included in this example

Lets assume that the figure 4.2 is the representation of a "small" 3D data matrix, made by those five channels considering the second dimension, the time code in the first dimension and the trials related to the "INV-R" condition in the third one. Lets call X the matrix extracted from a slice of the 3D matrix (slice having the time as the rows and the trials as the columns), through the singular value decomposition it is possible to rewrite X as

$$X = UD\Gamma'$$

[1] Wang et al. 2014

where Γ' is the eigenvectors matrix and each column γ'_i is a unit norm vector containing the contributions of each channel to the first principal component; in order to pusue our goal, only the first column is considered and possibly its eigenvalue. Collecting the first eigenvectors from each slice analyses, a matrix of eigenvectors is built and is called G. Now, having

$$G = [\gamma'_1 \dots \gamma'_i \dots \gamma'_{124}]$$

where[2] each column is the first eigenvector from the i-th channel, the principal components analysis is applied on that G' matrix, so that is possible to consider

$$G = U_G D_G \Gamma_G$$

With these analysis results, it can be useful to have a graphic representation. The figure 4.2 can be considered the graphic representation of each slice of the X matrix.

These graphics are based on the manual classification, coloring in blue the valid trials and in red the invalid ones. Considering this situation it is to be expected to have at least 6 unreliable trials basing that count on channel 2 representation.

The next step is to compare the manual classification with the results obtained from the previous analyses.

In figure 4.3 each point represents a trial and the label "_U" stands for those trials manually identified as "unreliable". The arrows are unit norm vectors and do not indicate any particular correlation because of the centering applied in the preprocessing phase. In an hypothetically perfect situation, it is to be expected that higher scores of the supplied data on the principal components may correspond to an outlier trial

[2] 124 is the number of channel considered in this experiment. Usually EEG experiments consider 129 channels, but in this particular case data concern children in tender age (6-months-old) and the lack of those channels information is dued to the exlusion of two sensors from positioning (those usually positioned below the eyes) that the subjects could not stand and two channels because of their null variance.

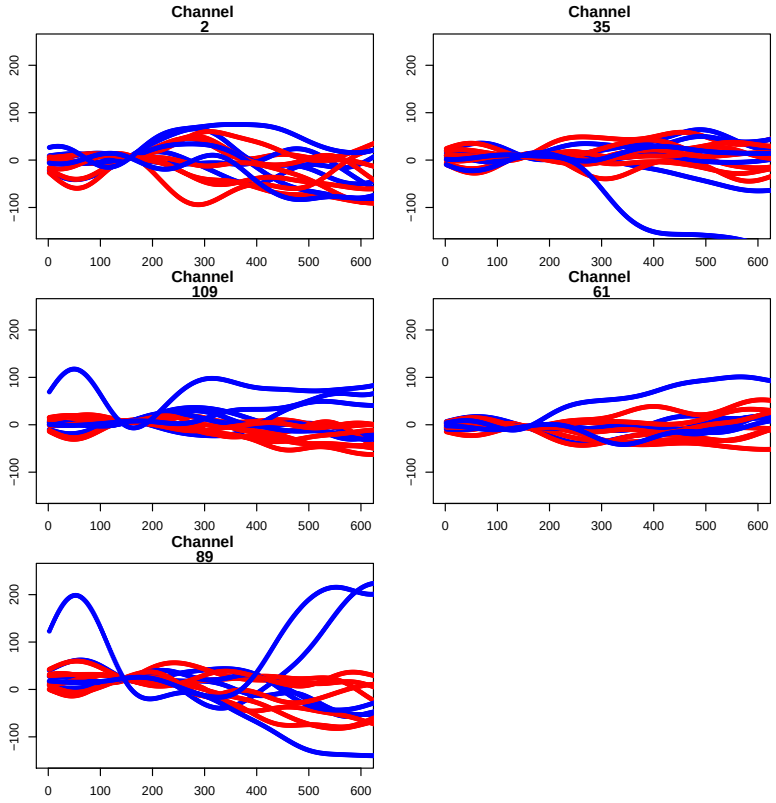

Figure 4.2: Signal related to 5 channels extracted from ID 4 trials

represented in the biplot by dots far from the center of the axes. At this point of the procedure, considering all the subjects became mandatory: a logistic regression is estimated considering `u` the response variable (assuming "1" if the trial has been manually defined "unreliable", "0" otherwise) and all the subjects, a distance (that will be later explained) and the interation between them as the indipendent variables. That kind of model can be useful under multiplex good points:

- it can be processed through the `auc` function in `R` to evaluate the area under the curve (a value between 70% and 100% is a good result), sensibility and specificity, and it is possible to define a threshold useful to classify each trial;

- considering the residuals extracted from the fitted model, it is possible to run a leave-one-out cross validation to evaluate the classification function, defining sensibility and specificity closer to reality.

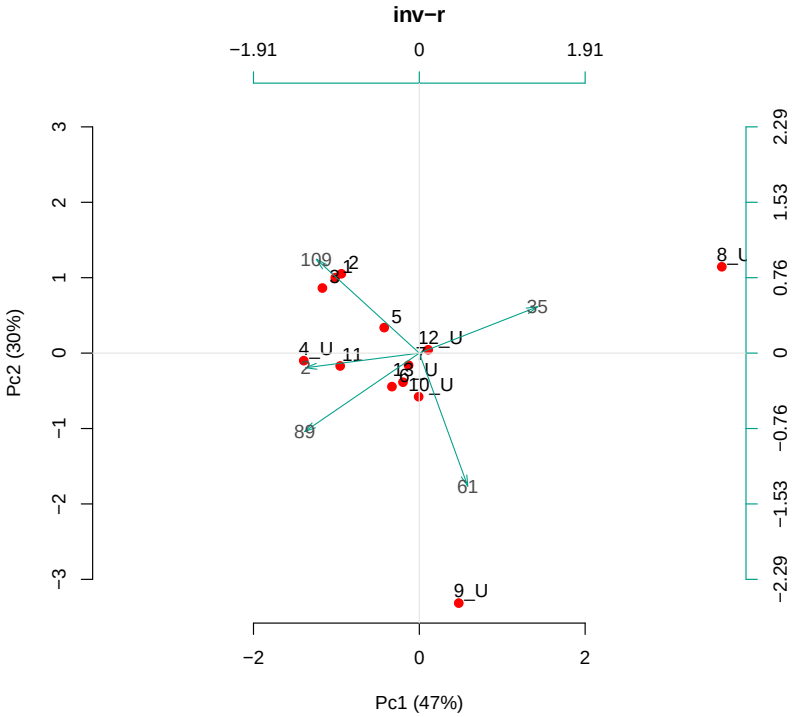

Figure 4.3: Biplot related to the "INV-R" condition of the ID 4 trials obtained by using `center=OverAll, scale=FALSE`

4.2 Procedure

4.2.1 Data initial conditions

In the first place, the EEG dataset was made three different subjects: each directory contained a file for each trial the subject ended with success.[3] This detail makes everything statistically more complex because not disposing of the same number of trials for each subject doesn't make the sample perfectly comparable, but this isn't exactly this work purpose. Some researchers, examining the videos related to each trial, kept track of the trial containing a blink or a saccade: those trials were marked as "unreliable". The purpose is to detect those trials through data processing and to measure the probability we have to classify correctly each trial.

4.2.2 Exploring data

Talking about tecnical issues in preprocessing, the first one we bumped into manipulating data was how to keep trace of the characteristics of each trial (type, number and reliability) absorbing these informations from the file name. So we wrote a few lines of code, based on the codelines of each file in order to select and keep only the needed characters. The following step was to built a 3D data matrix, one for each subject, containing in the first dimension each value recorded during the experiment, in the second dimension a list of the 125 channels used to record the brain electrical activity and in the third one we keep a label to identify the trials.

[3] Richards 2005. Testing was done only if the subjects maintained an alert, awake state during the procedure (eyes open, no fussing or crying, responding to the protocol). If the infant became fussy a short break was taken and the presentations were paused and then restarted. [...] The session was continued as long as the infants were not fussy in order to obtain as many trials as possible.

4.2.3 Filtering

To smooth the signal, we decided to apply a Butterworth low-pass filter. To ensure to this experiment to be replicable, we provide parameters we used: we assigned `n=3` to define the filter order, `W=1/40` as critical frequencies of the filter, considering that for digital filters, W must be between 0 and 1 where 1 is the Nyquist frequency and `type=‘‘low’’` to require the low-pass filter. To apply this filter to our data, we use the `R` function `butter` containing the function `filtfilt` in a double `for` cycle to fiter data by trial and by channel, replacing the new filtered data in the former 3D matrix structure.

4.3 Principal Components Analysis

In order to apply the Principal Components Analysis, the common solution is to consider standardized data. Besides, lacking of a referring model in EEG literature about what mean suits the best to center our data and about the possibility to use scaled data or not, we act like each configuration is equally good until proven otherwise. We decided to use two kind of mean:

- over all: the mean is calculated by each channel for each trial and then is subtracted from each channel in the original data.

- by trial: the mean is calculated by condition so we calculate four “big” means, one for each INV-L, INV-R, VAL-L, VAL-R condition, and then each mean is subtracted from the channels by condition.(This is the default mean `R` uses to treat PCA but we used this standard, splitting our sample into four groups to take advantage of each condition). If center is `TRUE` then centering is done by subtracting the column means (omitting `NAs`) of `x` from their corresponding columns.

Also the scale parameter can have the logic values `TRUE`/`FALSE`. If scale is `TRUE` then scaling is done by dividing the (centered) columns of `x` by their standard deviations if center is `TRUE`, and the root mean square otherwise. In other words, considering the center parameter set as `TRUE`, the analysis is based on covariance matrix with scale set as `FALSE` and on correlation matrix otherwise.

This undefined situation about centering and scaling data lead us to produce four versions of everything we had to estimate and classify.

With the scaled and centered data matrix, we produced an estimate of the signal based on the mean by channel and by condition, without principal components extraction.

4.4 Outliers Detection

After manipulating the signal according to our needs by centering and scaling the dataset, the principal components are estracted and the related loadings are used to graphically find the outliers (the analytical-mathematical analysis of the issue will soon follow with the analysis of the area under the ROC curve) . A simple example of a biplot can be the following figure 4.4. According to the legend, it seems that in this case manual and mathematical classification are equal regarding trials "INV-R 12_U" and "INV-R 11_U" but they are in contrast about trials "INV-R 10_U" and "INV-R 7_U" which are mathematically considered "reliable".

4.4.1 Area under the ROC curve

After the first principal components extraction, we apply the Singular Value Decomposition to our eigenvectors by condition and we use these new results in a biplot to actually see the outlier trial(s). These analyses were conducted by subject and is produced a biplot for each

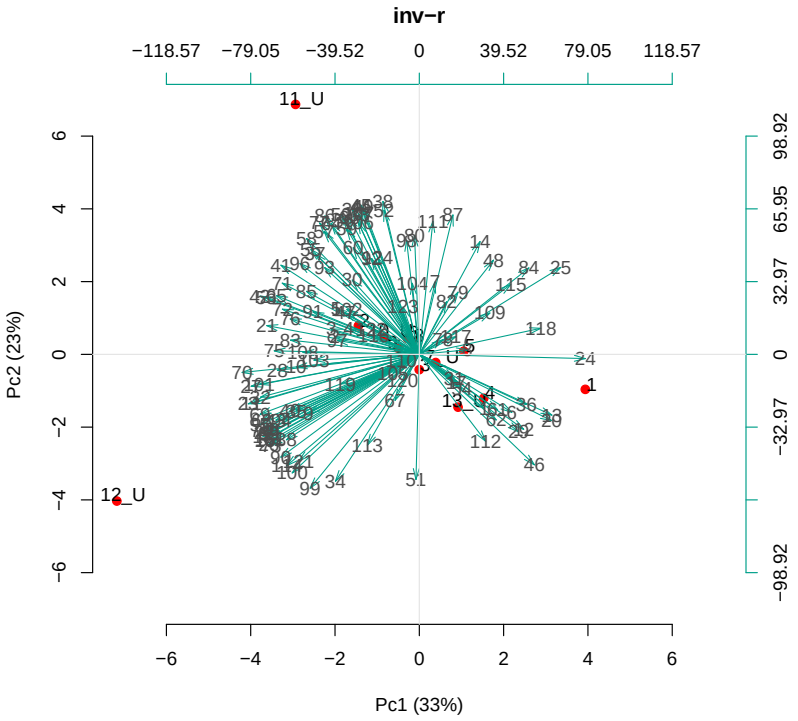

Figure 4.4: ID 10: biplot considering all the channels in the "INV-R" condition. Coded with "1" are those trials manually labelled as "reliable", and with "2" are those trials manually labelled as "unreliable". The `center=OverAll, scale=FALSE` case of preprocessing is shown here.

condition. During this analysis we keep memory of some distances, previously appointed, that will be useful later in the mathematical part:

- `mod1dir`: in this variable we save the product between the opposite of the first score and the sign of the mean of the components of the first loading as explained by the formula

$$-pc1\$u[i,1] \times sign\left(\sum_{j=1}^{124} pc1\$v[j,1]\right)$$

  where `pc1$u` indicates the scores of the SVD and `pc1$v` indicates the eigenvectors matrix;

- `mod1`: in this variable we save the absolute value of the first score;

- `mod2`: in this variable we save the square root of the sum by rows of the squared sum of the first and the second scores as explained by the formula

$$\sqrt{\sum_{j=1}^{n} \big(pc1\$u[j,1] + pc1\$u[j,2]\big)^2}$$

  where `pc1$u` indicates the scores of the SVD and `n` is the number of trials for that subject considering each condition.

We save these values also to represent a further graphic scenario by using the boxplot as in figure 4.5.

The horizontal line that appears in each boxplot corresponds to the estimated best cut according to each distance. The precise values are the following.

4.4.2 Classification

In this phase we discuss the mathematical aspect of outliers detection, using the Receiving Operating Characteristics curve. For each

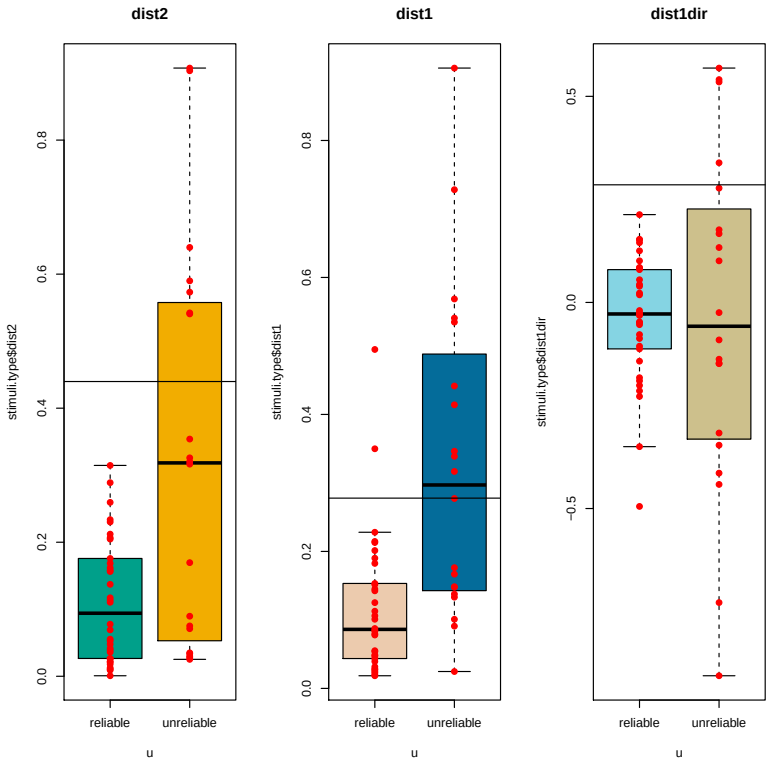

Figure 4.5: Graphics referred to the ID 4 obtained by using `center=OverAll, scale=FALSE` case of preprocessing. Each boxplot is related to the `dist2`, `dist1` and `dist1dir` distances respectively

of the previous saved distances, a logistic model regression is used to estimate our probability to consider a trial unreliable when it is not. As explained in the dedicated section, we handle the real data and the estimated ones to calculate specificity and sensitivity related to each possible threshold. This leads us to consider as the best cut the threshold related to the maximum value of the sum between sensitivity and specificity. This observation is consequent to the fact that we prefer a model having area under the curve which tends to one, the closer to one is the area, better is the model so we tend to consider acceptable those models having area between 0.5 and 0.7 and good the other ones having area over 0.7. The lack of a reference in the literature about preprocessing configuration leads us to test this procedure four times: the best results achieveable, in our opinion, are those obtained by using `center=OverAll, scale=FALSE`.

Considering this configuration, we achieved the best value of area under the curve (AUC) related to the `mod2` model as shown in figure 4.6. There are also some numbers supporting this theory, as in table 4.4.2 where there are reported all the values estimated in the previous paragraphs. There are also some new results about the sensibility and the specificity evaluated by leave-one-out cross-validation, the cross-validation error and relative error, everyone done for each distance. The `mod2` model has been considered the best also because of its low relative error in cross-validation procedure.

model	AUC	best cut	sens	spec	xsens	xspec	xerr	relxerr
mod1dir	0.70	0.28	0.66	0.67	0.57	0.60	0.41	1.39
mod1	0.75	0.28	0.63	0.76	0.55	0.73	0.32	1.09
mod2	0.75	0.44	0.49	0.89	0.46	0.86	0.26	0.88

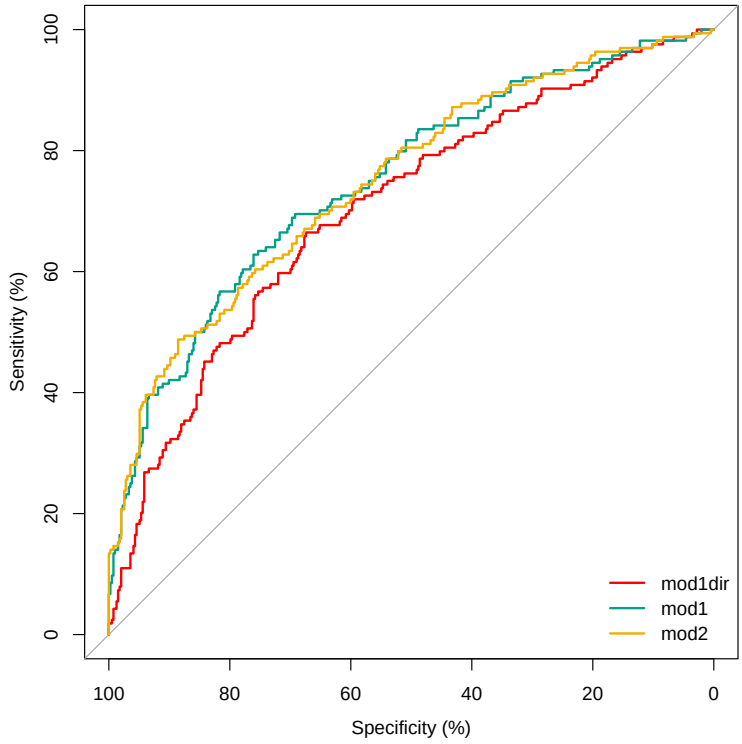

Figure 4.6: Graphics referred to the area under the ROC curve obtained by using `center=OverAll, scale=FALSE`

Conclusions

The work discussed in these pages is to be considered at its embryonic stage. There are some good aspects to be considered but there are also some improvements to be made. The pros are:

- this method is automatic, replicable and appliable to young subjects or to those subject who can stand less electrodes than the number stated by EEG standard protocols;
- through this method it is possible to consider "reliable" or not a trial basing this classification on mathematical analysis, without wasting time watching the video-recorded trials;
- if there are some errors in classification, they are certainly due to the method and are not to be considered caused by manual mis-labelling insertion.

The cons is that this method does not distinguish between the types of artifact, a removed trial can be disturbed by anything.

Bibliography

[1] C. Aaen-Stockdale et al. "Biological motion perception is cue-invariant". In: *Journal of Vision* 8.8 (2008), pp. 6–6.

[2] A. Achim, F. Richer, and JM. Saint-Hilaire. "Methodological considerations for the evaluation of spatio-temporal source models". In: *Electroencephalography and clinical Neurophysiology* 79.3 (1991), pp. 227–240.

[3] P. Berg and M. Scherg. "A multiple source approach to the correction of eye artifacts". In: *Electroencephalography and clinical neurophysiology* 90.3 (1994), pp. 229–241.

[4] P. Berg and M. Scherg. "Dipole modelling of eye activity and its application to the removal of eye artefacts from the EEG and MEG". In: *Clinical Physics and Physiological Measurement* 12.A (1991), p. 49.

[5] A. Gevins et al. "Deblurring". In: *Journal of clinical neurophysiology* 16.3 (1999), pp. 204–213.

[6] Bergen fMRI Group. *ERP vs EEG*. 2015. URL: `http://fmri.uib.no/index.php?option=com%5C_content%5C&view=article%5C&id=82:%20eeg-and-erp%5C&catid=52:key-notes%5C&Itemid=123`.

[7] C. A. Joyce, I. F. Gorodnitsky, and M. Kutas. "Automatic removal of eye movement and blink artifacts from EEG data using blind component separation". In: *Psychophysiology* 41.2 (2004), pp. 313–325.

[8] TP. Jung et al. "Removing electroencephalographic artifacts by blind source separation". In: *Psychophysiology* 37.02 (2000), pp. 163–178.

[9] S. J. Luck. "An Introduction to the Event-Related Potential Technique (Cognitive Neuroscience)". In: (2005).

[10] S. J. Luck and M. Girelli. "Electrophysiological approaches to the study of selective attention in the human brain". In: *The attentive brain* (1998), pp. 71–94.

[11] S. J. Luck, G. F. Woodman, and E. K. Vogel. "Event-related potential studies of attention". In: *Trends in cognitive sciences* 4.11 (2000), pp. 432–440.

[12] D. A. Overton and C. Shagass. "Distribution of eye movement and eyeblink potentials over the scalp." In: *Electroencephalography and Clinical Neurophysiology* 27.5 (1969), pp. 546–546.

[13] J. Pernier, F. Perrin, and O. Bertrand. "Scalp current density fields: concept and properties". In: *Electroencephalography and clinical neurophysiology* 69.4 (1988), pp. 385–389.

[14] J. E. Richards. "Localizing cortical sources of event-related potentials in infants' covert orienting". In: *Developmental science* 8.3 (2005), pp. 255–278.

[15] PK. Sadasivan and N. D. Dutt. "SVD based technique for noise reduction in electroencephalographic signals". In: *Signal Processing* 55.2 (1996), pp. 179–189.

[16] R. L. Smith. "Multivariate Analysis". In: (1999).

[17] N. F. Troje. "Biological motion perception". In: *The senses: A comprehensive reference* 2 (2008), pp. 231–238.

[18] R. N. Vigário. "Extraction of ocular artefacts from EEG using independent component analysis". In: *Electroencephalography and clinical neurophysiology* 103.3 (1997), pp. 395–404.

[19] L. Wang et al. "The feet have it: Local biological motion cues trigger reflexive attentional orienting in the brain". In: *NeuroImage* 84 (2014), pp. 217–224.

List of Figures