Signal processing methods for EEG data classification

by
Andreas Soteriou Varnavas
M.Eng(Hons)

A Thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy of University of London and
Diploma of Imperial College

Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
Imperial College London
University of London
2008

Abstract

The scope of this thesis is to determine appropriate features of a person's electroencephalographic (EEG) data and the way in which they can be used to predict their performance in an "oddball" experiment. We classify a person's performance in one of the following classes: "success" or "failure", depending on the reaction time related with it. Predicting a person's performance means finding the correct class where the latter belongs to, using the person's EEG data corresponding to a time period before the reaction takes place.

The problem is addressed in various ways as far as the feature construction process is concerned, whereas a Gaussian classifier is used in all cases. First, the raw time signals and the magnitude of their Fourier Transform are used as features. Then the number of these features is reduced, using various feature selection methods in combination with Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Non Negative Matrix Factorization (NMF). Subspace methods, using PCA and NMF to construct different spaces for the two classes, are also used to perform the desired classification. Features are also constructed using a Time-Frequency representation of the EEG signals. In this case we propose two novel algorithms which analyze the magnitude of the Time-Frequency representation using NMF, in a single or multi-trial basis, and the coefficients of selected NMF components are used as features.

Finally, a novel algorithm performing the desired classification based on the construction of signals characterising each of the classes is proposed. These characteristic signals are constructed linearly combining the EEG signals of the various channels, minimising the variance of the time samples over the trials belonging to the same class. A novel algorithm is also proposed for selecting the appropriate channels to be used in the construction of the characteristic signals. This algorithm is based on the identification of

the channels showing the least interference from background brain activity.

The maximum classification rate produced for one of the 11 subjects in our study is 97.22%. However the rates usually vary between 70% and 80%. Considering the difficulty of the problem, this is encouraging. However, with these classification rates, real applications should only consider the generation of notification signals to increase the attention of operators and not involve any critical, automatic decision making process. Moreover, the variability in the methods and channels being optimal across the various subjects, implies that in a real case, a "tailor made" system should be designed for each user.

Acknowledgment

First of all, I would like to express my gratitude to my supervisor, Professor Maria Petrou. Her personal guidance and encouragement throughout the whole time of research and writing of this thesis were really important to me.

I would like to give my thanks to the Centre of Human Sciences of QinetiQ for supplying the EEG data used in this thesis.

I would also like to thank my friends from Imperial College (and outside) for all the interesting conversations we had throughout these years.

Last but not least, I would like to thank my parents and my brother, Petros, for their love, support and understanding.

Contents

List of Figures

List of Tables

Statement of Originality

As far as the author is aware, the following aspects of the thesis are believed to be original contributions:

1. The feature construction algorithms using Non Negative Matrix Factorization for the analysis of single and multi-trial time-frequency representations of EEG signals presented in sections 4.5 and 4.6, respectively.

2. The classification algorithm using class characteristic signals remaining as invariant as possible over the trials of the same class presented in section 5.1.

3. The channel selection algorithm indicating the EEG channels believed to exhibit the largest correlations with the cognitive processes of an ongoing task presented in section 5.2.

Moreover, the classification algorithm of section 3.6.1 using Non Negative Matrix Factorization to construct subspaces for the two classes, although it has been used in Image Processing, it is the first time being used with spectral features in EEG processing. The rest of the methods in chapter three are well known techniques for feature reduction but it is the first time being used for the specific classification problem.

Papers published or submitted

1. Andreas Varnavas and Maria Petrou. Human's Performance Prediction with Wavelet Analysis of EEG Data. *Proceedings of the* 15^{th} *International Conference on Digital Signal Processing*, pp.167-170, July, 2007.

2. Andreas Varnavas and Maria Petrou. Performance Prediction using EEG and Trial-Invariant Characteristic Signals. *Proceedings of the* 30^{th} *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, August 2008.

3. Andreas Varnavas and Maria Petrou. Human Performance Prediction using EEG: A Comparative Study. *submitted to IEEE Transactions on Biomedical Engineering.*

4. Andreas Varnavas and Maria Petrou. An EEG Channel Selection Algorithm for Human Performance Prediction. *submitted to IEEE Transactions on Biomedical Engineering.*

Abbreviations

AR: AutoRegressive

ARMA: AutoRegressive Moving Average

ARX: AutoRegressive with eXogenous input

BCI: Brain Computer Interface

CWT: Continuous Wavelet Transform

DFT: Discrete Fourier Transform

DWT: Discrete Wavelet Transform

EEG: Electroencephalographic

EOG: Electro-oculogram

EP: Evoked Potential

ERP: Event Related Potential

FIR: Finite Impulse Response

fMRI: functional Magnetic Resonance Imaging

HPM: Human Performance Monitoring

KPCA: Kernel Principal Component Analysis

LDA: Linear Discriminant Analysis

LMS: Least Mean Squares

LPCA: Linear Principal Component Analysis

MA: Moving Average

NMF: Non Negative Matrix Factorization

PCA: Principal Component Analysis

PF: Performance Factor

RBF: Radial Basis Function

SNR: Signal to Noise Ratio

SVM: Support Vector Classification

WT: Wavelet Transform

Chapter 1

Introduction

Electroencephalography is one of the easiest, cheapest and most widely used ways of recording the electrical activity of the brain. Since its discovery, there is a high interest of associating the recorded signals of the brain with the cognitive processes or the physical/psychological condition of the person at the time of recording. A usual way of studying the relation between a person's cognitive processes and the recorded, electroencephalographic (EEG) signals is through "oddball" experiments. During an oddball experiment a person monitors a screen, where various stimuli are presented, and has to respond by pressing a button upon the presentation of a specific type of stimulus. In this thesis we propose ways of using a person's EEG signals in order to "predict" their performance in such an experiment.

In section 1.1 we give a description of the area of electrophysiological studies of cognition and the applications related to it. In section 1.2 we present the basic concepts of electroencephalography and the experimental procedures that are related to our study. In section 1.3 we define the problem we aim to solve, we explain our approach to it and give a detailed description of the experiment and the data we are going to use. The objective of this thesis is presented in section 1.4. Finally, in section 1.5 the outline of the thesis is given.

1.1 Research Area

Apart from the obvious clinical applications, EEG data can also be used for cognitive studies a special application of which is behavioural prediction. We will discuss next, in brief, the meaning of cognitive studies and then the meaning of behavioural prediction.

1.1.1 Electrophysiological studies of cognition

An Electrophysiological study of cognition tries to find correlations between a person's EEG data and their cognitive processes or physical/psychological condition at the time. The generalization of these correlations can give us a high level description of a person's unknown mental process or condition, provided that we have access to their EEG data.

These mental processes and conditions can be divided into two categories. In the first category belong cases where the subject receives an external stimulus, most often visual or auditory. If this evokes a mental process then it is usually a cognitive one that has to do with recognizing and evaluating the stimulus. Parameters of the response, following the mental process, give a high level description of the latter. These parameters may be connected with the accuracy, the confidence and the time of the response. In the cases when the external stimulus evokes emotional conditions, the output usually is a binary high level description (e.g. "like" or "dislike").

In the second category belong cases when the subject receives no external stimulus. The mental processes that we would like to describe this time can be cognitive processes of different nature, such as those evoked by writing a letter or solving a mathematical problem. The description of the mental process in this case is in fact the subject's high level action that the process is related with. Another interesting case is imagining of different kinds of action. These actions should belong of course to a finite set, usually binary. Finally, examples of mental conditions belonging to this category are mainly clinical ones, where the description of the condition is the existence or not of a mental disease.

Several potential applications of the correlation of EEG data with both kinds of

mental processes and conditions have been proposed. As far as the first category is concerned, an interesting type of applications are the ones related to the field of *Human Performance Monitoring* (HPM). There are many taks such as air traffic control or military applications where an officer's efficient response to an external stimulus is of high importance for safety reasons. It is natural the officer's performance to vary during the time period of the task, depending on their physiological condition. In such cases, a machine that would be able to indicate when a person's performance falls below an acceptable limit using their EEG data, could alert that a further action should be taken. Such an action could be the replacement of the officer with a new one.

Applications in the field of *Brain Computer Interface* (BCI) concern data coming from the second category. The goal in this case is to provide the ability to people to interact with a machine using their thought. This can be potentially useful to people facing a number of disabilities, from movement impairment to "locked in" syndrome. Finally various medical applications have been proposed, concerning the detection of mental diseases at an early stage, such as Alzheimer's disease. These applications belong to both categories, as in some of them there is a use of an external stimulus whereas in other there is not.

1.1.2 Human performance prediction

A challenge similar to human performance monitoring is to predict a person's performance in an action following an external stimulus, using their EEG data. The notion of prediction implies that the estimation of the person's performance should be computed before the action has taken place. This imposes two important constraints compared with the approaches followed for human performance monitoring:

- Only the part of the EEG signals preceding the person's action can be used for the desired prediction.
- No averaging can take place among the EEG data corresponding to consecutive stimuli presentations as the person's performance upon the presentation of a *single*

stimulus should be predicted.

An efficient algorithm for the prediction of human's performance should be used in safety critical applications, enabling the machine to take a default action to minimize the losses if it indicates that the person's performance is not going to be adequate.

1.2 Electroencephalography

In this section we present some basic information concerning the way EEG data are collected from the instrumentation point of view and from the experimental design point of view. We also explain what the event related potential is and the way its components are related to various cognitive processes. Relations between the frequency bands of an EEG waveform and various cognitive processes are also presented.

1.2.1 The Electroencephalogram

Electroencephalography is the first and most popular way of non-invasively observing human brain activity. The electroencephalogram is the recording of the electrical signals naturally produced by the brain, using electrodes placed on the subject's scalp. The recorded electrical potentials are produced by extracellular synaptic trans-membrane currents in neuronal dendrites. For a detailed study of the human brain electrophysiology we refer to [8, 84].

The standard system for placing the electrodes on the scalp is the 10-20 system. According to it, the nasion, inion, left and right pre-auricular points are used as reference points. We give below a description of the visible surface of the human scull that these points refer to.

- The *nasion* is the distinctly depressed area directly between the eyes and superior to the bridge of the nose.
- The *inion* is the bulging part at the lower rear part of the human skull.

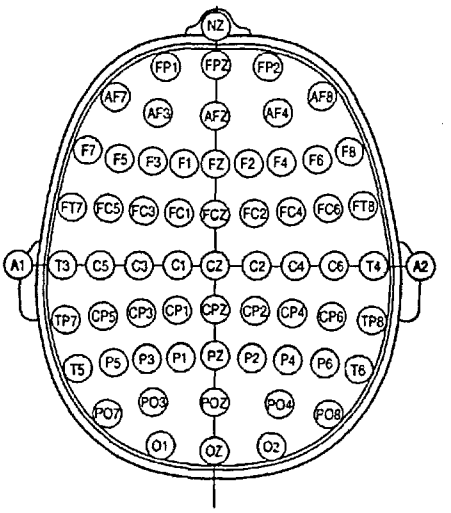


Figure 1.1: The Electrode nomenclature according to the International Federation of Clinical Neurophysiology's 10-20 system.

- The *pre-auricular* point is the point in front of the ear, between the ear's opening and the cheek.

The electrodes are placed at fixed percentages of the distances between these reference points. The commonly used distance between two electrodes is 10% or 20 % of a scalp measurement. The locations of the electrodes are determined by an abbreviation which uses letters that refer to different regions on the scalp and numbers for relative locations. Figure 1.1 shows the electrode nomenclature in the 10-20 system. For more information concerning the standards of EEG recordings we refer the reader to [65].

The term *channel* refers to a pair of two recording electrodes between which the difference in electrical potential is being measured. There are two different ways of recording the potential on the scalp. In *bipolar recording*, voltage differences are recorded between adjacent electrode sites, whereas in *referential recording*, the voltage differences are recorded between the various electrode sites and a common reference electrode. The reference electrode is usually placed at a location far from the source of electrical brain activity (e.g. ear, mastoid or chin).

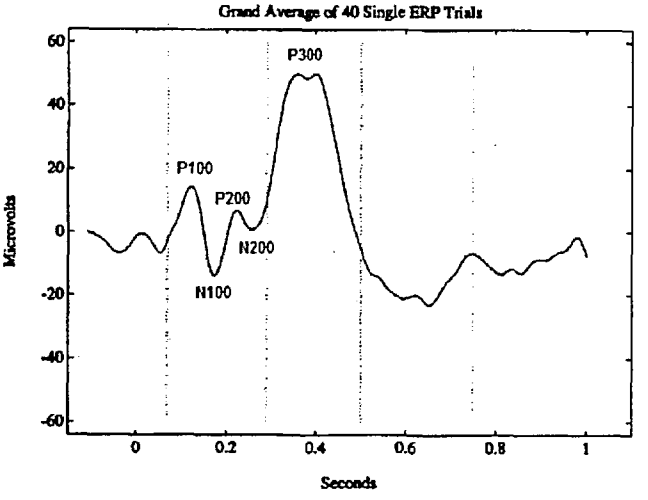


Figure 1.2: The ERP signal as acquired by averaging 40 EEG signals corresponding to different trials of an oddball experiment, (taken from [51]).

1.2.2 Event related potentials

The changes on the ongoing electroencephalogram due to the occurrence of an event (e.g. an external sensory stimulus), which needs to be evaluated by the subject, constitute the Event Related Potential (ERP). This signal is of great interest as it is directly related to the cognitive process caused by the stimulus.

The waveform of the ERP signal is the result of the superposition of the potentials of the neurons involved in the related cognitive processes. A description of the synchronization processes of the neurons that lead to the creation of the ERP signals can be found in [14]. In the time domain the ERP signal has a characteristic pattern (see Figure 1.2 [51]) which is usually acquired by computing a grand average of EEG signals, time locked to the event, across trials. This of course assumes that the ERP signal possesses the same amplitude and phase (latency) each time the event is repeated and that the background EEG is a zero mean stochastic process.

The ERP signals are described in terms of the succession of the components that follow the event that evoked them. These components are characterized by their ampli-

tudes and latencies from the stimulus onset and are named as 'Xno', where X is 'P' or 'N' if the amplitude is positive or negative, respectively, and 'no' denotes the approximate time, in milliseconds, at which the component usually appears after the stimulus. The ERP components are known to have functional meaning in relation to the cognitive processes in progress, some of which are given below [15,64,67,74]:

- The P100 and N100 components are mainly generated in the sensory cortex and are associated with the encoding of basic stimulus features. They can be used as an indication of the point in the EEG sequence where selective attention begins to emerge. Their amplitude is larger when the stimulus presented is the attended one.
- The N200 component is associated with sensory processing and with evaluation of stimulus information required for selecting a proper reaction. As such, it always precedes the reaction and is closely related to the reaction time.
- The P300 component has the largest amplitude and duration. It indexes memory storage and serves as a link between stimulus characteristics and attention. Concerning memory storage the amplitude of P300 has been found to be related to the initial encoding of the stimulus features in memory and has been used to predict whether a specific stimulus will be correctly recalled upon later presentation [30]. Moreover, the amplitude of P300 is related to the presentation frequency, stimulus sequence, stimulus quality, attention, and task relevance of the stimulus. It usually coincides or follows the person's reaction.

1.2.3 EEG frequency bands

As any waveform, the EEG recorded in each channel can be analyzed in terms of its frequency components. It has been found that the energy of the EEG in the frequency domain is concentrated in specific bands depending on the mental state and the cognitive processes that the subject undergoes. This is because of the synchronous depolarization of cells of neurons, involved in the foresaid mental state or cognitive processes, which results in the generation of rhythmical electrical activity. The above frequency bands are

named *natural rhythms* of the brain and their relation with various cognitive processes is summarized briefly as follows [7]:

- *Delta* (0.5-3.5 Hz): It is related to signal detection and decision making processes.
- *Theta* (4-7 Hz): It is highly correlated with associative processing. It is present in oddball experiments 300 msec after the target of interest is shown and contributes to the formulation of the P300 component.
- *Alpha* (8-13 Hz): It is associated with memory related processes. In oddball experiments the energy of the alpha rhythm before the stimulus onset strongly affects the N100 and P200 components.
- *Beta* (13-40 Hz): It is related to a wide range of mental activities such as integrated thinking, computing mathematical problems, planning and high level information processing.

1.2.4 Experimental procedures

The EEG data, which are used for the study of the cognitive processes related to the presentation of an external stimulus, are acquired through experiments that can be described as follows. Each person (subject) participating in the experiment receives a sequence of stimuli, each one separated from the other by a random period of time, which is long enough for the subject to respond. The stimuli are of the same nature (e.g. visual or auditory) but are divided into a finite number of types according to their characteristics. Each type of stimulus requires a different reaction from the subject. The EEG data of the subject are recorded for a specified duration of time, starting before the occurrence of each stimulus and ending after it. The part of the experiment over this period for a specific subject is named *trial*. Consecutive trials of the same subject constitute an *epoch*. Apart from the EEG data, various parameters characterizing the reaction of the subject are also recorded. The most common one is the time of the reaction.

A usual experiment of this type is the *"oddball" experiment*. In this case a series of two different types of events is presented to the subject. The occurrence of the event of

one type is more frequent than that of the other. The subject has to respond in some way only to the less frequent event (target event) as soon as possible. Therefore, the recorded subject's EEG data are related to their cognitive process of recognizing the rare event.

1.3 Problem definition

The scope of this study is to determine appropriate features of a person's EEG data and the way in which they can be used to predict their performance in a visual "oddball" experiment. A person's performance belongs to one of the following three classes: "success", "medium" or "failure", depending on the reaction time related to it. Predicting a person's performance means finding the correct class where the latter belongs to, using the person's EEG data corresponding to a time period before the reaction takes place. As we shall explain in section 1.3.4 we are only interested in correctly classifying performances belonging to classes "success" and "failure".

1.3.1 The oddball experiment and the raw data

The experimental data used in this study have been supplied by the Centre of Human Sciences of QinetiQ. They concern the EEG signals of 11 subjects which were recorded during an "oddball" visual detection experiment, as well as their corresponding reaction times.

The oddball experiment took place as follows: each subject monitored a screen where various visual stimuli were presented in a random order. The subject had to respond by pressing a button only when a specific type of stimulus (target) was presented. No reaction was required for the other type of stimuli (non-targets). The frequency of appearance of the target event was 1/10. The time period between two consecutive stimuli was random and lasted for at least 2304 msec. Overall each subject was shown 6 blocks of stimuli, with each block containing 400 stimuli (39 targets and 361 non-targets). The duration of the experiment for each subject was 2 hours, i.e. 20 minutes per block.

An 18 channel EEG, with the electrodes placed according to the 10/20 system

(their positions are marked with grey circles in Figure 1.1), was recorded for each subject during the whole period of 2 hours that the experiment lasted. The sampling frequency was $F_s = 500$ Hz and the type of recording referential. Then, blocks of duration of 2304 msec (1152 samples) were extracted from the originally recorded EEG. Each block (trial) concerns the presentation of a single stimulus and it is time locked to it as follows. The first 256 msec (128 samples) correspond to pre-stimulus activity and the remaining 2048 (1024 samples) to post-stimulus activity. Thus we were finally given 234 and 2166 trials for each subject, corresponding to the presentations of target and non-target stimuli, respectively. The trials of target stimuli were accompanied with the corresponding reaction time, which is the time duration between the stimulus onset and the requested press of the button.

We are interested here in predicting a subject's performance in the trials where a target stimulus was presented. Therefore, for this study only this type of trial will be used. Thus with the term "trial" from now on we shall mean a trial where a target stimulus is presented.

1.3.2 Definition of the classes of categorization

Since we want to predict a person's performance in each trial, we first have to decide the possible descriptions that this may have. Ideally a person's performance in a trial should be described by the corresponding reaction time and it would be the latter we should try to predict. However, because of the difficulty of the problem, arising from the need of early truncation of the EEG signals at an early point preceding the subject's reaction, we decided to focus our research on discriminating only between the very quick and very slow responses. For this reason we chose to describe a person's performance in a trial as successful, medium or failed, depending on the relative speed with which the person reacted. This results in three classes for the trials: "success", "medium" and "failure", which are defined as follows. Class "success" includes the 25% of the trials of a specific subject with the fastest reaction times and class failure includes the 25% of the trials with the slowest reaction times. The 50% of the trials with intermediate reaction times constitute the class "medium" and as explained in section 1.3.4 are ignored.

Table 1.1: Number of valid trials per subject

Subject	S1	S2	S3	S4	S5	S6	S8	S10	S11	S13	S14
Valid trials	116	98	73	186	159	198	70	125	126	129	96

1.3.3 Data selection

As mentioned in section 1.3.1 the number of available trials per subject is 234. However, only trials processed by the eye movement correction algorithm will be included in the analysis. This algorithm was used routinely by the people who supplied the data. In some cases, however, it failed to perform the correction and those cases are identified in the data set supplied. The number of valid trials that finally remain for each subject is shown in Table 1.1. Let us note here that the labeling of the 11 subjects is the one that QinetiQ used in the provided data, i.e.: S1, S2, S3, S4, S5, S6, S8, S10, S11, S13, S14.

As described in section 1.3.1 the data of each trial correspond to a time period of 2304 msec having both a pre-stimulus and a post-stimulus part. However since our purpose is to predict the subjects' performance in each trial, we are restricted to use in the analysis only the part of the data corresponding to the time period before the reaction time. Moreover, in a real time application the reaction time of a subject in a trial is not a-priori known. Thus, we decided to use in our analysis the parts of the EEG signals that correspond to the period before the smallest reaction time available in the set of valid data. In the single-subject based approach, we use in this thesis, this is the smallest reaction time of the specific subject studied. Thus, we finally keep the parts of the EEG signals which correspond to the time period between the stimulus onset and the minimum decision time to react. With the term EEG signals from now on we shall refer to the above parts of the signals unless it is otherwise stated.

1.3.4 Statistics of the reaction times

In order to appreciate how challenging it is to separate the three classes, in the way they have been defined, we have to look at the distributions of the reaction times for each

subject. Some statistics of the reaction time for each subject can be seen in Table 1.2 and the corresponding histograms are presented in Figures 1.3 and 1.4.

It can be seen that most reaction times cluster around the median (or the average) value and it must be very difficult to develop a classifier that will separate them successfully. In such studies, what one is really interested in, is identifying the cases of extra slow response, i.e. separate the class "failure" from the other two classes. In order to improve the separability of the data we choose to remove from our analysis the trials belonging to class "medium". This is not necessarily a big problem as in a real application a "medium" performance could be classified as either "success" or "failure" without severe consequences.

However, even distinguishing between the two classes "success" and "failure" is not an easy task. In order to gain some appreciation of the difficulty of the problem, in Figure 1.5 we plot the average signal for the class "success", for Subject 1, in all 18 channels, alongside the average of the class "failure". In Figure 1.6 the averages of the magnitude of the spectrum of the EEG signals between the stimulus onset and the time point of truncation are plotted for Subject 1. The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The corresponding plots for the remaining 10 subjects are given in Appendix A.

By observing the graphs of Figures 1.5, 1.6 and those in Appendix A we see that there is a huge overlap between the areas where the values of the majority of the signals of the two classes lie. This supports our argument concerning the difficulty in the discrimination of the two classes. Judging from the degree of overlapping we can say that the recording of some channels are more useful than the recording of other and that the useful channels are not the same for all subjects. This leads to the speculation that one might have to develop a "tailor made" system for each subject.

In tables 1.3 and 1.4 the number of valid trials and the statistics of the reaction times for classes "success" and "failure" are given, respectively.

Table 1.2: Statistics of the reaction times in ms

Subject	Valid trials	Mean	Median	Minimum reaction time	Standard deviation	Mode
S1	116	539	522	360	94	472
S2	98	944	872	592	261	792
S3	73	624	564	432	209	508
S4	186	705	658	428	175	{616, 612, 588}
S5	159	668	636	412	152	564
S6	198	722	712	424	120	688
S8	70	515	504	368	95	564
S10	125	847	808	468	211	{872, 780}
S11	126	662	654	436	99	680
S13	129	830	776	456	222	732
S14	96	995	960	540	272	864

Table 1.3: Statistics of the reaction times in ms for class "success"

Subject	Valid trials	Mean	Median	Minimum reaction time	Standard deviation
S1	29	436	444	360	26
S2	25	721	732	592	52
S3	19	481	496	432	26
S4	47	560	568	428	37
S5	40	530	546	412	43
S6	50	587	600	424	48
S8	18	410	414	368	21
S10	32	658	686	468	78
S11	32	554	564	436	36
S13	33	615	628	456	56
S14	24	711	718	540	65

Table 1.4: Statistics of the reaction times in ms for class "failure"

Subject	Valid trials	Mean	Median	Minimum reaction time	Standard deviation
S1	29	668	648	588	72
S2	25	1312	1220	1000	251
S3	19	880	784	644	267
S4	47	929	844	764	204
S5	40	859	800	728	173
S6	50	875	842	784	99
S8	18	641	620	564	66
S10	32	1099	970	896	250
S11	32	789	758	708	85
S13	33	1139	1088	916	191
S14	24	1361	1258	1156	236

1.4 Thesis objective

The objective of this thesis is to develop methodology that will allow the detection of class "failure" with maximum specificity, distinguishing it from class "success". As mentioned in section 1.3.4 we are not interested in identifying trials belonging to class "medium" or even separating them from class "failure". For this reason the trials of class "medium" are removed from our analysis. Thus from now on we have trials that belong to one of only two classes: "success" and "failure" and we aim to classify them to the correct class. In a real application, of course, one cannot exclude class "medium", but in a real application if class "medium" is wrongly recognized as class "success" or class "failure" does not really matter.

1.5 Outline of the thesis

This thesis is structured as follows:

- In Chapter 2 we present a review of the literature concerning methods for enhancing the potential of the EEG evoked by an external stimulus. Thus, such methods could be used for constructing features correlated with a person's cognitive processes during an oddball experiment. Moreover, we review approaches proposed in the

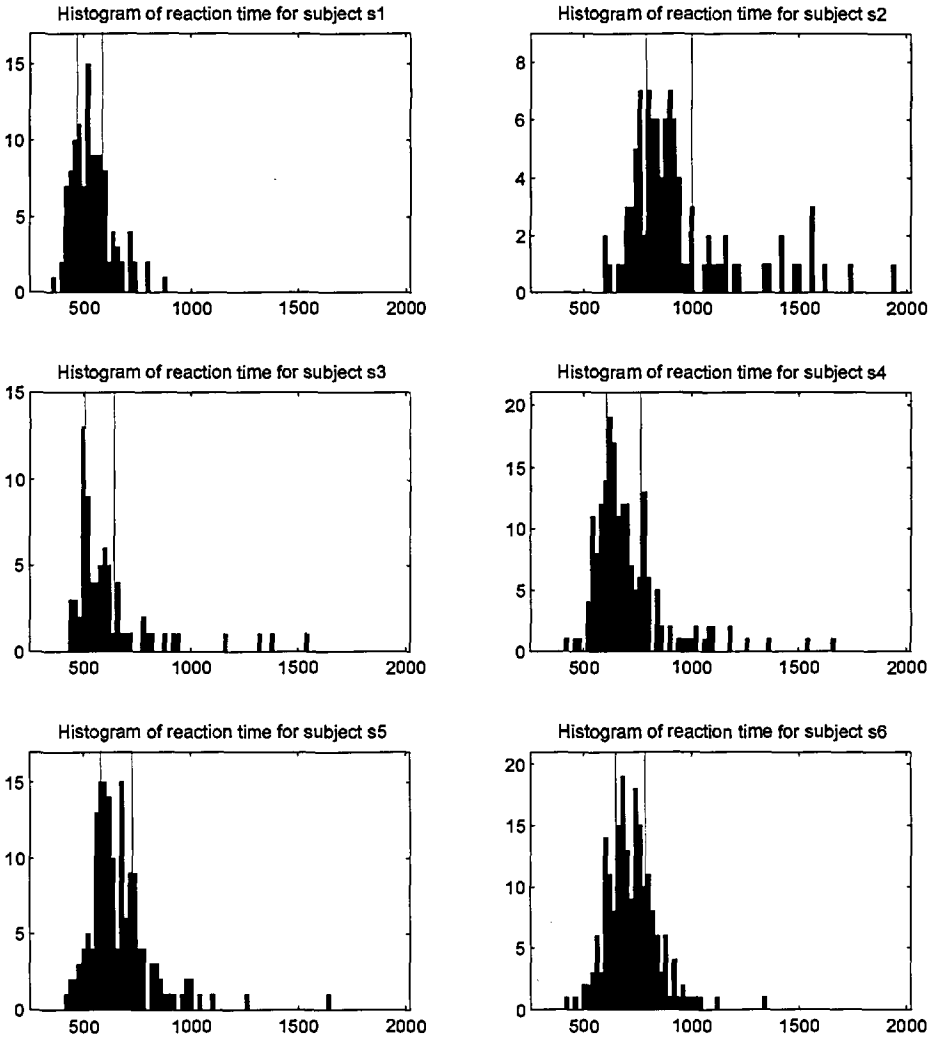


Figure 1.3: Histograms of the reaction time for subjects: S1, S2, S3, S4, S5, S6. The two red, vertical lines indicate the boundaries among classes "success", "medium" and "failure".

closely related to our problem field of Human Performance Monitoring.

- In Chapter 3 an initial approach to the problem is done using time and spectral features constructed from all channels and performing classification using a Gaussian Classifier. For a number of reasons we explain, we continue the analysis using spectral features corresponding to frequencies smaller than 40Hz. A number of methods

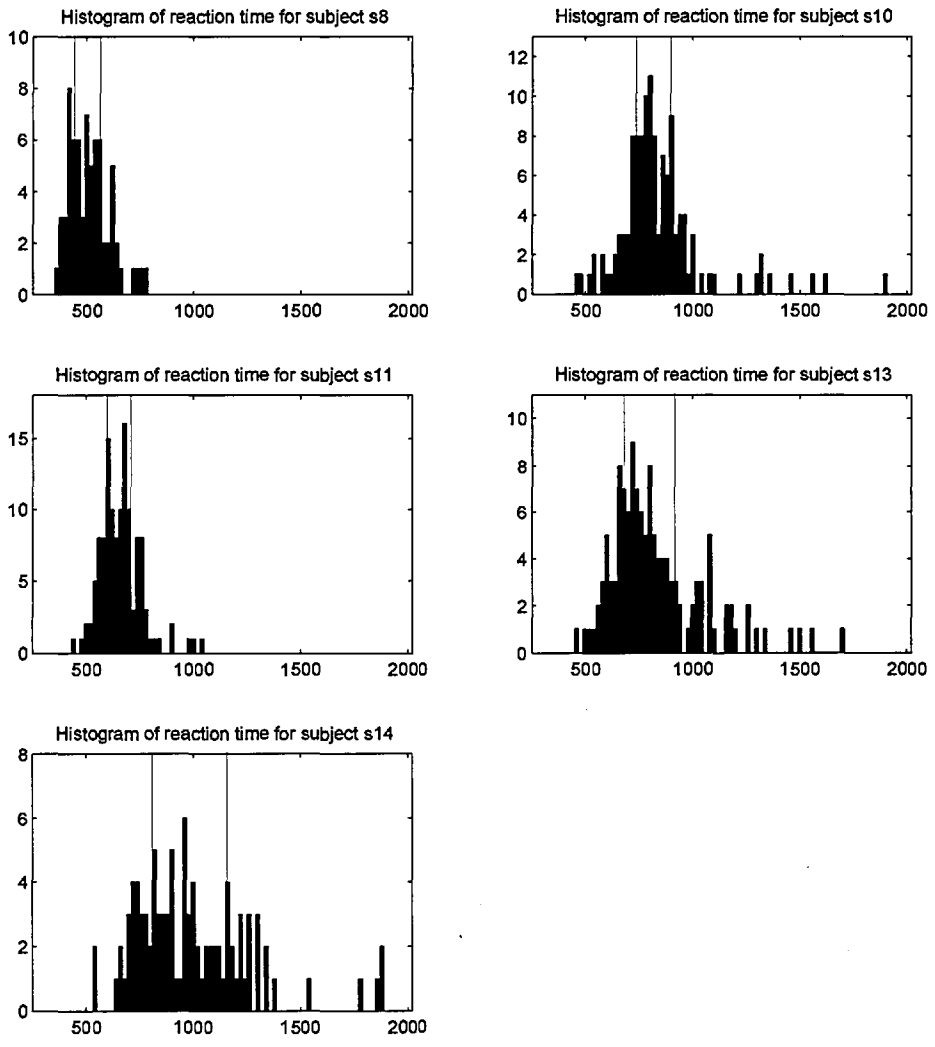


Figure 1.4: Histograms of the reaction time for subjects: S8, S10, S11, S13, S14. The two red, vertical lines indicate the boundaries among classes "success", "medium" and "failure".

are then used for feature reduction such as: selection of features according to their Euclidean distance between the two classes, Principal Component Analysis (PCA), Linear Discriminant Analysis and Non Negative Matrix Factorization (NMF). Finally subspace methods of classification using PCA and NMF are also investigated.

- In Chapter 4 the use of NMF for the analysis of time-frequency representations of

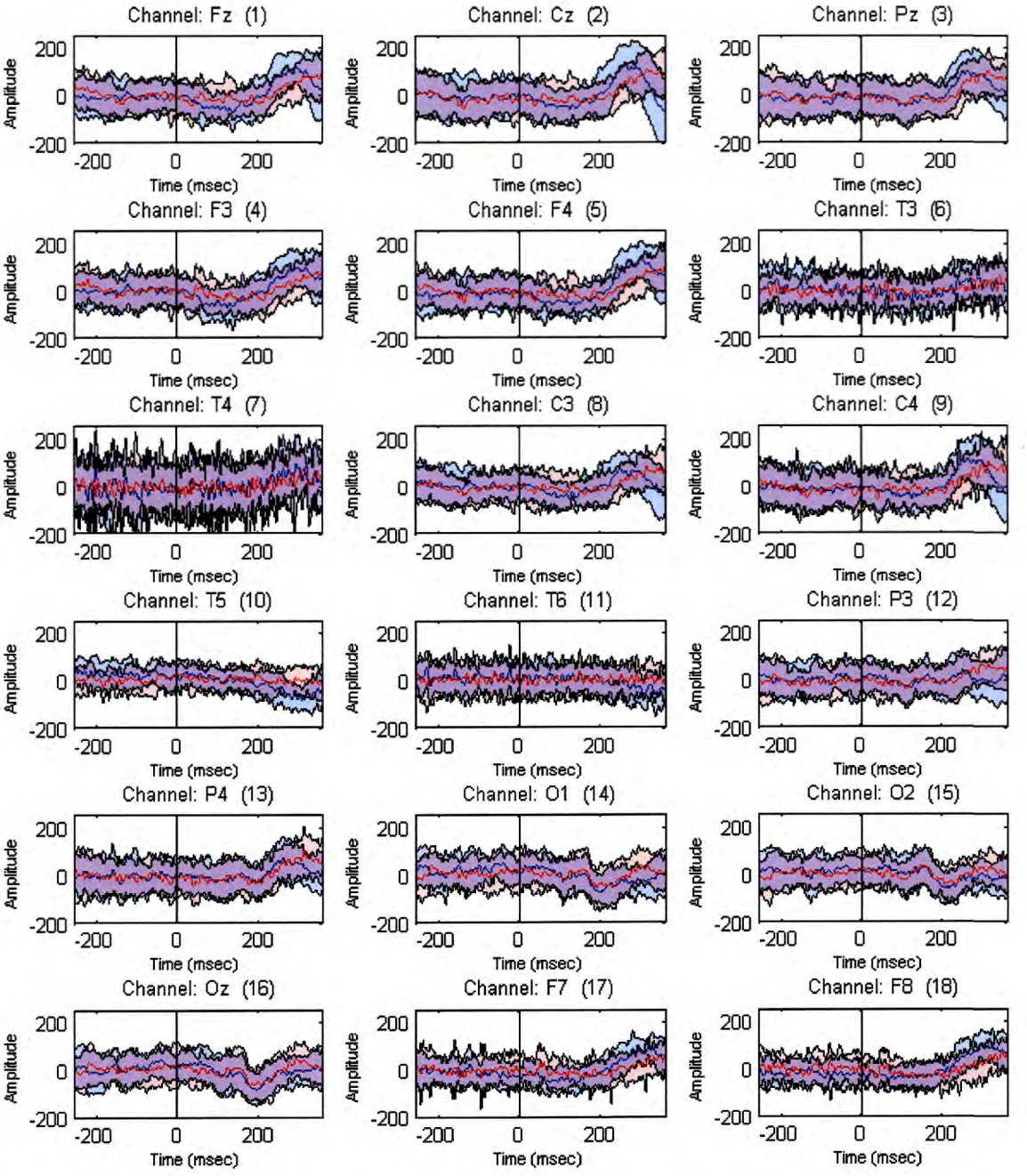


Figure 1.5: Subject 1. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

the EEG signals is proposed. The Continuous Wavelet Transform is used for the acquisition of these representations. We present two novel algorithms for feature

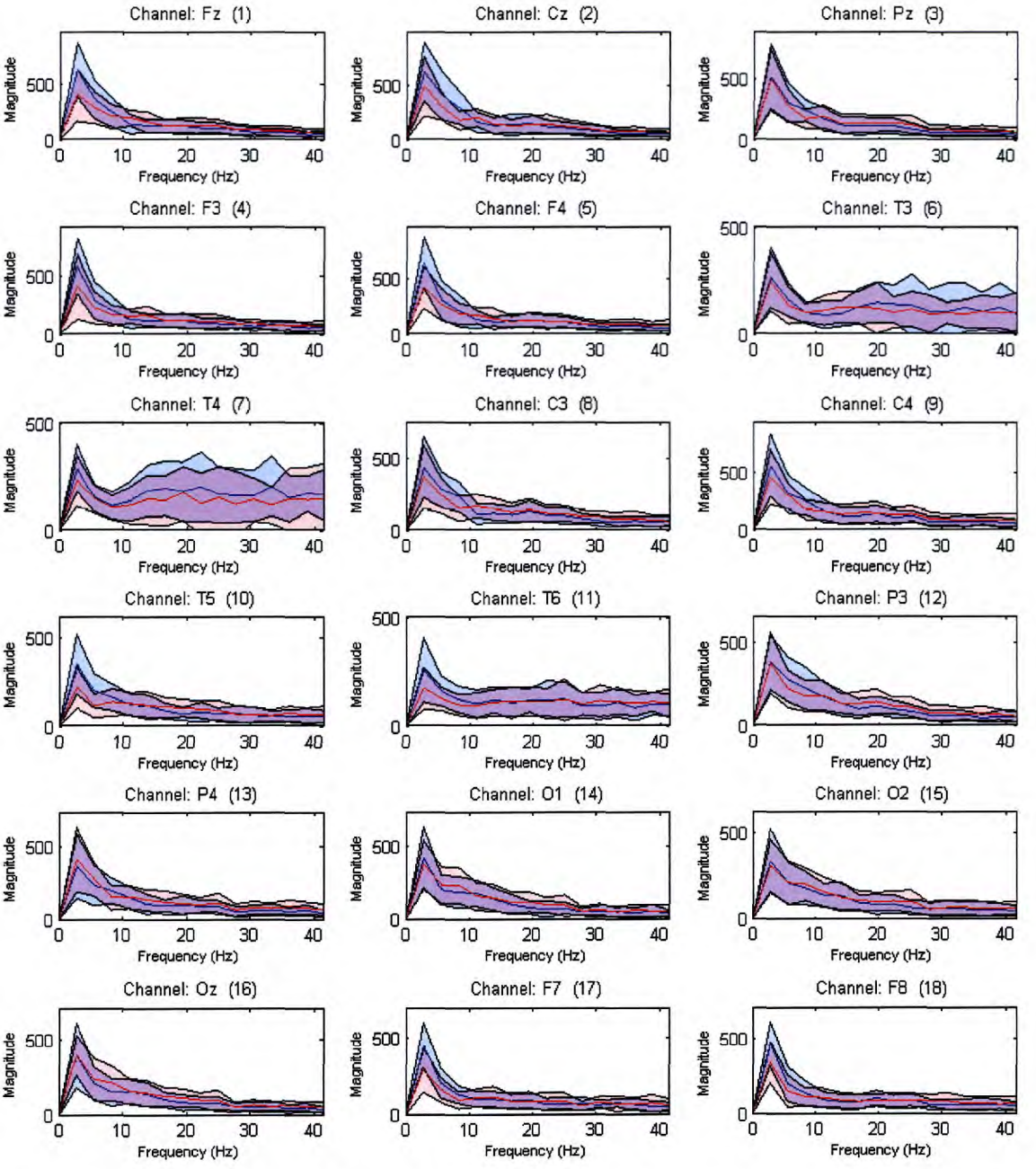


Figure 1.6: Subject 1. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

construction based on the analysis of trials in a single or multi trial basis. The nature of features meant to be captured in both cases is explained. Finally, we

compare the produced classification rates with the ones acquired when the features are constructed using directly the time signals.

- In Chapter 5 we propose a novel algorithm tackling the classification problem using signals constructed in such a way that they can be thought of as characterising each of the two classes. The idea is to construct a characteristic signal for each class to remain as invariant as possible over all data of the class. This is done by constructing a weighted signal for each trial, linearly combining the EEG signals of the various channels and then computing a mean signal for each class using the trials in the training set. The weights of the above combination are chosen so that the variance of the EEG samples over the trials in the training set belonging to the same class is minimised.

  In addition to that, we also propose a novel method for selecting the appropriate channels to be used for the classification purpose. The proposed algorithm is based on stretching and averaging the EEG signals of the available channels, to identify the ones that show the least interference from background activity.

- In Chapter 6, Kernel PCA methods proposed in the Human Performance Monitoring literature are applied to our problem for feature construction. A Support Vector Machine as well as a Gaussian Classifier are used for the classification task. These methods are compared with some of the main methods proposed in this thesis.

- Finally, in Chapter 7, we present the conclusions and describe some of the future work we intend to do.

Chapter 2

Literature survey

In order to predict a person's performance in an oddball experiment using their EEG data, one has to extract suitable features from them. These features should be related to the cognitive processes of recognizing the target stimulus, so as they can be used by a classification method to indicate a quick or late reaction. There are two ways of processing the EEG data to extract these features. The first one is to estimate the ERP from the EEG signals, i.e. the signal of the potential resulting from the activity of the neural subsystems involved in the process of recognizing the stimulus. Then various features can be extracted from it, e.g. the latencies and the amplitudes of its components, and used for the desired prediction. The second way is to skip the first step of enhancing the ERP and directly extract features from the raw EEG signals.

We present first in this chapter the most important techniques in the literature for the estimation of the Evoked Potential (EP) from an EEG signal. The evoked potential is the potential caused by stimulation of the somatosensory system. These methods can be used for any type of EP, as they make very general assumptions for its nature. Thus, they can be used for the estimation of the ERP which is a special type of EP occurring when the external stimulus has to be evaluated by the subject, i.e. the potential is connected with a cognitive process. In the rest of the chapter we review studies from the field of Human Performance Monitoring that use features of the EEG data to estimate a person's performance in an undertaken task.

2.1 Single trial estimation of the Evoked Potential

The most common and widely used model for the observed EEG signal measured after the stimulation of the somatosensory system is the additive noise model. It assumes that the observed signal is the sum of the evoked potential with independent noise. The source of this noise is the background EEG activity which is irrelevant with the occurrence of the stimulus. This model can be expressed as:

$$z = s + v \tag{2.1}$$

where z, s and v are the vectors containing the observed EEG signal, the evoked potential and the background EEG signal, respectively.

If we assume that the evoked potential is a deterministic signal, having the same form each time the somatosensory system of the same person is stimulated in the same way, then we could estimate it by averaging a large number of time locked observations coming from different trials [5]. This of course requires the additional assumption that the background activity is a zero mean stochastic process. Under these conditions it can be proven that averaging is the best estimating technique with respect to the least squares criterion [42].

The assumption, however, that the nature of the evoked potential is deterministic is not a realistic one. There is much evidence indicating that the evoked potential is a stochastic signal and that the variations between its instantiations are related to the differences in the condition of the subject over the trials [13]. In the special case of an ERP, trial variability may reveal changes in the cognitive processes involved in the recognition of the target event. This information is lost through averaging. For this reason, a large variety of techniques have been proposed for the estimation of the evoked potential from a single trial. We review in the rest of this section the most important of them.

2.1.1 Time invariant filtering

Time invariant filtering is the simplest way for estimating the evoked potential. A common method belonging to this category is low pass filtering using digital FIR (finite impulse response) filters. Such examples are the moving average filter proposed in [73] and the especially designed to detect peaks, such as P300, filter in [28]. Bandpass filtering has also been used in [59,72].

Time invariant Wiener filtering has also been proposed for the extraction of the evoked potentials [4,17]. The impulse response of such a filter is calculated minimizing the mean square error between the output of the filter and the desired output (i.e. the evoked potential). For its calculation the autocorrelation function of the input signal (i.e. the EEG signal) and the crosscorelation function between the EEG signal and the evoked potential have to be known. The invariant nature of the impulse response implies the assumption of stationarity for both the evoked response and the background EEG which is not usually true [46,94].

The problem with linear time-invariant filtering is the fact that the spectrum of the evoked potential and the background EEG activity usually overlap. This results in poor estimation of the evoked potential as it has been reported in [46,81].

2.1.2 Time varying filtering

Most of the approaches using time varying filtering for the extraction of evoked potentials are based on Wiener formalism. The first approach was proposed in [24], where the frequency domain of the observed signal is divided into subbands, for each one of which a Wiener-like filter is designed.

A more popular approach is presented in [93] where an optimal time-varying filter for evoked potential estimation is proposed. Assuming the model of Eq. (2.1) an estimate $\hat{s}(t)$ of the evoked potential at time t is computed as:

$$\hat{s}(t) = \boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{z} \tag{2.2}$$

where $\boldsymbol{h}_t$ is the impulse response of the filter at time t and $\boldsymbol{z}$ the observed signal. This means that the estimated evoked potential can be computed as:

$$\hat{\boldsymbol{s}} = \boldsymbol{H}\boldsymbol{z} \tag{2.3}$$

where the rows of $\boldsymbol{H}$ contain the impulse response of the filter at different times. Matrix $\boldsymbol{H}$ is obtained minimizing the mean square error $E\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2\}$. This results in the time varying Wiener filter equation, i.e.:

$$\boldsymbol{H} = \boldsymbol{R}_{\boldsymbol{sz}}\boldsymbol{R}_{\boldsymbol{zz}}{}^{-1} \tag{2.4}$$

where $\boldsymbol{R}_{\boldsymbol{sz}} = E\{\boldsymbol{s}\boldsymbol{z}^{\mathrm{T}}\}$ and $\boldsymbol{R}_{\boldsymbol{ss}} = E\{\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}\}$. Let us note here that if matrix $\boldsymbol{H}$ in Eq. (2.4) is a Toeplitz one, then we have the time invariant version of Wiener filter mentioned in section 2.1.1.

Usually the problem in time varying filtering is the estimation of the crosscorrelation matrix $\boldsymbol{R}_{\boldsymbol{sz}}$. A common procedure to overcome this problem is the use of an analytical model for the evoked potential. Such an approach is proposed in [93] for the construction of a filter called the "time-varying minimum mean square error filter". In that case the evoked potential is modelled as the superposition of signals with random peaks at random latencies. Then a parametric form of the autocorrelation matrix of the evoked potential $\boldsymbol{R}_{\boldsymbol{ss}}$ is calculated. This is used as the desired crosscorrelation matrix because for uncorrelated evoked potential and background EEG activity (with zero mean) we have $\boldsymbol{R}_{\boldsymbol{sz}} = \boldsymbol{R}_{\boldsymbol{ss}}$. The disadvantage, however, of this method is that in order to compute $\boldsymbol{R}_{\boldsymbol{ss}}$ from its parametric form, the probability densities of the peak locations and the means and variances of the peak amplitudes have to be known. Such information does not usually exist before the estimation of the evoked potential. An extention of this approach for the multichannel case is presented in [91].

Different time-varying methods are presented in [50,63]. In both cases the observed signal is divided in three segments. In [63] three different kinds of bandpass filter are used, one for each segment. In [50] a reference signal is first acquired through averaging. Then

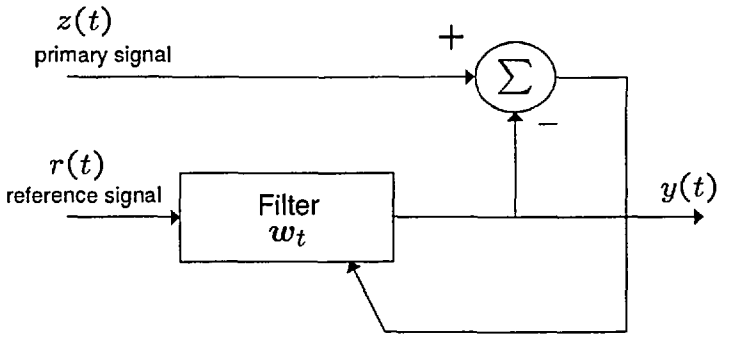


Figure 2.1: The adaptive filtering scheme

a filter is designed for each segment, with its transfer function matched to the spectrum of the reference signal in the corresponding segment.

2.1.3 Adaptive filtering

Adaptive filtering techniques [38] have been extensively used in the processing of the EEG signals for the extraction of the underlying evoked potential. The main advantage of adaptive filters is that they are self-designing filters that adjust their coefficients to track the desired signal. If a Least Mean Squares (LMS) algorithm is used for the update of the coefficients, then no knowledge about the statistics of the signal to be estimated or the existing noise is required. The first use of this kind of filtering for the estimation of evoked potentials was in [83].

The main scheme of the adaptive signal enhancer can be seen in Figure 2.1. The primary input signal, $z(t) = s(t) + v(t)$, is the observed EEG signal modeled as the sum of the evoked potential $s(t)$ and uncorrelated noise (background EEG) $v(t)$. The reference input signal $r(t)$ is a signal closely related to the signal we want to extract from the primary input, i.e. $s(t)$. The output of the filter $y(t) = \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t$ is the desired estimate of the evoked potential at time t. $\boldsymbol{w}_t = [w_t(0), ..., w_t(M-1)]^{\mathrm{T}}$ and $\boldsymbol{r}_t = [r(t), ..., r(t-M+1)]^{\mathrm{T}}$ are the time varying impulse response, of length M, of the adaptive filter and the reference vector, respectively.

The impulse response of the filter can be computed at each time instant using an iterative algorithm. Usually an LMS algorithm is used, minimizing the mean square

error between the output of the filter and the primary input. Using the normalized LMS algorithm [38] the filter coefficients are updated as follows:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \frac{\mu}{\alpha + \|\boldsymbol{r}_t\|^2} \boldsymbol{r}_t (z(t) - y(t)) \tag{2.5}$$

where α is a small positive constant and μ the step-size parameter constrained to $0 < \mu < 2$.

Several approaches have been proposed concerning the choice of the reference signal. In [19, 83] grand averaging is proposed in order to improve the power of the desired signal. In [66] a multireference method is presented where the signals from various physical channels are used as references for the estimation of the somatosensory potential of a peripheral nerve. A different approach is used in [88] where the reference signal is modeled as a combination of sines and cosines whereas in [47] a unit impulse sequence synchronized with the beginning of each iteration is used. Finally, adaptive filtering for the estimation of evoked potentials, has been used in relation with time varying Wiener filtering [95], Autoregressive modeling [60] and Neural Networks [32].

Let us note here that a critique on adaptive filtering is that the LMS algorithm used, is based on second order statistics and therefore is sensitive to the spread of the eigenvalues of the autocorrelation matrix of the reference signal. Moreover, as mentioned in [56], second order statistics fail to remove correlated or coloured noise. For these reasons learning algorithms based on third order statistics have been proposed for the adaptive filtering of evoked potentials [31, 56].

2.1.4 Parametric Modeling

A different approach for the estimation of evoked potentials is through AutoRegressive (AR) and Moving Average (MA) modeling. The additive model of Eq. (2.1) is again assumed. Then the evoked potential is modeled as an Autoregressive-Moving Average (ARMA) filtering of a deterministic signal, which has the evoked potential's pattern. This acts like a reference signal and is usually chosen to be the average of the observed EEG

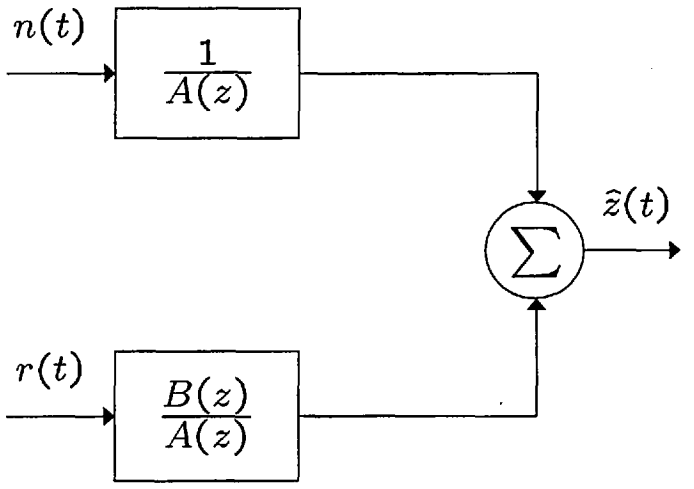


Figure 2.2: The parametric model for the observed EEG

signals of many trials. On the other hand the background EEG is modeled as an AR process. This scheme (see Figure 2.2) is named AutoRegressive model with eXogenous input (ARX) and was first proposed in [16].

The output of the scheme $\hat{z}(t)$, which is the modeled EEG signal $z(t)$, can be written as:

$$\hat{z}(t) = -\sum_{i=1}^{p} a_i \hat{z}(t-i) + \sum_{j=d}^{q+d} b_j r(t-j) + n(t) \tag{2.6}$$

where, p is the order of the AR model, q is the order of the MA model, d is the delay introduced by the MA model, $r(t)$ is the reference signal and $n(t)$ white noise. Let us note here that the AR component, i.e. the coefficients a_i, on the two branches of the model is the same, which compromises the generalization of the model. This is done for reasons of simplicity [16].

The estimated evoked potential $\hat{s}(t)$ can be computed as:

$$\hat{s}(t) = -\sum_{i=1}^{p} a_i \hat{s}(t-i) + \sum_{j=d}^{q+d} b_j r(t-j) \tag{2.7}$$

The values of the parameters (p,q,d) are chosen optimizing the Final Prediction Error [2]. For a specific value of these parameters the coefficients a_i and b_j of the model

are computed using a Batch Least Squares algorithm [27]. This algorithm minimizes the quadratic function Q of the estimation error $e(t) \equiv \hat{z}(t) - z(t)$, where $Q \equiv \frac{1}{N}\sum_{t=1}^{N}[e(t)]^2$, N is the number of samples to be estimated and $y(t)$ the observed EEG signal. The suitability of the computed coefficients is tested by means of the cumulative Anderson test on the whiteness of the error signal $e(t)$ [12]. In [16] a set of parameters is accepted only if $e(t)$ is white with confidence 95%.

Various other approaches using the ARX model have been presented. In [18] an additional ARMA filter is added to the scheme to model the Electro-oculogram (EOG) activity interfered in the recorded EEG signal. Let us mention here that the EOG signal is the electrical activity caused by the movement of a person's eyes and it can interfere with the EEG signal, especially on the electrodes placed on the front of the skull. The actual EOG signal of each trial, recorded on electrodes positioned on the skin near the eye, is given as input to the additional ARMA filter. In [49] the reference signal is first whitened using the inverse transfer function of an AR model which is trained to estimate the reference signal. Then the whitened reference signal is given as an exogenous input to an ARX model. The purpose of the prewhitening is to provide the ARX model with a wide band reference signal and prevent the least squares algorithm from estimating the EEG signal using mainly only the part of the AR modeling of the background EEG. The proposed scheme, which is named Robust-Evoked-Potential-Estimator, appears to achieve considerable Signal to Noise Ratio (SNR) improvements with respect to the original ARX model, in poor initial SNR conditions.

2.1.5 The linear observation model and regularization methods

Various methods of estimating evoked potentials are based on the regularization theory. These methods assume that the evoked potential can be expressed as a linear combination of a suitable, predefined set of vectors. This model is called the linear observation model and can be expressed as:

$$\boldsymbol{z} = \boldsymbol{H\theta} + \boldsymbol{v} \tag{2.8}$$

where z is the observed EEG signal, the columns of $\boldsymbol{H}$ are the predefined set of vectors, θ is the vector of the coefficients and v the background EEG. Thus the evoked potential is modeled as $s = \boldsymbol{H}\theta$. In order to identify the evoked potential, an estimation $\hat{\theta}$ for the vector θ is calculated using the observed data z and the evoked potential is computed as $\hat{s} = \boldsymbol{H}\hat{\theta}$.

Regularization methods come from the theory of ill posed inverse problems [33], where a solution has to be identified such as an arbitrarily small perturbation on the observed data will not cause an arbitrarily large perturbation on the solution. For the identification of this kind of solution regularization methods use, apart from the observed data, initial knowledge that they have about the type of solution. This knowledge is formulated in a mathematical way and is incorporated into the original constraint depending on the data, which has to be minimized.

The most popular regularization approach is Tikhonov's regularization method [37], for solving the weighted least square problem. According to it the estimation for the vector θ of Eq. (2.8) is:

$$\hat{\theta} = \arg\min_{\theta}\{\|\boldsymbol{L}_1(z - \boldsymbol{H}\theta)\|^2 + \alpha^2\|\boldsymbol{L}_2(\theta - \theta_e)\|^2\} \tag{2.9}$$

where $\boldsymbol{L}_1^{\mathrm{T}}\boldsymbol{L}_1 \equiv \boldsymbol{W}_1$, $\boldsymbol{L}_2^{\mathrm{T}}\boldsymbol{L}_2 \equiv \boldsymbol{W}_2$ are positive definite weighting matrices and θ_e is an initial estimate for θ, based on existing knowledge. This solution is called the generalized Tikhonov regularized solution; its constraint is constituted of two parts: the constraint of the ordinary weighted least square solution $\|\boldsymbol{L}_1(z - \boldsymbol{H}\theta)\|$ and the side constraint $\|\boldsymbol{L}_2(\theta - \theta_e)\|$. Parameter α determines the weight given to the side constraint. Usually $\boldsymbol{L}_2$ is chosen to be the second derivative operator which produces a smoother solution compared with the ordinary weighted least squares one.

The analytical form of the solution occurring from the minimization process is:

$$\hat{\theta} = (\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}_1\boldsymbol{H} + \alpha^2\boldsymbol{W}_2)^{-1}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}_1 z + \alpha^2\boldsymbol{W}_2\theta_e) \tag{2.10}$$

A special case for the selection of the columns of $\boldsymbol{H}$ is to be chosen as the first p

eigenvectors of the correlation matrix of the observed signals $\boldsymbol{z}$. For a prefixed amount of basis vectors, this choice minimizes the mean square error between $\boldsymbol{z}$ and $\hat{\boldsymbol{s}} = \boldsymbol{H}\hat{\boldsymbol{\theta}}$, when $\hat{\boldsymbol{\theta}}$ is calculated using the mean square solution, i.e. $\hat{\boldsymbol{\theta}} = (\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{z}$ and the elements of $\boldsymbol{z}$ are jointly Gaussian [44]. Let us call $\mathcal{S}$ the subspace spanned by the selected eigenvectors and $\boldsymbol{H}_{\mathcal{S}}$ the corresponding matrix. Then the linear observation model of Eq. (2.8) takes the form:

$$\boldsymbol{z} = \boldsymbol{H}_{\mathcal{S}}\boldsymbol{\theta} + \boldsymbol{v} \tag{2.11}$$

In [43] the model of Eq. (2.11) is used for the estimation of evoked potentials. For the estimation of $\boldsymbol{\theta}$ the Gauss-Markov (minimum variance) solution is used, i.e.

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{H}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{C}_v^{-1}\boldsymbol{H}_{\mathcal{S}})^{-1}\boldsymbol{H}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{C}_v^{-1}\boldsymbol{z} \tag{2.12}$$

which can be obtained from Eq. (2.10) setting $\boldsymbol{W}_1$ equal to the covariance matrix $\boldsymbol{C}_v$ of the background EEG and $\boldsymbol{W}_2 = 0$. The model of Eq. (2.11) has also been used in [23,49] using the least squares solution for the estimation of $\boldsymbol{\theta}$, i.e. $\hat{\boldsymbol{\theta}} = (\boldsymbol{H}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{H}_{\mathcal{S}})^{-1}\boldsymbol{H}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{z} = \boldsymbol{H}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{z}$. This can be derived from Eq. (2.12) with the white background EEG assumption $\boldsymbol{C}_v = \sigma^2\mathbf{I}$. When the least square solution is used in combination with the model of Eq. (2.11) we have the principal component regression method. Let us note here that all these methods are not regularization methods as the side constraint is not actually used.

In [44] a regularization method which combines the model of Eq. (2.8) with the space spanned by the columns of $\boldsymbol{H}_{\mathcal{S}}$ is presented. This method belongs to the category of subspace regularization methods. It assumes that the evoked potential belongs to a space spanned by the columns of $\boldsymbol{H}$ but is close to $\mathcal{S}$. The columns of $\boldsymbol{H}$ are chosen in some generic way, e.g. a set of Gaussian shaped vectors with different delays and preselected widths. Then the Tikhnov's solution of Eq. (2.9) is used with $\boldsymbol{L}_1^{\mathrm{T}}\boldsymbol{L}_1 = \boldsymbol{C}_v^{-1}$, $\boldsymbol{L}_2 = (\mathbf{I} - \boldsymbol{H}_{\mathcal{S}}\boldsymbol{H}_{\mathcal{S}}^{\mathrm{T}})\boldsymbol{H}$ and $\boldsymbol{\theta}_e = 0$. Let us note that this produces a side constraint of the form $\|(\mathbf{I} - \boldsymbol{H}_{\mathcal{S}}\boldsymbol{H}_{\mathcal{S}}^{\mathrm{T}})\boldsymbol{H}\boldsymbol{\theta}\|$, which is the distance of the evoked potential $\boldsymbol{s} = \boldsymbol{H}\boldsymbol{\theta}$ from $\mathcal{S}$. This is in accordance with the assumption that $\boldsymbol{s}$ is close to $\mathcal{S}$. Using Eq. (2.10) the

evoked potential is estimated as:

$$\hat{s} = H(H^{\mathrm{T}}C_v^{-1}H + \alpha^2 H^{\mathrm{T}}(\mathbf{I} - H_S H_S^{\mathrm{T}})H)^{-1}H^{\mathrm{T}}C_v^{-1}z \tag{2.13}$$

2.1.6 Wavelets

A promising tool for the extraction of evoked potentials is wavelets. The Wavelet Transform (WT), introduced in [34], enables the decomposition of a signal in a set of functions, the energy of which is localized in both time and frequency domains. Its main advantage is that the length of the time period where the energy of the decomposing functions is localized, varies depending on the range of frequencies of the function. More precisely, the length of the time window is enlarged in low frequencies and becomes smaller for the high ones, providing in this way any desirable trade-off between the time and the frequency resolution.

The usual procedure for the enhancing of the EP using the WT is to compute the coefficients of the observed EEG signal and then manipulate them according to various criteria, before using them to reconstruct the enhanced EP. For the analysis and synthesis process the concept of Multiresolution Analysis [61] is used. The first use of wavelets in EP extraction is in [6]. The authors decompose contaminated auditory EP and pure EEG signals and then use correlation and discriminant analysis to find the scales, the coefficients of which best discriminate the two types of signal. The more discriminative coefficients of the selected scales are used to reconstruct the "clear" EP.

In [68,69] the average signal of the single trial EPs/ERPs is first computed and decomposed using the WT. The coefficients with high value, within a predefined time range, are identified for each scale. Then the EP/ERP is extracted from the signal of each trial by decomposing it using the WT and reconstructing it using only the identified coefficients. In [11] the WT coefficients are manually selected according to a priori knowledge for the time-frequency content of the general pattern of the somatosensory/auditory evoked potentials which are enhanced. The authors also use the WT to propose a time-frequency extension of the a posteriori Wiener filter [89]. An adaptive filter using Multiresolution

Analysis is also proposed in [9] for the extraction of auditory EPs. In [75] the use of a mother wavelet function especially designed to match the shape of the evoked potential to be extracted, is proposed. The technique presented in [20] is used for the design of the mother wavelet function, which is a member of the class of Meyer wavelets

2.2 The use of ERP signals in human performance monitoring

There are a number of fields relating a person's EEG data with their ongoing cognitive processes in order to satisfy the needs of a specific application. The field closer to our problem is that of *Human Performance Monitoring*. The applications of this field try to estimate a person's performance in a specific task using their EEG data at the time. In this section we review the most important studies of the field.

In [85] the authors study the relation between the ERP and a person's performance in three visual display-monitoring tasks: signal detection, running memory and computation. The difficulty of each task is designed to vary resulting in a variation in the persons' performance. The indexes of performance are the reaction time, the confidence and the accuracy. A global measure for the characterization of performance, the Performance Factor (PF), is designed as a linear combination of the above indexes. Then, linear regression models are developed to relate the ERPs (specifically measures such as the amplitude and the latency of the ERP components) with the corresponding PFs. The models are distinguished by three factors: single subjects versus inter-subject based approach, relevance of the stimulus and SNR. The relevant stimuli are the ones connected with the undertaken task. The SNR of the ERPs are enhanced by computing the running average of ten consecutive ERPs. In such cases the ERP signals are related to the corresponding averages of the PFs. The results show that single subject based linear regression models, using ERPs elicited by the relevant stimuli and having sufficiently large SNR, can reliably estimate a person's performance factor.

The data from the signal detection task of [85] are processed in different ways

in [45, 70, 71, 86] to develop a model capable of providing reliable estimations of a person's performance in the task. In [45] ERP features are constructed through Principal Component Analysis (PCA) and modelling the observed ERP as an autoregressive (AR) process. The constructed features are related to human's performance using linear regression and Radial Basis Function (RBF) networks. The superiority of the combination of PCA features with the use of RBF networks for modeling is reported.

The use of high power coefficients of the Discrete Wavelet Transform (DWT) for the construction of features is investigated in [86]. The authors report that the use of DWT coefficients in linear regression models lead to the same performance with PCA coefficients using half as many free parameters. This results in models more resistant to overfitting and thus capable of generalizing better to new data. The superiority of DWT features over PCA coefficients and raw ERP data is also reported when neural networks are used as models for estimating the human's performance. This superiority of the DWT features is discussed as far as decorrelation and energy compaction properties are concerned. Finally, evidence that low frequency (delta band) activity, occurring at specific times and scalp locations, correlates with signal detection performance is presented.

In [70, 71] the use of Kernel PCA is proposed as a feature extraction method for a human's performance monitoring application. Kernel PCA [77] maps the observed feature vector $\boldsymbol{x}$ into a space $\mathcal{F}$ of higher dimension, using a non linear function $\phi(\boldsymbol{x})$. The dot product in $\mathcal{F}$ is computed through a kernel function, i.e. $K(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}).\phi(\boldsymbol{y})$. After the mapping, standard (linear) PCA is performed in $\mathcal{F}$ and the components across which high variance is achieved are selected for projection. The advantage of Kernel PCA over linear PCA is that higher order correlations between the selected variables are achieved and that more variables can be extracted in case the input data are more than the dimension of the original space. In the above studies Kernel PCA with Gaussian kernel is compared with linear PCA using Support Vector Regression and Kernel types of Regression and its superiority is reported.

Let us mention here that all the above studies use averages of the observed EEG signals in order to enhance the buried ERP. This restricts them to estimating the average

performance in a period of consecutive trials. The preprocessing of the data using the methods described in section 2.1, in order to enhance the single trial ERP and permit the performance estimation in single trial, basis is of great interest.

Chapter 3

Classification using dimensionality reduction and subspace methods

In this chapter we perform classification using very simple features for the characterisation of the subject's response in each trial and concentrate on a number of transforms of the initial feature space that are likely to enhance the separability of the two classes.

We begin performing classification on the initial feature space. This is the simplest way to approach the problem and the results achieved will be used as benchmark to evaluate the efficiency of more sophisticated methods. We then try to increase the classification results reducing the amount of features used, by selecting those with the highest discriminating capability. The latter is computed by the data of the trials in the training set. Then we use a number of methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Non Negative Matrix Factorization (NMF) to construct a new space of lower dimension for the representation of the features and investigate its suitability for the desired classification. Finally, we perform classification using PCA and NMF to construct different subspaces for the two classes.

3.1 Construction of the set of features

We propose here an algorithm that can be used for constructing the initial set of features for the signals of each trial. When feature construction is done for the purpose of classification, it is important to isolate from the elements of a class those characteristics that make them different from the elements of the other classes. In our case, we search those characteristics in the magnitude of the Fourier coefficients that are known to be related with the cognitive processes of recognising a target (see Section 1.2.3). Moreover, the magnitude of the frequency components meets the non negative requirement of NMF, so its use for processing the constructed features is feasible.

The proposed algorithm constructs the desired features as the magnitude of the Discrete Fourier Transform (DFT) of the EEG signals. The magnitude of the Fourier coefficients has been used as feature for discriminating between two different mental tasks in [57].

In our case the feature vector $\boldsymbol{f}_t$ of the t^{th} trial is constructed as follows.

- We first compute the mean value of vectors $\boldsymbol{x}_{n,t}$ and subtract it from their elements. Vector $\boldsymbol{x}_{n,t}$ contains the time samples of the EEG signal between the stimulus onset and the quickest reaction time of the subject, as recorded by the n^{th} channel.

- For each of the N channels used, we compute the DFT of the corresponding vector $\boldsymbol{x}_{n,t}$ containing the EEG signal. This is computed as:

$$\widehat{x}_{n,t}(k) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} x_{n,t}(m) e^{-j2\pi \frac{(k-1)(m-1)}{M}}, \quad k = 1, ..., M.$$

- We construct vector $\widehat{\boldsymbol{x}}_{n,t}$ containing the magnitude of the positive digital frequencies corresponding to an analog frequency smaller than 40 Hz. This is because the majority of the energy of an EEG signal is concentrated in the frequencies smaller than 40 Hz (see Table 3.1). Moreover it is the band 0-40 Hz of the EEG spectrum which is mainly known to be connected with a person's ongoing cognitive processes [7]. We exclude from these frequencies the one of 0 Hz as this is always equal to zero

Table 3.1: Mean percentage (%) of the EEG signals' energy in the range 0-40Hz for the 11 subjects in the 18 channels

Subject	1	2	3	4	5	6	8	10	11	13	14
Ch. Fz	99.84	99.34	98.83	98.4	99.01	99.15	99.33	99.23	98.2	99.55	99.82
Ch. Cz	99.79	98.62	99.78	98.86	98.7	98.12	98.74	98.73	99.65	99.36	97.8
Ch. Pz	99.85	99.02	99.84	99.23	99.03	99.65	97.74	99.5	98.87	99.74	99.57
Ch. F3	99.54	99.32	97.25	97.52	98.97	97	99.4	99.38	98.62	98.24	98.8
Ch. F4	98.05	97.91	98.92	97.17	99.2	94.75	98.7	98.31	99.64	99.78	98.84
Ch. T3	99.22	97.89	99.8	98.73	92.8	87.35	94.77	97.62	97.68	98.11	99.62
Ch. T4	92.42	98.32	97.99	99.2	97.47	92.31	97.33	98.7	95.53	98.08	94.74
Ch. C3	99.15	98.53	99.81	99.15	98.41	99.37	97.75	98.66	99.84	99.3	99.87
Ch. C4	99.8	99.55	99.39	99.03	98.93	99.34	98.19	98.68	93.63	99.78	99.9
Ch. T5	99.16	99.91	98.83	98.87	95.15	96.36	98.41	99.34	98.88	98.69	99.19
Ch. T6	99.34	98.62	98.65	98.97	97.5	98.82	96.96	99.03	99.66	98.87	97.74
Ch. P3	99.9	99.48	98.22	92.75	97.45	99.13	99.02	99.5	97.26	99.04	99.67
Ch. P4	99.89	98.53	99.76	97.35	97.03	99.76	97.73	98.55	95.27	99.48	99.87
Ch. O1	99.79	99.77	98.98	98.21	99.04	98.51	98.41	97.33	95.98	99.66	99.09
Ch. O2	99.54	99.01	99.65	98.66	99.16	98.86	97.62	98.56	96.33	98.5	99.78
Ch. Oz	99.56	99.58	98.63	98.27	98.94	97.23	97.89	99.08	95.83	99.29	98.28
Ch. F7	98.61	99.81	97.81	98.12	99.33	96.77	96.94	96.08	96.37	97.24	99.62
Ch. F8	99.38	98.41	96.56	99.28	97.52	97.02	94.11	97.6	95.73	98.29	93.51

because of the subtraction of the mean value of the signal.

Let us note that a sample $\widehat{x}(k)$, with $k = 1, ..., \lfloor \frac{M}{2} \rfloor + 1$ since we are interested only in positive frequencies, corresponds to the continuous frequency $\frac{k-1}{M} F_s$. We repeat here that $F_s = 500$ Hz is the sampling frequency of the continuous EEG.

- The feature vector $\boldsymbol{f}_t$ is constructed by concatenating vectors $\widehat{\boldsymbol{x}}_{n,t}$ across channels, i.e. $\boldsymbol{f}_t = (\widehat{\boldsymbol{x}}_{1,t}; ...; \widehat{\boldsymbol{x}}_{N,t})$.

3.2 The Gaussian Classifier

In most cases in this chapter the Gaussian classifier [26] is used for the purpose of classification. We give here a brief description of the way that this classifier works.

Let us assume that we have a training set which is constituted from N dimensional data samples belonging to M different classes. We denote with $\boldsymbol{x}_i$ the random vector of the i^{th} class. The data samples in the training set are instantiations of these random

vectors. The Gaussian classifier assumes that $\boldsymbol{x}_i$ follow a multivariate normal distribution, which means that the probability density function of each one is given by:

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{R}_i|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{R}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i)\right], \quad i = 1,...,C \tag{3.1}$$

where $\boldsymbol{m}_i$ and $\boldsymbol{R}_i$ are the mean vector and the covariance matrix of $\boldsymbol{x}_i$. In the following sections these quantities are calculated from the samples in the training set.

A sample $\boldsymbol{x}_t$ in the testing set is assigned by the classifier to the i^{th} class C_i if:

$$P(C_i|\boldsymbol{x} = \boldsymbol{x}_t) = \max_{j=1,...,M} P(C_j|\boldsymbol{x} = \boldsymbol{x}_t). \tag{3.2}$$

Using Bayes' theorem:

$$P(C_j|\boldsymbol{x} = \boldsymbol{x}_t) = \frac{P(\boldsymbol{x} = \boldsymbol{x}_t|C_j)P(C_j)}{P(\boldsymbol{x} = \boldsymbol{x}_t)} \tag{3.3}$$

Assuming that each sample has the same a-priori probability of belonging to any class we have that:

$$\arg\left\{\max_{j=1,...,M} P(C_j|\boldsymbol{x} = \boldsymbol{x}_t)\right\} = \arg\left\{\max_{j=1,...,M} P(\boldsymbol{x} = \boldsymbol{x}_t|C_j)\right\} = \arg\left\{\max_{j=1,...,M} p_j(\boldsymbol{x}_t)\right\} \tag{3.4}$$

Thus using Eq. (3.2), (3.4) and the negative ln of Eq. (3.1) a sample $\boldsymbol{x}_t$ is assigned to class C_i if:

$$L_i(\boldsymbol{x}_t) = \min_{j=1,...,M} L_j(\boldsymbol{x}_t) \tag{3.5}$$

where $L_i(\boldsymbol{x})$ is defined for class C_i as:

$$L_i(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{R}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i) + \ln|\boldsymbol{R}_i| \tag{3.6}$$

In some of the cases in this chapter we assume that the covariance matrices of the classes have specific forms which make Eq. (3.6) having a simpler form. In the case that the random variables $x_{i,k}, k = 1,...,N$ of each random vector $\boldsymbol{x}_i$ are uncorrelated, Eq.

(3.6) reduces to:

$$L_i(\boldsymbol{x}) = \sum_{k=1}^{N} \frac{1}{{\sigma_{i,k}}^2}(x_k - m_{i,k})^2 + \ln\left(\prod_{k=1}^{N} {\sigma_{i,k}}^2\right) \tag{3.7}$$

where $m_{i,k}$ and $\sigma_{i,k}^2$ are the mean and the variance of $x_{i,k}$, respectively. Moreover if we assume that the variance of $x_{i,k}$ is the same in all classes and equal with σ_k^2, we take obtain:

$$L_i(\boldsymbol{x}) = \sum_{k=1}^{N} \frac{1}{{\sigma_k}^2}(x_k - m_{i,k})^2 \tag{3.8}$$

where the second term in Eq. (3.7) was omitted because it is the same for all classes and plays no role in the classification process.

3.3 Classification in the initial feature space

We present in this section the classification results we get performing classification in the initial feature space. The set of features characterising each trial is constructed as described in Section 3.1, i.e. using the magnitude of the frequencies smaller than 40Hz. In addition to that, we construct the features here in two more ways: a)keeping the magnitude of all frequencies up to 250Hz and b)using the time samples instead of the magnitude of the spectrum. In the case of time samples two ways are used for the classification task. The first one is the Gaussian classifier, as done with the previous types of features. The second one is through correlation of the testing time signals with the mean time signal of each class as computed from the trials in the training set.

The classification process takes place with respect to each subject separately. The trials of each subject are divided in two sets in a random way and the two subsets created are used as training and testing sets. We make sure that there is an equal number of trials from each class in each one of the two sets. The processes of the random splitting of the available trials and the one of subsequent classification are repeated 500 times in order to reduce the variability of the results. The classification results are computed as the average of the classification results of each repetition.

As far as the classification task is concerned using the Gaussian classifier and the

criterion of Eq. (3.6) we have the problem that the covariance matrix of each class is not invertible. This is because the dimensionality of the feature space is larger than the available samples of each class in the training set. The number of trials in the training set of each subject and the dimensionality of the feature space for each one of the feature construction algorithms can be seen in Table 3.2. In order to overcome the problem of the non invertibility of the covariance matrices we assume that the features are mutually uncorrelated and use Eq. (3.7) for the definition of $L_i(\boldsymbol{x})$. In a second round of experiments the variance of the features is assumed the same for all classes and Eq. (3.8) is used.

The correlation type of classification, which is used only for the time features, can be described as follows. A mean signal is first computed for each class using the signals of the trials in the training set. The mean signals are normalised to have energy equal with one. In order to classify each trial in the testing set, the energy of its corresponding signal is normalised and then the inner product of the normalised signal with the mean signal of each of the two classes is computed. Let as denote with a the correlation of the testing signal with the mean signal of class success, with b the corresponding correlation with class "failure" and with d their difference, i.e. $d = a - b$. Then the testing trial is classified to class "success" if $d > t$, to class "failure" if $d < t$ and a random choice is made if $d = t$, where t is a suitably chosen threshold. The value of t is evaluated in the following way. We repeat the correlation process described earlier for each trial in the training set, computing each time the mean signals of the two classes using the signals of the remaining trials. We denote with d_s the value of difference d when the trial belongs to class "success" and with d_f when it belongs to class "failure". We then compute the mean values of d_s and d_f over the trials belonging to class "success" and "failure" respectively and the value of the threshold t is computed as $t = \frac{(\overline{d_s}+\overline{d_f})}{2}$.

The rate of correct classification for each combination of feature construction algorithm and type of classification are presented in Figures 3.1 and 3.2. We can see that the correct classification rate varies across subjects, with the best being achieved for subject 14 and the worst for subject 3. In the majority of cases the time domain features achieve the best classification accuracy. As far as the accuracy of the features constructed from

the magnitude of the spectrum is concerned, the algorithm keeping only the frequencies smaller than 40Hz outperforms the one using all the available frequencies. This is something we expected as the EEG signal is a lowpass signal so the high frequencies contain mostly noise which deteriorates the classification. The only exception is subject 6. This can be explained observing the magnitude of the spectrum of subject 6 across frequencies and noticing that in contrast with the other subjects, subject 6 has a significant amount of energy in frequencies larger than 40Hz. Finally, we observe that in most cases the classification accuracy is better when the variance of the features is assumed the same for the two classes. This is probably because the variance of the features is the same for the two classes in the majority of cases, something which we fail to estimate because of the small size of the testing set. This can be seen in Figures 1.5, A.1,-A.10 and 1.6, A.11,-A.20, in which the variance of each class is computed from all the available trials of each class (i.e. both testing and training sets). Thus, in the majority of the experiments following in this chapter, the features will be assumed to have the same variance in the two classes.

In order to check whether the difference in the estimated classification rates of two algorithms is large enough to permit us conclude about the superiority of one of them, we perform the following test of significance. We compute the observed level of significance of the hypothesis that the algorithm with the larger estimated classification rate has a true classification rate equal or smaller than the other algorithm. If this level of significance is small enough ($< 5\%$) then the difference in the classification rates is significant and we can assume that the algorithm with the larger estimated classification rate is superior. We refer to Appendix C for more details concerning the computation of the observed level of significance.

We performed this test here for the algorithm having the best estimated classification rate for each subject (see Figures 3.1 and 3.2) against all other algorithms and the results are presented in Table 3.3. We denote in bold the cases for which the observed level of significance is smaller than 5%, i.e. the cases we can be confident (at a level of 95%) that the algorithm with the best estimated classification rate is actually superior. One can see that in most cases the significance level is not small enough to give us such

Table 3.2: Number of trials of the same class in the training set and dimension of the feature space per subject

Subject		1	2	3	4	5	6	8	10	11	13	14
Number of trials per class in the training set		14	12	9	23	20	25	9	16	16	16	12
Dimens. of initial feature space	Spectrum (<40Hz)	270	432	324	324	306	306	270	342	324	342	396
	Spectrum	1620	2664	1944	1926	1854	1908	1656	2106	1962	2052	2430
	Time	3240	5328	3888	3852	3708	3816	3312	4212	3924	4104	4860

confidence. However, we have to note here that because of the small number of available trials, a small difference in the true classification rate of two algorithms is not likely to be adequate to support the superiority of one of them with such a high level of confidence.

In the remaining sections of this chapter we use only the magnitude of the frequencies smaller than 40Hz as features. Although the time samples give in general better results in the initial features space, the magnitude of the spectrum has a number of characteristics that makes it more appealing for this study. First of all, the resulting number of features is much smaller which is an advantage especially in our case that the size of the testing set is small. Moreover the positive nature of the features will enable us to use the Non Negative Matrix Factorization for their processing. Finally, since the magnitude of the spectrum is known to be connected with various cognitive processes, we want to investigate the possibility of a suitable transformation to exploit this connection and boost the classification accuracy in our problem.

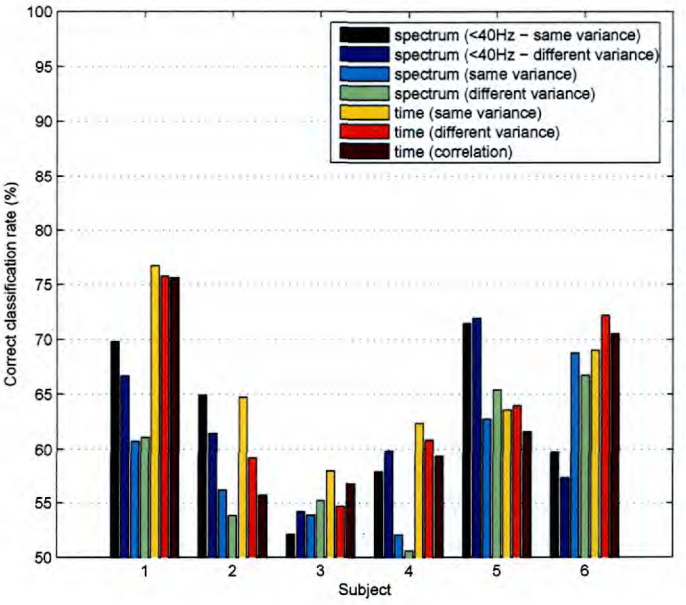


Figure 3.1: Classification accuracy for subjects 1,2,3,4,5,6.

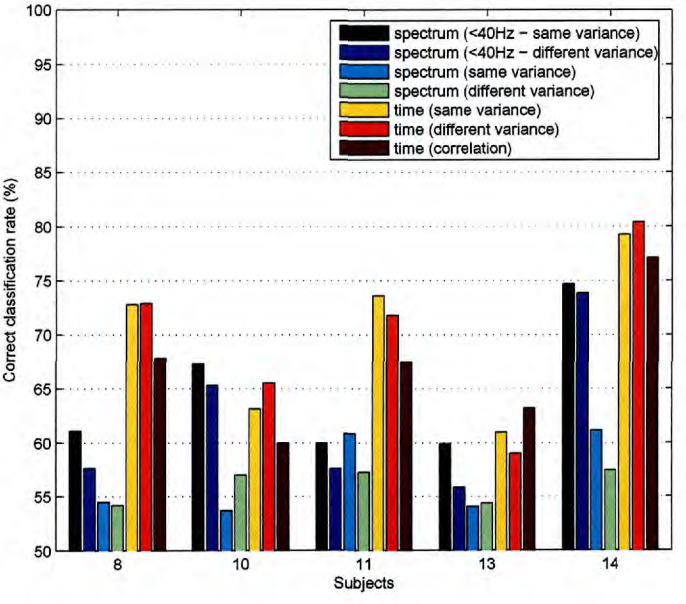


Figure 3.2: Classification accuracy for subjects 8,10,11,13,14.

Table 3.3: Observed level of significance (%) for the hypothesis that the algorithm with the best estimated classification rate for each subject is equivalent or inferior than the other algorithms.

Subjects	Algorithms						
	Spectrum (0-40Hz)		Spectrum (0-250Hz)		Time		
	same variance	different variance	same variance	different variance	same variance	different variance	correlation
1	20.49	11.93	**3.37**	**3.66**	-	45.27	44.65
2	-	36.56	19.16	13.55	49.26	28.42	17.91
3	31.19	37.61	36.69	40.98	-	39.21	46.07
4	26.49	36.35	7.78	5.27	-	41.55	33.87
5	47.41	-	11.02	18.78	13.18	14.23	8.41
6	**3.13**	**1.34**	29.9	20.19	31.12	-	39.76
8	14.53	8.7	5.15	**4.87**	49.54	-	32.07
10	-	40.58	5.68	11.47	30.99	41.59	19.47
11	5.03	**2.73**	6.2	**2.5**	-	40.94	22.38
13	35.11	19.87	14.75	15.48	39.73	31.37	-
14	25.26	22.34	**1.79**	**0.66**	44.6	-	34.92

3.4 Classification selecting the most discriminating features

In this section we investigate the possibility of increasing the classification accuracy keeping only a subset of the constructed features. These features are selected with respect to their discriminating capability, with the latter being measured from the data in the training set.

We begin first by reducing the number of channels used. We use the following three ways to measure the discriminating capability of a channel:

- The Bhattacharyya distance between the classes "success" and "failure" which is measured as:

$$d = \frac{1}{8}(\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}} \left(\frac{\boldsymbol{R}_1 + \boldsymbol{R}_2}{2}\right)^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2) + \frac{1}{2} \ln \frac{\left|\frac{\boldsymbol{R}_1+\boldsymbol{R}_2}{2}\right|}{\sqrt{|\boldsymbol{R}_1||\boldsymbol{R}_2|}} \tag{3.9}$$

  where $\boldsymbol{m}_1$, $\boldsymbol{m}_2$, $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are the mean vectors and the covariance matrices of the feature vectors of the two classes constructed from the corresponding channel.

- The Bhattacharyya distance between the two classes, assuming that they both have

the same covariance matrix $\boldsymbol{R}_c$. This makes Eq. (3.9):

$$d = (\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}} \boldsymbol{R}_c^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{3.10}$$

- The Euclidean distance between the mean feature vectors of each class, i.e:

$$d = (\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}} (\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{3.11}$$

Obviously the larger the distance is between the two classes for a specific channel, the larger its discriminating capability will be. We perform the classification experiments splitting the set of available trials into two halves and using them as training and testing sets. We use the trials in the training set to measure the discriminating capability of each channel. Then 18 sets of channels are constructed with the n^{th} set, $n = 1, ..., 18$, containing the n channels with the largest discriminating capability. The classification experiments are performed for each set of channels separately, using each time only the available channels for the feature construction of each trial. The whole process is repeated 500 times for randomly different compilations of the training and testing sets and the classification accuracy is measured through averaging. Let us repeat here that the features are constructed according to the process described in Section 3.1.

Because of the small size of the training set, the features in each channel are assumed mutually uncorrelated to ensure the invertibility of the covariance matrices in Eq. (3.9), (3.10) and (3.11). When Eq. (3.10) and (3.11) are used for the computation of the distance between the two classes, Eq. (3.8) is used for the classification task (same variance for each feature in the two classes). On the other hand, when Eq. (3.9) is used for the computation of the distance between the two classes, Eq. (3.7) is used for the classification task (different variance for each feature in the two classes). The results of correct classification can be seen in Figure 3.3.

One can observe that in some cases (e.g. subjects 2, 3, 14) the classification accuracy is increased for a reduced number of channels. In the majority of cases the best results are produced when the Euclidean distance is used to estimate the discriminating

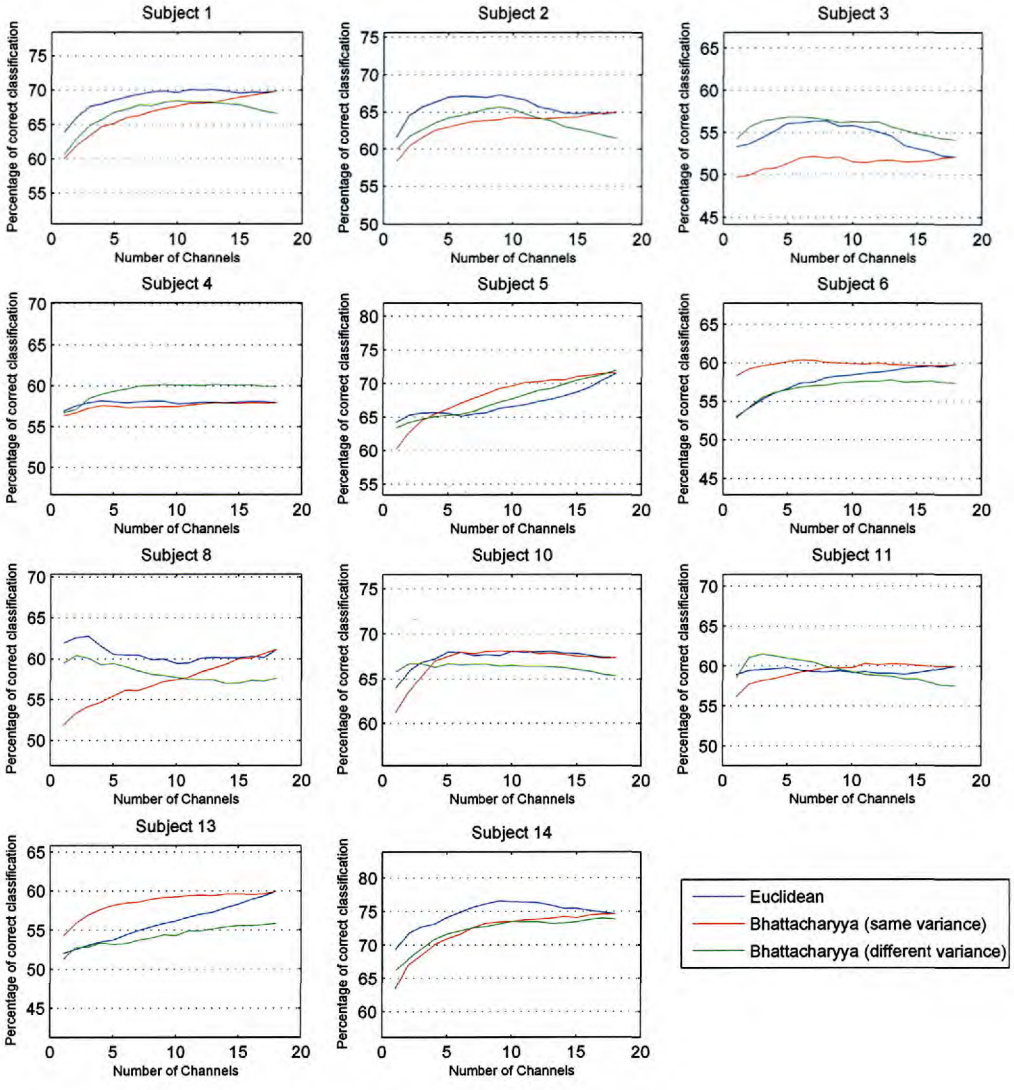


Figure 3.3: Classification accuracy as a function of the number of channels used.

Table 3.4: Number of channels achieving maximum classification accuracy per subject

Subject	1	2	3	4	5	6	8	10	11	13	14
Number of channels	11	9	7	4	18	18	3	13	18	18	9

capability of the channels. This is probably because of the inadequate estimation of the features' variances from the training set. The number of channels that achieve maximum classification accuracy per subject, when Euclidean distance is used, can be seen in Table 3.4.

We next compute the distance between the two classes for each feature separately, i.e. for the magnitude of each frequency (smaller than 40 Hz) of the signal recorded on each channel. Eq. (3.9), (3.10) and (3.11) are again used but this time the mean vectors are the mean values of the feature and the covariance matrices are its variances. The classification experiments are repeated in the same way as above, using this time a reduced number of features achieving maximum distance between the two classes. In the beginning we select a number of channels, according to Table 3.4, having the maximum Euclidean distance and take into account only the features constructed from these channels. In a second round of experiments, the features from all channels are taken into account. The results of correct classification across the number of features that are used can be seen in Figure 3.4.

One can observe that in the vast majority of subjects (all subjects except 5 and 13) the classification accuracy is increased for a reduced number of features comparing with the classification accuracy when all features are used. This effect is more prominent for subject 8, something that agrees with Figure A.16 in which one can see that there are specific frequencies much more suitable for separating the two classes than others. Moreover, let us note here that the Euclidean distance is proven again more suitable for evaluating the discriminating capability of the features. Finally, we can see that reducing first the number of channels to the one giving better results and then reducing the number of features gives slightly better results compared with the case that a reduced number of features are selected from all the channels.

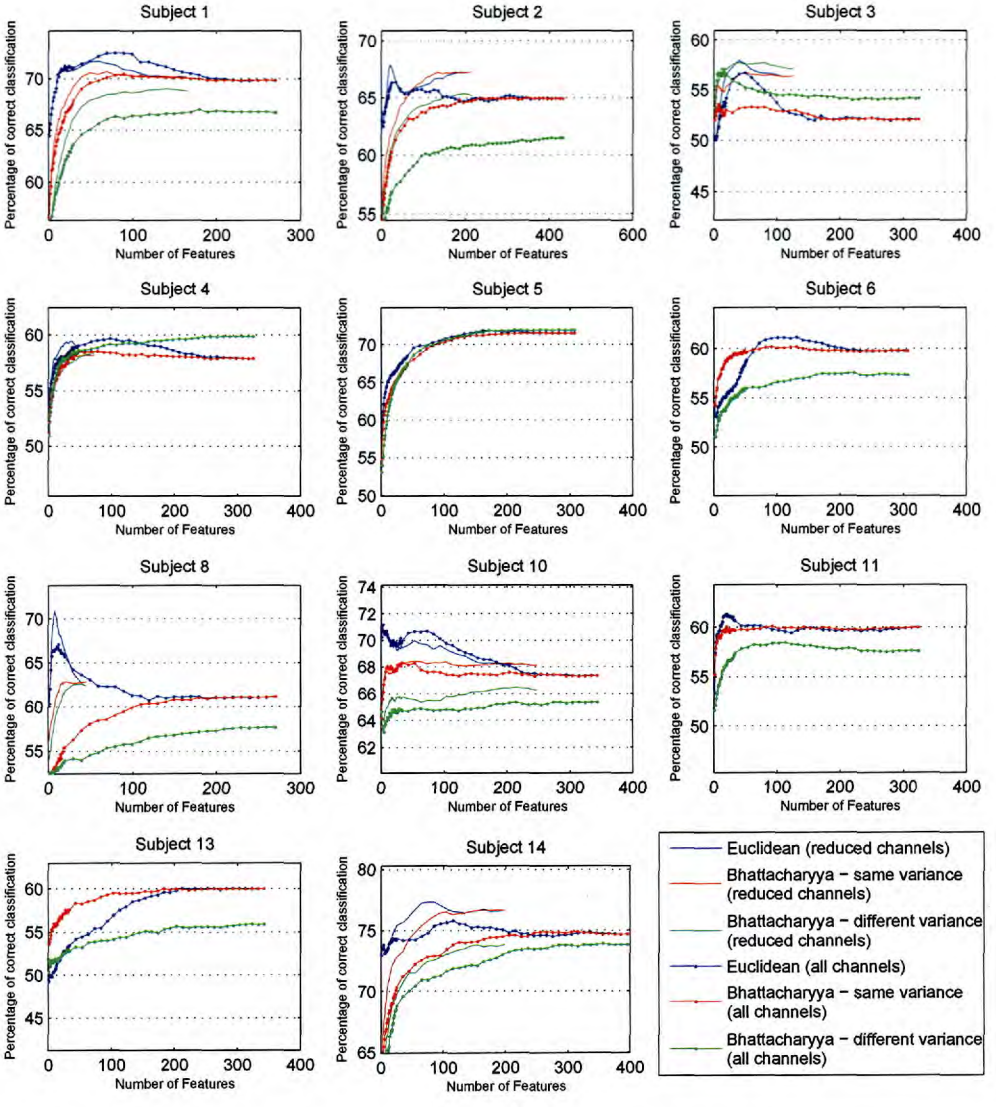


Figure 3.4: Classification accuracy as a function of the number of features used.

3.5 Classification in a feature space of reduced dimensionality

In this section we use a number of methods to construct a new feature space for the two classes, which is of a lower dimensionality than the original one. We then perform classification using the new features. The motivation for the construction of a new feature space of lower dimensionality is twofold. First, the reduction in dimensionality can prevent overfitting that could result from our inability to populate a highly dimensional feature space with a limited number of samples in our training set. Second, the various criteria that are used for the construction of the new space may construct more discriminative features and enhance the classification accuracy.

We give below a brief description of the various methods used for the construction of the new spaces. Then we present the classification results acquired.

3.5.1 Methods

Principal Component Analysis

Having a random vector $\boldsymbol{f} = (f_1, ..., f_Q)^{\mathrm{T}}$, containing Q random variables f_i, PCA finds a linear $Q \times Q$ transformation $\boldsymbol{B}^{\mathrm{T}}$ which decorrelates them and at the same time projects them onto axes ranked in order of decreasing variance: $\boldsymbol{h} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{f}$. The new variables are called principal components of $\boldsymbol{f}$. The rows of $\boldsymbol{B}^{\mathrm{T}}$ are directional vectors, with norm one. It can be proven [39] that the matrix $\boldsymbol{B}^{\mathrm{T}}$ which achieves this, has as rows the eigenvectors of the covariance matrix of $\boldsymbol{f}$, denoted as $\boldsymbol{C_f} = E\{(\boldsymbol{f} - \overline{\boldsymbol{f}})(\boldsymbol{f} - \overline{\boldsymbol{f}})^{\mathrm{T}}\}$. An estimation of $\boldsymbol{C_f}$ is usually computed as $\boldsymbol{C_f} = \frac{1}{N-1}(\boldsymbol{F} - \overline{\boldsymbol{F}})(\boldsymbol{F} - \overline{\boldsymbol{F}})^{\mathrm{T}}$, where $\boldsymbol{F}$ is a $Q \times N$ matrix the columns of which are different instantiations of $\boldsymbol{f}$ and $\overline{\boldsymbol{F}}$ is a $Q \times N$ matrix the columns of which are the mean feature vector $\overline{\boldsymbol{f}}$ as estimated by the available instantiations. The variance of the retained variable h_i is the eigenvalue of the i^{th} corresponding eigenvector. Thus, dimension reduction can take place by keeping the P variables h_i, computed as the projections of $\boldsymbol{f}$ on the eigenvectors with the P largest eigenvalues, i.e. the principal components with the largest variance.

PCA has been used to process EEG features before classification in various cases, such as [54,55,58]. In our case it is used as follows.

- Let us denote with $\mathbb{S}_f = \{\boldsymbol{f}_i\}, i = 1, ..., L$ and $\mathbb{S}_{\widetilde{f}} = \{\widetilde{\boldsymbol{f}}_j\}, j = 1, ..., \widetilde{L}$ the training and the testing set, respectively, both containing different instantiations of the feature vector $\boldsymbol{f}$. The mean feature vector $\overline{\boldsymbol{f}}$ is computed, using the elements of the training set and the $Q \times L$ matrix $\boldsymbol{F} = (\boldsymbol{f}_1 - \overline{\boldsymbol{f}}, ..., \boldsymbol{f}_L - \overline{\boldsymbol{f}})$ is constructed.

- We perform PCA on the feature vector $\boldsymbol{f}$ calculating the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{C_f} = \frac{1}{L-1}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$. We then construct the $P \times Q$ matrix $\boldsymbol{B}^{\mathrm{T}}$, the rows of which are the eigenvectors corresponding to the P largest eigenvalues and the new feature vector for each trial in the training set is computed as $\boldsymbol{h}_i = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{f}_i$. Parameter P is chosen such that a large percentage of the total variance in the data is encapsulated.

- Finally the new feature vectors $\widetilde{\boldsymbol{h}}_j$ of the trials in the testing set are constructed from the original ones $\widetilde{\boldsymbol{f}}_j$ by projecting on the selected eigenvectors, i.e. $\widetilde{\boldsymbol{h}}_j = \boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{f}}_j$.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [29] is a well known technique which finds the transformation maximizing the between-class scatter to within-class scatter. For a classification problem of K equiprobable classes the within-class scatter is described by matrix $\boldsymbol{S_w} = \sum_k \boldsymbol{R}_{\boldsymbol{f}_k}/K$, where $\boldsymbol{R}_{\boldsymbol{f}_k}$ is the covariance matrix of the k^{th} class. The between-class scatter is described by matrix $\boldsymbol{S_b} = \frac{1}{K}\sum_k(\overline{\boldsymbol{f}}_k - \overline{\boldsymbol{f}})(\overline{\boldsymbol{f}}_k - \overline{\boldsymbol{f}})^{\mathrm{T}}$, where $\overline{\boldsymbol{f}}_k$ is the mean feature vector of the k^{th} class and $\overline{\boldsymbol{f}}$ the mean feature vector with respect to all classes. The desired transformation $\boldsymbol{B}$ is defined as the transformation maximising function:

$$J(\boldsymbol{B}) \equiv \frac{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{S_b}\boldsymbol{B}}{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{S_w}\boldsymbol{B}} \tag{3.12}$$

Matrix $\boldsymbol{B}$ is constructed using the eigenvectors of matrix $C = \boldsymbol{S_w}^{-1}\boldsymbol{S_b}$, corresponding to non zero eigenvalues, as columns. For a K class problem the number of such eigenvectors is $K - 1$. In our case LDA is used as follows:

- Let us denote with $\mathbb{S}_f = \{\boldsymbol{f}_i\}, i = 1, ..., L$ and $\mathbb{S}_{\widetilde{f}} = \{\widetilde{\boldsymbol{f}}_j\}, j = 1, ..., \widetilde{L}$ the training and the testing set respectively, both containing different instantiations of the feature vector $\boldsymbol{f}$.

- Having two classes in our problem the inequality $P \leq \widetilde{L} - 2$, should be true to ensure the invertibility of $\boldsymbol{S_w}$, where $\widetilde{L}$ is the size of the training set and P is the dimensionality of the feature vector. For this reason we reduce the dimensionality of the original feature vector, either using PCA and keeping the coefficients in the directions of larger power or by selecting a number of discriminating features using the Euclidean distance as in Section 3.4. Thus, a new feature vector $\boldsymbol{f'}$ of dimension P is constructed. The new training and testing sets are denoted as: $\mathbb{S}_{f'} = \{\boldsymbol{f'}_i\}, i = 1, ..., L$ and $\mathbb{S}_{\widetilde{f}'} = \{\widetilde{\boldsymbol{f}}'_j\}, j = 1, ..., \widetilde{L}$.

- Matrices $\boldsymbol{S_w}$ and $\boldsymbol{S_b}$ are constructed using the vectors in the training set $\mathbb{S}_{f'}$. Then the eigenvectors of $C = {\boldsymbol{S_w}}^{-1}\boldsymbol{S_b}$ corresponding to non zero eigenvalues are used for the construction of the desired transformation $\boldsymbol{B}$. Since there are only two classes in our problem, $\boldsymbol{B}$ is a column vector.

- The new feature vectors of the training and testing sets are constructed as $\boldsymbol{h}_i = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{f'}_i$ and $\widetilde{\boldsymbol{h}}_j = \boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{f}}'_j$, respectively.

Non Negative Matrix Factorization

We describe here how to use the recently proposed technique of Non Negative Matrix Factorization (NMF) [52] to construct a new, low dimensionality feature space for the desired classification.

Let us consider a space of feature vectors $\mathbb{F} \subseteq \mathbb{R}^{Q(+)}$, where $\mathbb{R}^{Q(+)}$ is the space of real, non negative numbers of dimension Q, and a $Q \times N$ matrix $\boldsymbol{F}$, the columns of which are different instantiations of feature vectors belonging to $\mathbb{F}$. NMF factorizes $\boldsymbol{F}$ in the form $\boldsymbol{F} \approx \boldsymbol{BH}$, where $\boldsymbol{B}$ and $\boldsymbol{H}$ are $Q \times P$ and $P \times N$ matrices with non negative elements. There are a variety of algorithms that have been proposed for this factorization. We use here the simplest one [53] which minimizes the Euclidean distance $\|\boldsymbol{F} - \boldsymbol{BH}\|$

under the update rules:

$$B_{ij} \leftarrow B_{ij} \frac{(FH^{\mathrm{T}})_{ij}}{(BHH^{\mathrm{T}})_{ij}} \qquad H_{jk} \leftarrow H_{jk} \frac{(B^{\mathrm{T}}F)_{jk}}{(B^{\mathrm{T}}BH)_{jk}} \tag{3.13}$$

The columns of matrix $\boldsymbol{B}$ form a basis for the approximation of space $\mathbb{F}$. These columns are usually called *parts*, since they are constructively (because of the non negativity of both their elements and the coefficients) combined to form a feature vector. Moreover, each column of $\boldsymbol{H}$ contains the coefficients for the construction of the feature vector at the corresponding column of $\boldsymbol{F}$, and therefore can be thought of as its encoding. Because of the non negativity constraint of the algorithm, both matrices $\boldsymbol{B}$ and $\boldsymbol{H}$ are usually sparse. This is especially obvious when $P \ll Q$. The sparsity of the columns of $\boldsymbol{B}$ implies that each 'part' of the feature vectors incorporates specific only features. Moreover, the sparsity of the columns of $\boldsymbol{H}$ shows that each of the feature vectors mainly needs few of the 'parts' for its formation, i.e. it belongs to a subspace of the space spanned by the columns of $\boldsymbol{B}$. If we find an approximation with these characteristics for a space of feature vectors $\mathbb{F}$ and it is adequate, then we say that space $\mathbb{F}$ has a meaningful part based representation.

If the nature of the original feature space $\mathbb{F}$ permits a meaningful part based representation, then the encodings of the feature vectors, i.e the columns of $\boldsymbol{H}$, can have a large classifying capability. This is because their elements denote the importance that has the corresponding basis vector in the representation of the original feature vectors and this importance will be similar for trials belonging to the same class (i.e. objects of the same class are composed of the same parts). Thus, the columns of $\boldsymbol{H}$ can be used as the new feature vectors according to which we shall perform the classification. We give below a detailed description of the algorithm that is used for the construction of the new feature vectors in our case. Similar algorithms have been used in [57] for EEG data classification and in [36] for image classification.

- Let us denote with $\mathbb{S}_f = \{\boldsymbol{f}_i\}, i = 1, ..., L$ and $\mathbb{S}_{\widetilde{f}} = \{\widetilde{\boldsymbol{f}}_j\}, j = 1, ..., \widetilde{L}$ the training and the testing set, respectively, both containing different instantiations of feature

vector $\boldsymbol{f}$. We construct the $Q \times L$ matrix $\boldsymbol{F}$ the columns of which are the elements of the training set, i.e. $\boldsymbol{F} = (\boldsymbol{f}_1, ..., \boldsymbol{f}_L)$.

- We factorize $\boldsymbol{F}$ using NMF to find the $Q \times P$ matrix $\boldsymbol{B} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_P)$ and the $P \times L$ matrix $\boldsymbol{H} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_L)$ which satisfy $\boldsymbol{F} \approx \boldsymbol{BH}$. $P < Q$ is a free parameter of the algorithm. The set of vectors (parts) $\{\boldsymbol{b}_k\}$, $k = 1, .., P$, is a basis which can span a good approximation of the feature space and $\{\boldsymbol{h}_i\}$, $i = 1, ..., L$, are the encodings of the feature vectors of the trials (i.e. the coefficients used to construct the feature vectors as a linear combination of vectors $\boldsymbol{b}_k$). These encodings are the new feature vectors of the trials in the training set, that will be finally used in the classification process.

- Finally, in order to find the new feature vectors $\widetilde{\boldsymbol{h}}_j$ from the feature vectors $\widetilde{\boldsymbol{f}}_j$ in the testing set, we first construct matrix $\widetilde{\boldsymbol{F}}$, the columns of which are vectors $\widetilde{\boldsymbol{f}}_j$. We then use the iterative algorithm of NMF to factorize $\widetilde{\boldsymbol{F}}$ in the form $\widetilde{\boldsymbol{F}} \approx \boldsymbol{B}\widetilde{\boldsymbol{H}}$. The difference with the case of the training set is that now matrix $\boldsymbol{B}$ is not updated at each iteration but is fixed with the values that have been previously computed, as suggested in [36]. Let us mention that this means that the encoding $\widetilde{\boldsymbol{h}}$ of a feature vector $\widetilde{\boldsymbol{f}}$ depends on the fixed matrix $\boldsymbol{B}$ and not on the other feature vectors in the testing set.

3.5.2 Results

In all cases below, the classification experiments take place as follows. The original feature space is constructed and then the methods described above are applied to create a new feature space. The set of available trials is splittted into two halves with the one being used as training and the other as testing set. The classification task is performed using the Gaussian classifier. The whole process is repeated 500 times for randomly different compilations of the training and testing sets and the classification accuracy is measured through averaging.

Table 3.5: Size of original feature space and training set per subject

Subject	1	2	3	4	5	6	8	10	11	13	14
Size of training set	28	24	18	46	40	50	18	32	32	32	24
Dimension of original feature space (All features)	270	432	324	324	306	306	270	342	324	342	396
Dimension of original feature space (Reduced features)	27	23	17	45	39	49	17	31	31	31	23

Classification results using Principal Component Analysis

We present here the classification results we obtain when the new feature space is constructed using PCA. Two different ways are used for the construction of the original feature space. In the first way, the features are constructed as described in Section 3.1. In the second way the dimensionality of the original feature space is reduced by keeping only a subset of size $L-1$, of the features producing, each one separately, the largest Euclidean distance between the two classes, as it is computed from the training set. Let us repeat here that the Euclidean distance between the two classes with respect to a certain feature is the Euclidean distance between the means of the feature in the two classes. L is the number of trials in the training set. The dimension of the feature space in the second case is chosen as $L-1$ in order to guarantee a non singular feature covariance matrix in the PCA which will follow. The size of the original feature space in the two cases as well as the size of the training set can be seen in Table 3.5.

After constructing the original feature space, PCA is applied to construct the new one as described in Section 3.5.1. The number of the new features constructed, which are the principal components of the original feature vector, equals the number $\widetilde{L}$ of the non zero eigenvalues of the covariance matrix of the original feature vector. $\widetilde{L}$ sets of features are constructed, with the n^{th} set, $n = 1, ..., \widetilde{L}$, containing the n features with the largest variance, i.e. those corresponding to the directions of the largest eigenvalues. Since the new features are mutually uncorrelated, Eq. (3.8) is used for the classification task with the Gaussian classifier. This also means that we assume that the feature vectors of the two classes have the same covariance matrices. Experiments estimating a separate covariance

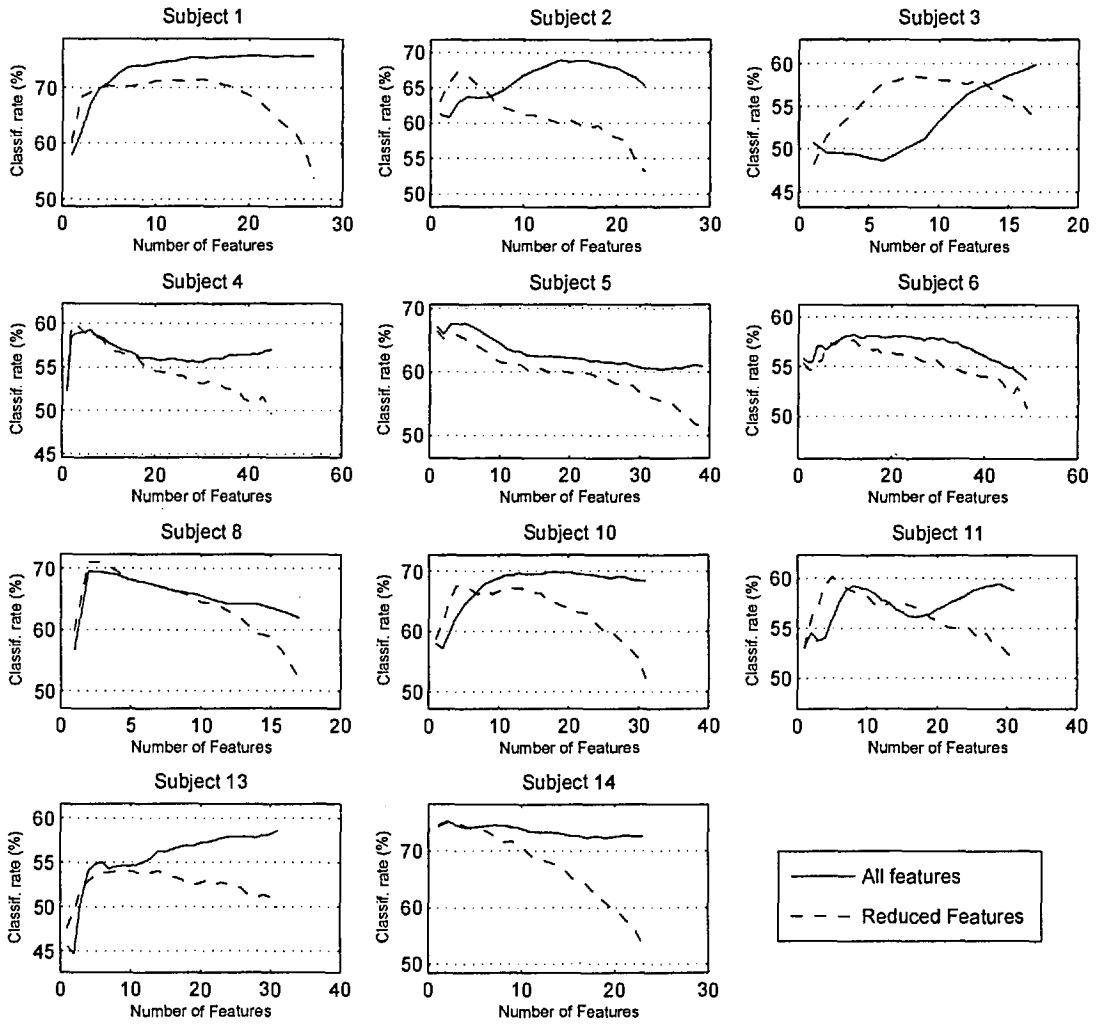


Figure 3.5: Classification accuracy in the space constructed with PCA as a function of the number of principal components used.

matrix for each class were performed as well, but the results were worse (or the same) so they are not presented here. The classification accuracy as a function of the number of features kept can be seen in Figure 3.5.

Observing Figure 3.5, we can see that for most subjects the classification accuracy is maximised for a certain number of principal components kept and then it starts decreasing. Moreover, we can see that there is no significant improvement for the case when the number of the original features is initially reduced. In order to evaluate the classification results we obtain using PCA, we compare them with the results that the use of the initial features produces (i.e those acquired in Sections 3.3 and 3.4). The number of principal

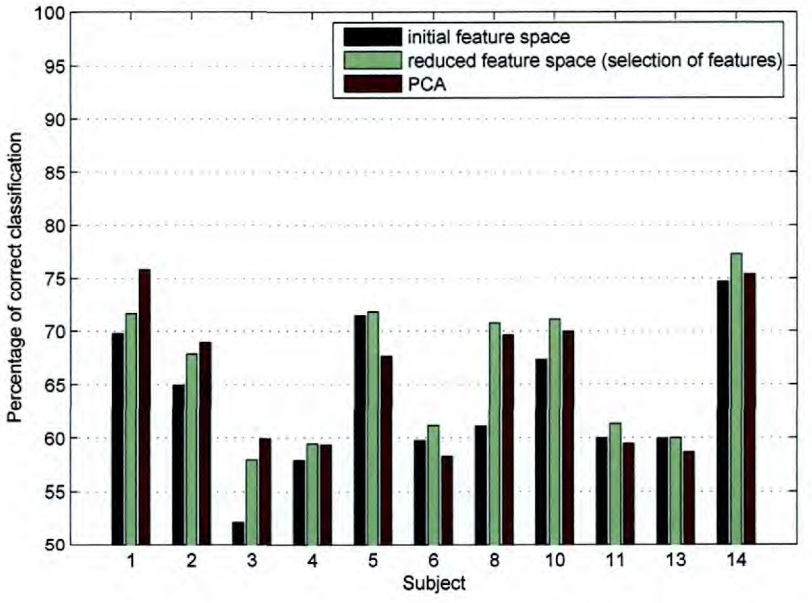


Figure 3.6: Comparison of the classification accuracy achieved in the initial feature space, reduced feature space (selection of specific features) and PCA.

components or initial features giving maximum classification accuracy is used for the purpose of comparison. For the case that PCA is used, the original features are not first reduced, since this is the case that gives best results for the majority of subjects. The results can be seen in Figure 3.6.

As seen from Figure 3.6 PCA improves the classification accuracy for the majority of subjects compared with the classification accuracy achieved in the initial feature space. As far as the comparison with the reduced feature space created from the selection of specific features is concerned, PCA improves the classification accuracy only for Subjects 1, 2 and 3.

Classification results using Linear Discriminant Analysis

We present here the classification results in the space constructed by LDA. As explained in Section 3.5.1 the dimensionality of the original feature vector is first reduced either by keeping the directions of largest power through PCA or by selecting the most discriminating features according to the Euclidean distance of the two classes in the training set. This is done in order to ensure the invertibility of the matrix $\boldsymbol{S_w}$. The Gaussian classifier, i.e.

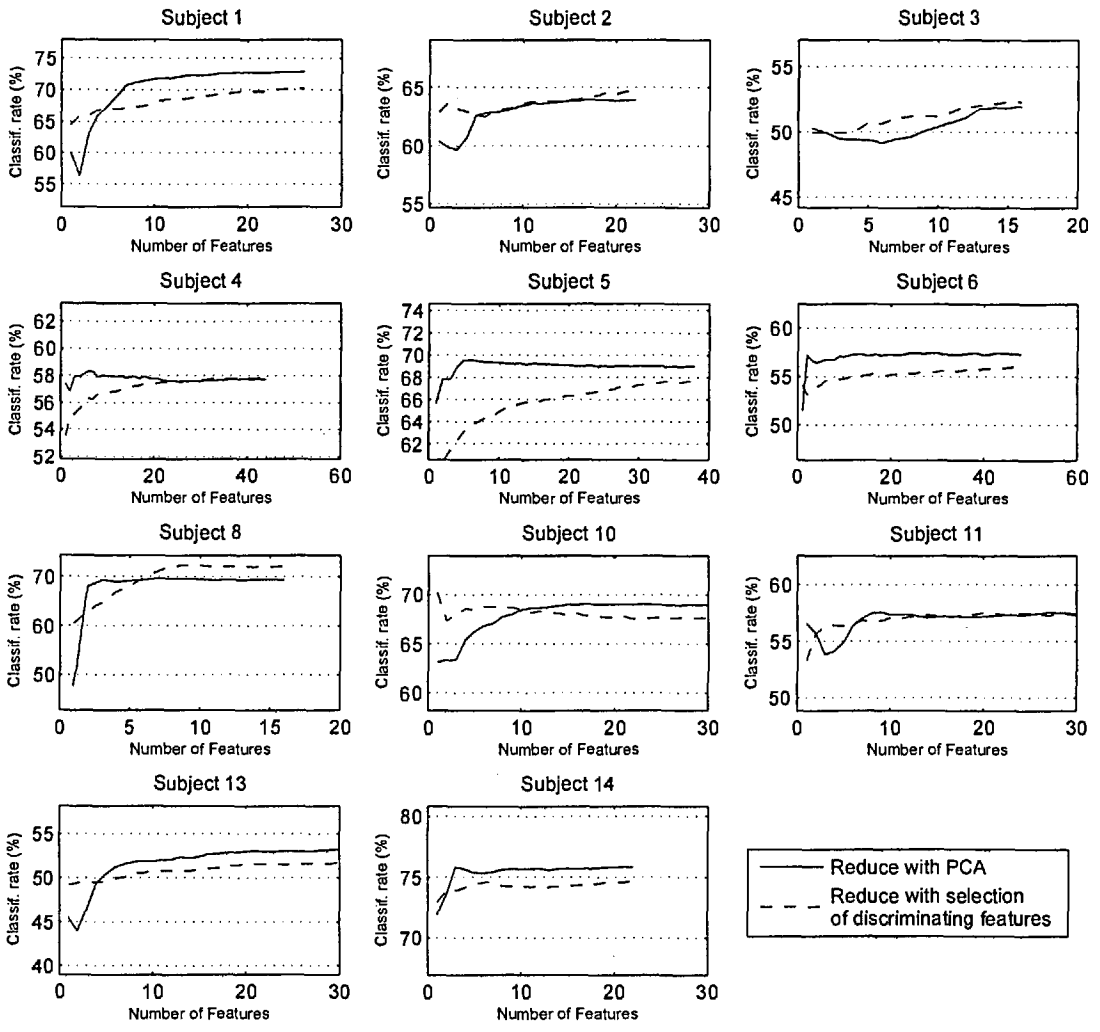


Figure 3.7: Classification accuracy in the space constructed with LDA as a function of the number of initial features used.

Eq. (3.8) since the feature vector constructed with LDA is actually a scalar, is used for the task of classification. The classification results as a function of the number of features initially used can be seen in Figure 3.7. Let us repeat here that Eq. (3.8) assumes that the feature variable in the two classes has the same variance. Experiments estimating different variance for the feature variable of each class were carried out but since the results were similar or worse they are not presented here.

Observing Figure 3.7 we can see that in general, better results are acquired when the dimensionality of the original feature vector is reduced through PCA. Only for Subject 8 the reduction through selection of the most discriminating features is clearly superior.

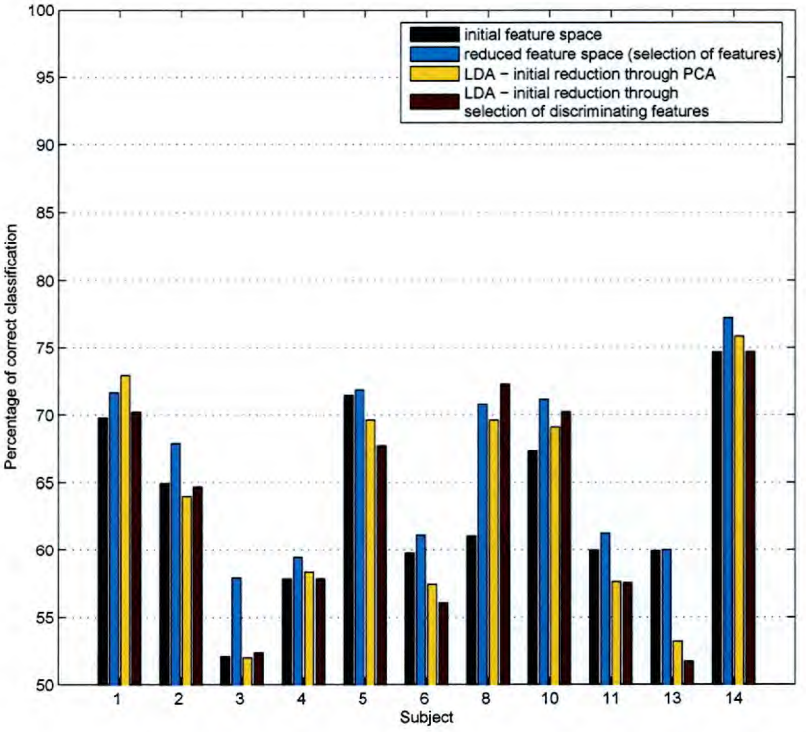


Figure 3.8: Comparison of the classification accuracy achieved in the initial feature space, the reduced feature space (selection of specific features) and the space constructed with LDA (2 cases of initial reduction of feature vectors).

We next compare the classification results we get using the space constructed with LDA with the results coming from the use of the initial feature space (i.e those acquired in Sections 3.3 and 3.4). In all cases the number of features giving the maximum accuracy is used. The results can be seen in Figure 3.8. We can see that the results with LDA are better only for Subjects 1 and 8 as for all other cases the method using directly the features with the largest distance produces higher classification accuracy.

Classification results using Non Negative Matrix Factorization

We present here the classification results obtained in a feature space constructed using NMF. We describe below the three different ways we used NMF to construct the new space. In the first two cases NMF is applied in two different features spaces for various levels of reduction (i.e. different sizes of the NMF basis). In the third case the original feature space is reduced with NMF only once and then the NMF coefficients (i.e. new

features) are further reduced based on the producing Euclidean distance between the two classes. More precisely we have:

- Case 1 ("All features"): The original feature space of Section 3.1 is analysed using NMF. The number of NMF basis vectors and consequently the dimensionality of the new space varies between 1 and the size L of the training set of each subject.
- Case 2 ("Reduced features"): The original feature space of Section 3.1 is first reduced, computing the Euclidean distance between the two classes resulting from the use of each feature and keeping the L features corresponding to the L largest distances. Then the reduced space is analysed with NMF L times with the number of NMF basis vectors varying between 1 and L. The new feature spaces are constituted each time from the coefficients of the NMF basis vectors.
- Case 3 ("Selected subset of NMF components"): The original feature space of Section 3.1 is first analysed using NMF, for L basis vectors. The new feature space of size L produced (i.e. coefficients of NMF) is further reduced L times, to a size varying from 1 to L. This is done computing the Euclidean distance between the two classes resulting from the use of each feature and keeping each time the features corresponding to the largest distances.

As far as the classification task, is concerned the Gaussian classifier is used, with a common diagonal covariance matrix for the NMF features of the two classes. Experiments with a separate covariance matrix for each class were carried out as well, but the results were worse so they are not presented here. The classification accuracy produced across the number of NMF features used for the three cases can be seen in Figure 3.9. There is no specific tendency common in all subjects as far as the relation between the dimensionality of the space and the classification accuracy is concerned.

For reasons of simplicity we use only the first case for the construction of the original feature space (all features constructed as described in Section 3.1 are used) to compare the results with the cases when classification is performed in the initial feature space (i.e. Sections 3.3 and 3.4). In all cases the number of features giving the maximum accuracy is

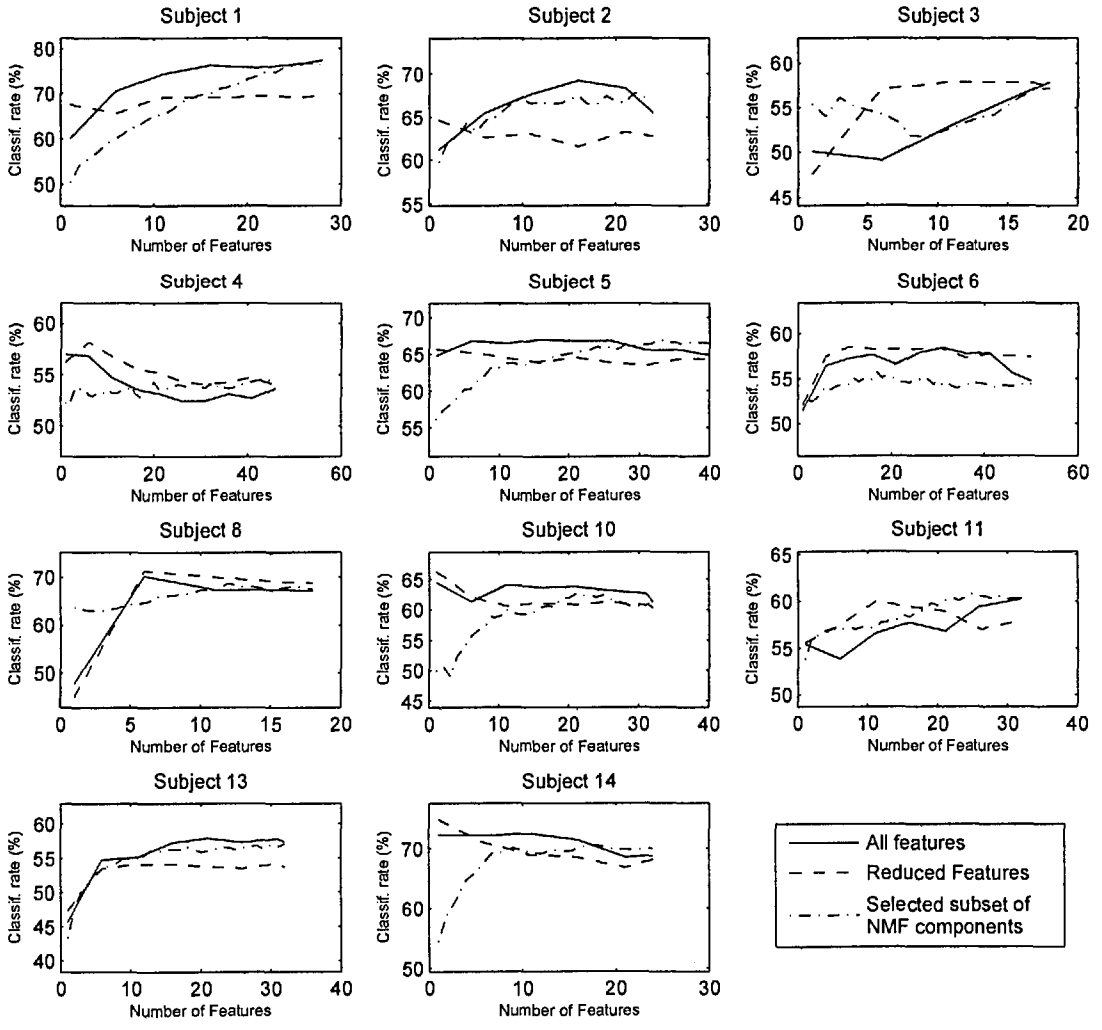


Figure 3.9: Classification accuracy on the space constructed with NMF as a function of its dimensionality.

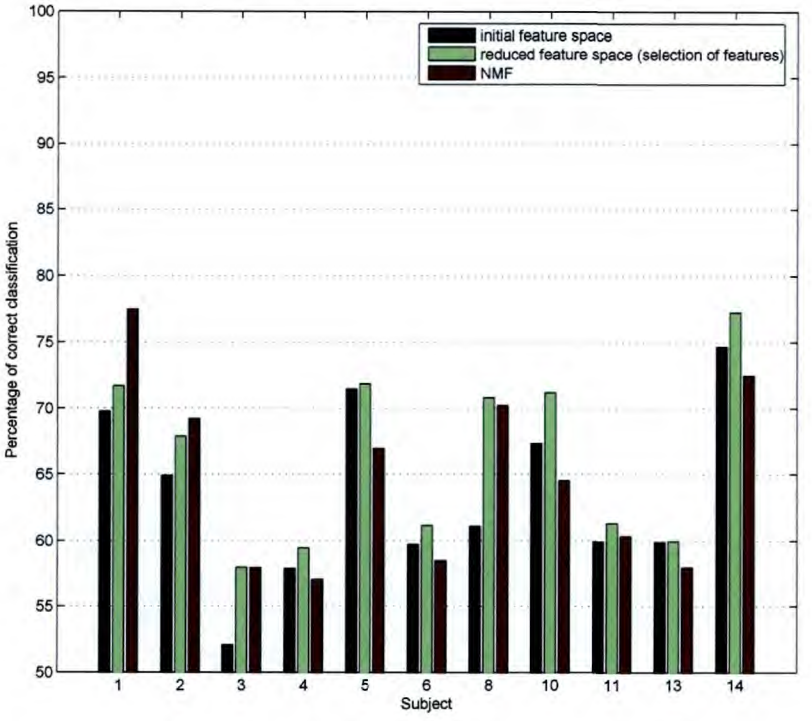


Figure 3.10: Comparison of the classification accuracy achieved with the initial feature space, the reduced feature space (selection of specific features) and the space constructed with NMF.

used. The results comparing the methods can be seen in Figure 3.10. As one can observe, there is an improvement for Subjects 1 and 2 but for the rest of the subjects the method using the space constructed with NMF is outperformed by the one using the space of the features with the largest distance.

3.6 Classification using subspace methods

In this section we perform classification constructing a different subspace for each of the two classes in our problem. For the construction of the subspace of each class two methods are used: Principal Component Analysis and Non Negative Matrix Factorization. In order to classify a trial in the testing set, we project its feature vector on the two subspaces and choose the class the subspace of which gives the best approximation for the feature vector.

We first give a brief description of the methods used and then present the acquired results.

3.6.1 Methods

Principal Component Analysis

Let us assume that we have a Q dimensional random vector $\boldsymbol{f}$ and we want to find a P dimensional subspace (i.e. $P < Q$), such that the mean square error between $\boldsymbol{f}$ and its projection on the subspace is minimized. It can be proved [25] that the desired subspace can be found performing PCA on $\boldsymbol{f}$, as described in Section 3.5.1, but using this time the correlation matrix $\boldsymbol{R_f} = E\{\boldsymbol{f}\boldsymbol{f}^{\mathrm{T}}\}$ of the random vector $\boldsymbol{f}$. Thus, the desired subspace is spanned by the P eigenvectors corresponding to the P largest eigenvalues of $\boldsymbol{R_f}$.

We use here this technique to construct two subspaces for the two classes of our problem, using the feature vectors in the training set. If the subspaces constructed are representative for each class and different between them, then they are suitable for classification. The classification process can be described as follows:

- Let us assume that we have L_1 trials in our training set belonging to the class "success" and L_2 trials belonging to the class "failure". We construct the feature matrices $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ the columns of which are the feature vectors $\boldsymbol{f}$ of the trials belonging to the classes "success" and "failure", respectively. The correlation matrices of the two classes are computed as $\boldsymbol{R_{f1}} = \frac{1}{L_1-1}\boldsymbol{F}_1\boldsymbol{F}_1^{\mathrm{T}}$ and $\boldsymbol{R_{f2}} = \frac{1}{L_2-1}\boldsymbol{F}_2\boldsymbol{F}_2^{\mathrm{T}}$.

- We perform PCA on the feature vector of each class computing the eigenvalue decomposition of matrices $\boldsymbol{R_{f1}}$ and $\boldsymbol{R_{f2}}$. Matrices $\boldsymbol{B}_1^{\mathrm{T}}$ and $\boldsymbol{B}_2^{\mathrm{T}}$, constructed using as columns the P eigenvectors of $\boldsymbol{R_{f1}}$ and $\boldsymbol{R_{f2}}$ corresponding to the P largest eigenvalues, span the subspace of the "success" and "failure" class, respectively. P is a free parameter of the algorithm and it should be chosen such that the ratio of the sum of the rejected eigenvalues over the sum of the retained eigenvalues is small.

- For each trial in the testing set we project the equivalent feature vector $\widetilde{\boldsymbol{f}}$ on the constructed subspaces. The distance between $\widetilde{\boldsymbol{f}}$ and its projection on the two spaces, i.e. $|\widetilde{\boldsymbol{f}} - \boldsymbol{B}_i\boldsymbol{B}_i^{\mathrm{T}}\widetilde{\boldsymbol{f}}|$, $i = 1, 2$, is computed and the trial is classified to the class which gives the minimum distance.

Non Negative Matrix Factorization

Another way to construct the desired subspaces for the two classes is using NMF. The idea of constructing suitable subspaces for classification using NMF has been proposed in [35] for image classification. As explained in Section 3.5.1, NMF constructs a basis for a random Q dimensional vector $\boldsymbol{f}$, the elements of which are always positive, minimizing the Euclidean distance:

$$\|\boldsymbol{F} - \boldsymbol{BH}\| \quad \text{where} \quad \|\boldsymbol{A} - \boldsymbol{B}\| = \sum_{ij}(A_{ij} - B_{ij})^2. \tag{3.14}$$

$\boldsymbol{F}$ is a matrix the columns of which are the available instances of $\boldsymbol{f}$, the $P < Q$ columns of $\boldsymbol{B}$ are the basis vectors and the columns of $\boldsymbol{H}$ the coefficients for the approximation of the corresponding instance of $\boldsymbol{f}$. This is similar to the minimization criterion of the mean square error that PCA uses. However, the main difference is that in NMF the elements of the basis vectors and the coefficients are forced to be non negative.

If the nature of $\boldsymbol{f}$ is "suitable" then the above approximation can be quite good. This means that the instances of $\boldsymbol{f}$ have a good approximation on the subspace spanned by the columns of $\boldsymbol{B}$, under the constraint of non negativity of the coefficients. Thus, constructing one such subspace for each of the two classes, an unknown feature vector can be classified to the class the subspace of which produced the best approximation. We present below a detailed description of the way that the classification process takes place.

- Let us assume that we have L_1 trials in our training set belonging to the class "success" and L_2 trials belonging to the class "failure". We construct the feature matrices $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ the columns of which are the feature vectors $\boldsymbol{f}$ of the trials belonging to the classes "success" and "failure", respectively.

- We factorise both of the above matrices using NMF to find two $Q \times P$ matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ satisfying $\boldsymbol{F}_1 \approx \boldsymbol{B}_1\boldsymbol{H}_1$ and $\boldsymbol{F}_2 \approx \boldsymbol{B}_2\boldsymbol{H}_2$. The columns of matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are the bases for the classes "success" and "failure", respectively. The size P of the two bases is a free parameter of the algorithm and it has to satisfy $P < Q$ and

$P < L_1, P < L_2$.

- For each trial in the testing set, we project the equivalent feature vector $\widetilde{\boldsymbol{f}}$ on the space spanned by the two computed bases. The projection of a vector $\widetilde{\boldsymbol{f}}$ on the space spanned by the linear independent columns of a matrix $\boldsymbol{B}$, i.e. the vector in this space having the minimum distance from $\widetilde{\boldsymbol{f}}$, is given by $\boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{f}}$. However the contributions of the vectors of the basis in this way are not necessarily constructive, because the coefficients $(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{f}}$ are not necessarily non negative. Thus, we prefer to use the approach proposed in [36] and use the NMF algorithm to get $\widetilde{\boldsymbol{f}} \approx \boldsymbol{B}_1\widetilde{\boldsymbol{h}}_1$ and $\widetilde{\boldsymbol{f}} \approx \boldsymbol{B}_2\widetilde{\boldsymbol{h}}_2$, keeping matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ fixed. Vectors $\boldsymbol{B}_1\widetilde{\boldsymbol{h}}_1$ and $\boldsymbol{B}_2\widetilde{\boldsymbol{h}}_2$ are the projections of vector $\widetilde{\boldsymbol{f}}$ on the space of classes "success" and "failure", respectively.

- The distances $|\widetilde{\boldsymbol{f}} - \boldsymbol{B}_i\widetilde{\boldsymbol{h}}_i|$, $i = 1, 2$, are calculated and the trial with the feature vector $\widetilde{\boldsymbol{f}}$ is classified to the class which gives the minimum distance.

3.6.2 Results

In all cases below, the classification experiments take place as follows. The original feature space is constructed and then the methods described above are applied to construct a subspace for each class. The set of available trials is splitted into two halves with the one being used as training and the other as testing set. The classification task is performed projecting the testing feature vectors on the two subspaces and classifying them according to the minimum distance between the original and the projection. The whole process is repeated 500 times for randomly different compilations of the training and testing sets and the classification accuracy is measured through averaging.

Classification results using Principal Component Analysis

We present here the classification results we obtain when the subspace for each class is constructed using PCA. The initial feature space is constructed according to Section 3.3. Then the eigenvalue decomposition of the correlation matrix of each class, as computed

Table 3.6: Size of initial feature space and samples per class in the training set for each subject

Subject	1	2	3	4	5	6	8	10	11	13	14
Samples per class in the training set	14	12	9	23	20	25	9	16	16	16	12
Size of initial feature space	270	432	324	324	306	306	270	342	324	342	396

from the samples in the training test, is used for the construction of a subspace for each class. Because of the limited number of samples in the training set, which is smaller for each class than the number of features (see Table 3.6), we cannot "see" the whole space of each class. However, even for the subspace observed, PCA shows that the power is concentrated in very few dimensions (see Figure 3.11). Therefore the construction of a smaller subspace than the one observed, capable of providing a good approximation for the feature vectors of each class, is feasible.

The classification accuracy obtained across the number of eigenvectors kept for the construction of each subspace is presented in Figure 3.12. For certain subjects (i.e. 5, 6 and 8) the classification accuracy is maximised when few eigenvectors are used for the construction of the subspaces and it is decreased thereafter. For other subjects (i.e. 1 and 14) the classification accuracy does not really depend on the number of eigenvectors used. In order to evaluate the effectiveness of the method, we compare the classification results (the maximum one achieved for each subject) with the results acquired using the initial feature space (Section 3.3) and those using the most discriminating features (Section 3.4). The comparison can be seen in Figure 3.13. As one can observe, there is an improvement for Subjects 1, 8 and 11. However, in the majority of subjects performing Gaussian classification in a space constructed using features having the largest distance produces the best results.

Classification results using Non Negative Matrix Factorization

We present here the classification results we get when the subspace for each class is constructed using NMF. The initial space is constructed as described in Section 3.3. The

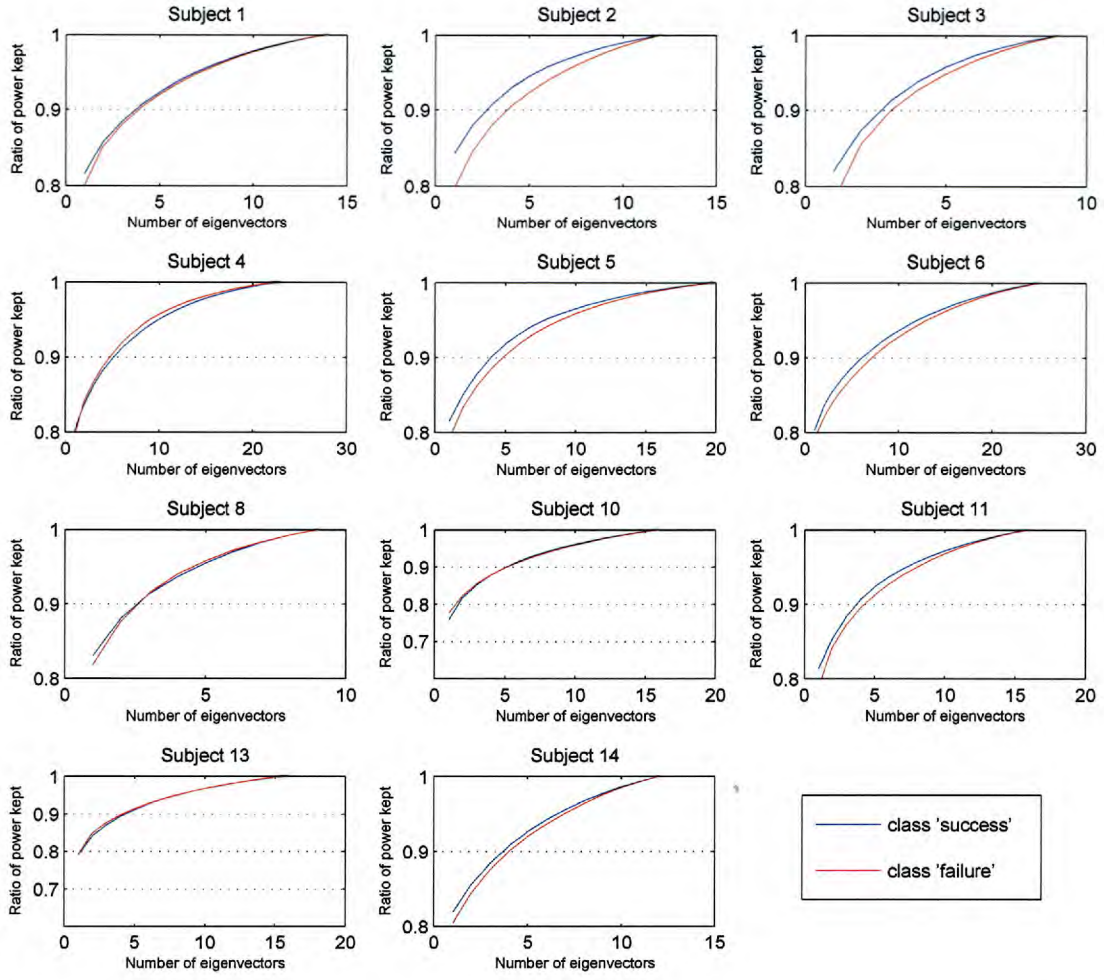


Figure 3.11: Ratio of power kept as a function of the number of eigenvectors corresponding to non zero eigenvalues used.

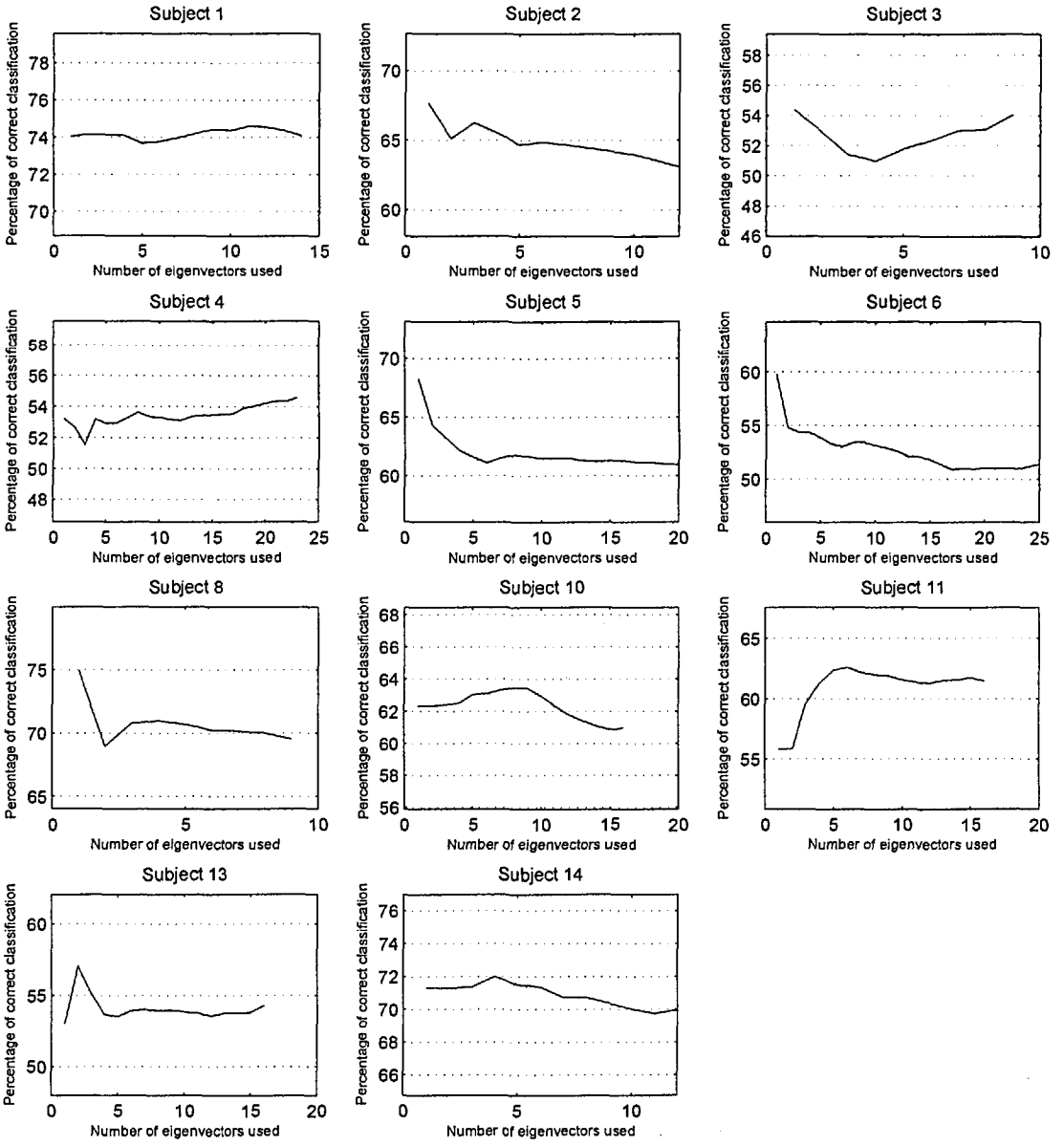


Figure 3.12: Classification accuracy constructing different subspaces for the two classes using PCA.

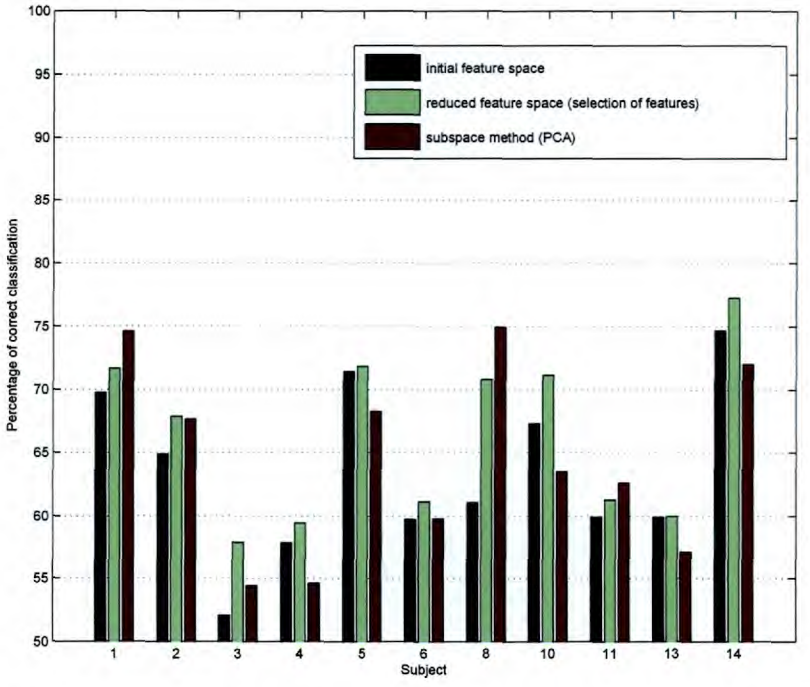


Figure 3.13: Comparison of the classification accuracy achieved in the initial feature space, the reduced feature space (selection of specific features) and the different subspaces for each class using PCA.

dimensionality of the constructed subspaces varies between one and the number of samples for each class in the training set (which is always smaller than the dimensionality of the original space - see Table 3.6). The percentage of correct classification as a function of the dimensionality of the subspaces can be seen in Figure 3.14. We can observe that the results are similar with those obtained when the subspaces were constructed using PCA (see Figure 3.12). As in that case, for certain subjects (i.e. 5, 6 and 8) the classification accuracy is maximised when few dimensions are used for the construction of the subspaces, whereas for other (i.e. 1 and 14) the classification accuracy does not really depend on the dimensionality of the subspace.

A comparison between the results achieved here and those performing classification in the original feature space or the one constructed from a subset of selected features can be seen in Figure 3.15. As in the case in which the subspaces are constructed with PCA, there is an improvement for subjects 1, 8 and 11 whereas for the other subjects performing Gaussian classification in a space constructed using the features having the

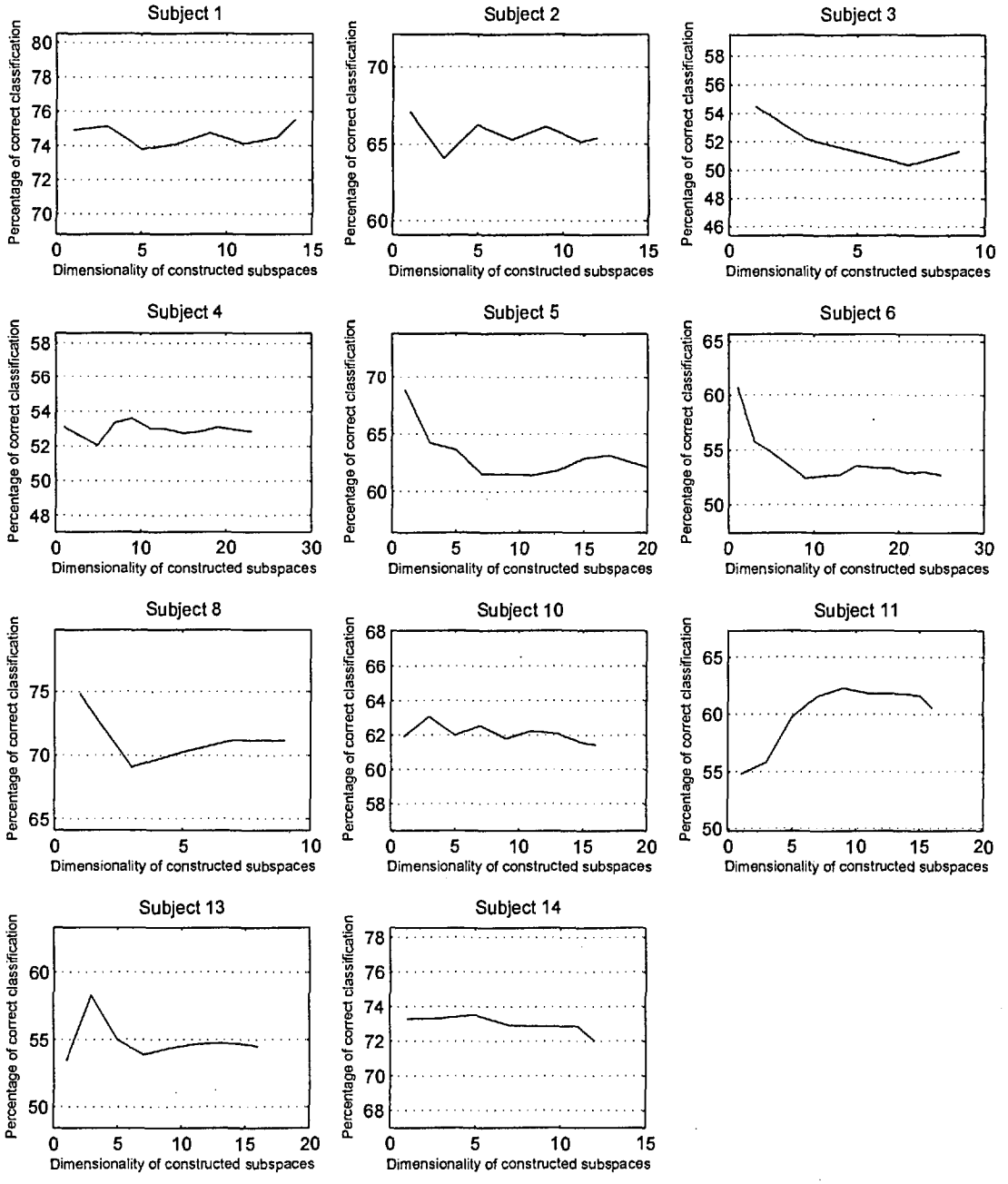


Figure 3.14: Classification accuracy constructing different subspaces for the two classes using NMF.

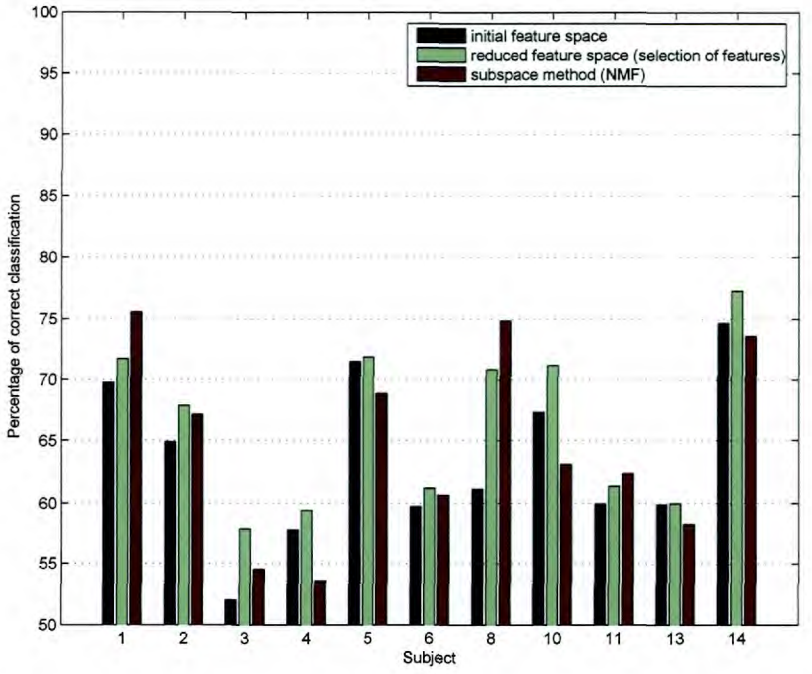


Figure 3.15: Comparison of the classification accuracy achieved in the initial feature space, a reduced feature space (selection of specific features) and in different subspace for each class using NMF.

largest distance produces the best results.

3.7 Conclusions

From the experiments presented in this chapter various conclusions can be drawn that motivate the work presented in the next chapter.

1. In some cases frequency features produce better results than time domain features and vice versa.

2. In general, reducing the number of features used, increases the correct classification rates. However, no algorithm appears to be universally optimal for the process of feature selection.

3. There are no universally good channels or frequency bands.

4. In general, frequencies below 40 Hz are the most useful.

For the last two reasons, in the next chapter we shall use time - frequency joint representations, tailor made to individual subjects, using frequencies below 40 Hz.

Chapter 4

Classification analyzing EEG wavelet transforms with NMF

In this chapter we use Non Negative Matrix Factorization (NMF) to analyse the time-frequency representation of the EEG signals of the trials in order to construct features to be used in the prediction problem. The time frequency representations are acquired using a discretised version of the Continuous Wavelet Transform (CWT). In sections 4.1 and 4.2 the CWT and the Discrete Wavelet Transform (DWT) are presented whereas in section 4.3 the reasons for choosing CWT for our experiments are explained. In section 4.4 the classification rates acquired with time domain features are computed. Then in sections 4.5 and 4.6 two algorithms analysing the time frequency representations of the EEG signals with NMF in two different ways are presented and the corresponding classification results are calculated. Finally, in section 4.7 we compare the classification rates acquired with the three algorithms.

4.1 The continuous wavelet transform

The Continuous Wavelet Transform (CWT) enables the decomposition of a function $f(t)$ into a set of functions $\psi_{a,b}(t), a \in \mathbb{R}^+, b \in \mathbb{R}$, which are localised in time and frequency providing thus a time-frequency representation of f. It is defined as:

$$W_f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t)\psi\left(\frac{t-b}{a}\right) dt \tag{4.1}$$

As it can be seen from this definition the atoms of the decomposition $\psi_{a,b}(t) = a^{-1/2}\psi\left(\frac{t-b}{a}\right)$, named wavelets, are shifted and dilated versions of a function $\psi(t)$, named mother wavelet. Wavelets must satisfy the following condition:

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0. \tag{4.2}$$

Their Fourier transform should be continuously differentiable. These two conditions are sufficient for the existence of the reconstruction formula, which states that any function $f \in L^2(\mathbb{R})$ can be written as:

$$f(t) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} W_f(a,b) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) db \frac{1}{a^2} da \tag{4.3}$$

where

$$C_\psi = \int_0^{+\infty} \frac{|\widehat{\psi}(\omega)|^2}{\omega} d\omega < +\infty \tag{4.4}$$

with $\widehat{\psi}(\omega)$ being the Fourier transform of ψ. Relation (4.4) is named "admissibility condition".

The way that the wavelet transform provides a time-frequency representation of a function can be described as follows. The scaling parameter a of a wavelet atom determines its scale. The notion of scale is related to that of frequency, with large values of scale corresponding to bands of small frequencies and small values of scale corresponding to bands of high frequencies. Moreover the value of scale determines the width of the wavelet both in time and frequency. As a decreases, i.e. we are moving to high frequencies, the width in time decreases whereas the width in frequency increases. The time-frequency resolution remains the same for all values of a, as it depends only on the choice of the mother wavelet, but the intrinsic variability of the time and frequency windows gives us a good tradé-off between the time and frequency resolutions, depending on the range of frequencies we are looking at, each time. Thus the scaling parameter a determines the

frequencies we are looking at. On the other hand, the shifting parameter b changes the time localisation centre of the wavelet, associating the corresponding wavelet coefficients with a specific time period.

4.2 The discrete wavelet transform

The CWT is highly redundant as it represents a function of one variable as a combination of functions of two variables. This redundancy can be reduced by sampling the wavelet coefficients, i.e. keeping those corresponding to specific scales and time shifts. The scaling parameter is discretised as $a = a_0^m$, for a fixed positive value $a_0, (a_0 \neq 1)$. The discretisation of the shifting parameter should depend on a_0^m in order to use large steps for large scales, for which the width of the wavelet in the time domain is large and vice versa. Because the width of the wavelet is proportional to its scale, the shifting parameter is discretised as $b = nb_0a_0^m$. Setting these parameters to (4.1) the Discrete Wavelet Transform (DWT) is defined as:

$$D_f(m,n) = a_0^{-m/2} \int_{-\infty}^{+\infty} f(t)\psi(a_0^{-m}t - nb_0)dt, \quad m,n \in \mathbb{Z} \tag{4.5}$$

In this case the atoms of the decomposition constitute a discrete set of functions $\psi_{m,n}(t) = a_0^{-m/2}\psi(a_0^{-m}t - nb_0)$. With the appropriate choice of the mother wavelet $\psi(t)$ and constants a_0 and b_0, the above set is a frame for space of $L^2(\mathbb{R})$ (for more details we refer to [22], section 3.1). This means that any function $f \in L^2(\mathbb{R})$ can be reconstructed from its discrete wavelet coefficients using the dual set of functions $\widetilde{\psi}_{m,n}$ such as:

$$f(t) = \sum_{m,n} D_f(m,n)\widetilde{\psi}_{m,n} \tag{4.6}$$

The DWT is still a redundant transformation. However, there are special choices of ψ, a_0 and b_0 for which $\psi_{m,n}$ constitute an orthonormal basis for $L^2(\mathbb{R})$. In this case the redundancy is eliminated and any L^2-function can be approximated with arbitrarily large precision by a finite linear combination of functions $\psi_{m,n}$.

4.3 Comparison of CWT and DWT for the analysis of a discrete signal

We describe in this section how CWT and DWT can be applied on a discrete signal of finite duration and compare between the decompositions that the two transforms produce.

In order to apply CWT (see Eq. (4.1)) a type of interpolation should be applied between the samples of the discrete signal. The simplest approach is to use piecewise constant interpolation. In this case the CWT of a signal $f(n)$, $n = 1, ..., N$ can be computed as follows:

$$W_f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t)\psi\left(\frac{t-b}{a}\right) dt = \sum_{n=1}^{N-1} \int_{n}^{n+1} f(t)\psi\left(\frac{t-b}{a}\right) dt \qquad (4.7)$$

and assuming that $f(t) = f(n)$ for $t \in [n, n+1]$ we have:

$$W_f(a,b) = \frac{1}{\sqrt{a}} \sum_{n=1}^{N-1} f(n) \int_{n}^{n+1} \psi\left(\frac{t-b}{a}\right) dt. \qquad (4.8)$$

The integrals of relation (4.8) can be evaluated numerically. Thus the wavelet coefficients, at any desired scale and for a number of shifts equal to the number of samples of f, can be computed. Frequency F_a corresponding to a scale a can be computed from the following formula [1]:

$$F_a = \frac{F_c}{aT_s}, \qquad (4.9)$$

where F_c is the central frequency of the mother wavelet and T_s the sampling period of the discrete function f. The central frequency of the mother wavelet is the frequency with the maximum energy in the Fourier transform of the signal and captures its main oscillations.

In the case of the DWT, the wavelet coefficients can be computed using the scheme of multiresolution analysis [61]. Once an orthogonal wavelet has been chosen, the coefficients of a half-band lowpass and a half-band highpass filter can be computed. Convolving the discrete signal with these filters we get two sets of coefficients, corresponding to two frequency bands: $[0, F_s/4]$ and $[F_s/4, F_s/2]$, where F_s is the sampling frequency of the discrete signal. The output of the highpass filter is the wavelet coefficients, corresponding

to scale 2. The output of the lowpass filter constitutes the "scaling" coefficients which describe the lowpass part of the signal. The filtering procedure can be repeated with the "scaling" coefficients, getting at each level of decomposition two sets of coefficients, corresponding to two sequential frequency bands with half the bandwith of the original one. After each convolution, the resulting sequence is downsampled by a factor of two, so the resolution in time is halved as we move to lower frequencies.

To illustrate the differences between the CWT and DWT, we use them to decompose two different types of signal. The mother wavelet used in both cases is the quadratic biorthogonal spline with 7 vanishing moments ([62], section 7.4.3), and the DWT is applied up to the fifth level of decomposition.

The first signal has duration 5.12 sec, the first half of it is constructed as the superposition of 10 cosines with frequencies 1,2,...,10 Hz and the second half is constructed as the superposition of 5 cosines with frequencies 11,...,15 Hz. The amplitudes of the cosines are normalised so that the two parts have the same energy. The sampling frequency is 50 Hz and the number of samples is 256. The signal in the time domain as well as the absolute value of its continuous and discrete wavelet transforms can be seen in Figures 4.1 and 4.2.

For the presentation of the CWT and DWT in figure 4.2 a time-frequency representation is used instead of a time-scale one. The same type of representation is also used for analysis with NMF in the next sections of this chapter. This is because the correlations between the EEG signals and a person's undergoing cognitive processes are described in the literature with respect to various frequency bands of these signals. Thus, it is natural to work with frequencies rather than scales. The correspondence from scale to frequency for the CWT takes place through Eq.(4.9). This is the frequency around which the spectrum of the wavelet at the corresponding scale is localised. For the DWT the j^{th} level of decomposition corresponds to scale 2^j and as we said before is associated with the frequency band $[F_s/2^{j+1}, F_s/2^j]$ (F_s is the sampling frequency of the signal).

In the second signal which we use as an example, we used two Gaussian functions to resemble the components of the ERP. This signal has a duration of 510 msec and is

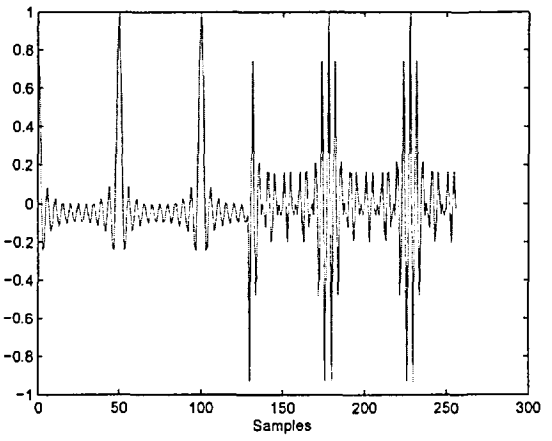
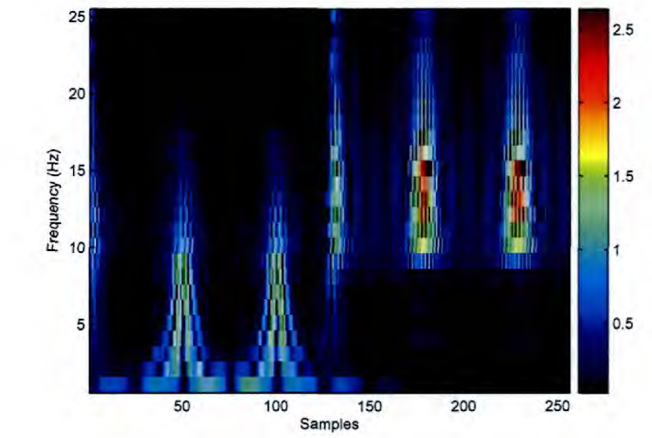


Figure 4.1: A superposition of cosines of different frequencies

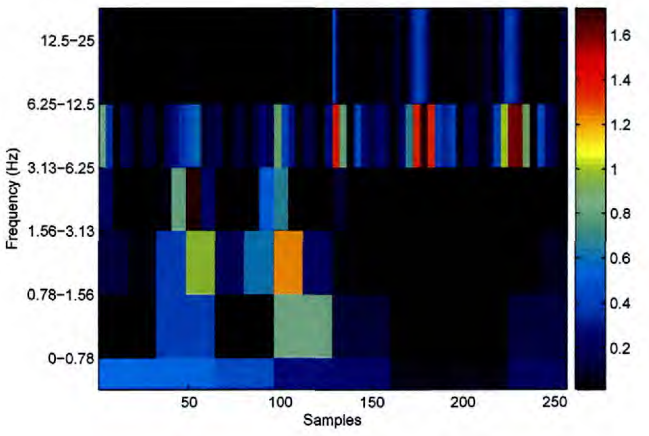
constituted of two Gaussian functions of opposite amplitude centered at 200 msec and 250 msec. Both functions have a width of 100 msec. We superimposed on this signal a true EEG signal from our data, acquired before the stimulus onset, to simulate the background EEG activity. We scaled the amplitude of this signal in order to have SNR=5dB. The sampling frequency used is 500 Hz, which produces 256 samples. The time domain representation of this signal as well as the absolute value of its continuous and discrete wavelet transforms can be seen in Figures 4.3 and 4.4.

In both cases it is obvious that the CWT provides a better insight of the time-frequency components of the signals compared with the DWT. This is due to its redundancy. This is more obvious in the second case, in which the time duration of the signal is small and the sampling frequency large. In that case the CWT managed to isolate the peaks of the two Gaussian functions, the energy of which is concentrated in the low frequencies. On the contrary, the DWT failed to isolate these components, because of its very small resolution in the low frequencies. Since the EEG signals of the oddball experiment we want to analyse are very similar to this one, concerning their duration and their frequency content, we are going to use the CWT to acquire their time-frequency decomposition.

In the results presented in the following sections of this chapter, the second derivative of the Gaussian probability density function is used as a mother wavelet. This function



(a)



(b)

Figure 4.2: Wavelet transform of the signal in Fig. 4.1 (a) CWT (b) DWT.

was not used here for the comparison between CWT and DWT as it does not have realisation of discrete filters for the implementation of DWT. However, since we decided to use the CWT in the following experiments, both the Gaussian and the spline mother wavelets were tested. The results with the Gaussian mother wavelet were generally better. We decided to present only those in order to reduce the volume of the results and make the comparison between the different methods easier.

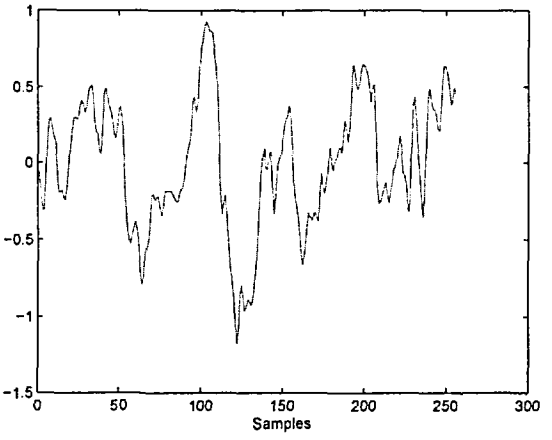
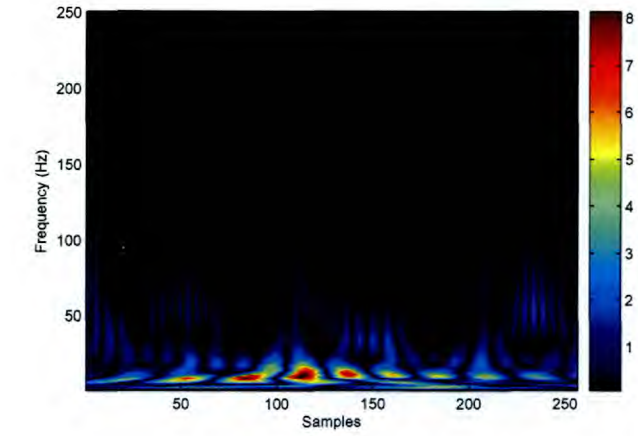


Figure 4.3: A simulated ERP signal

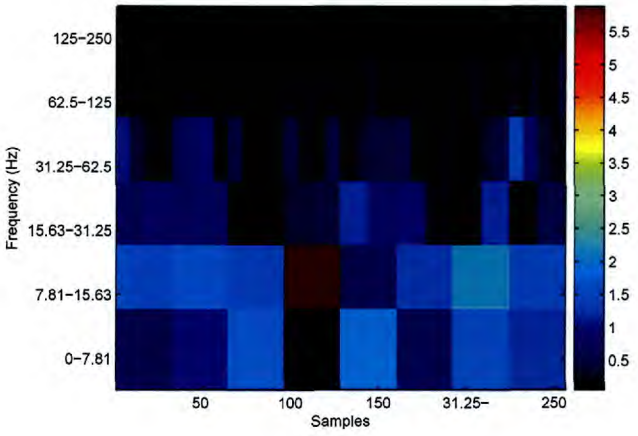
4.4 Classification in the time domain

Before using NMF to analyze the wavelet representations of the EEG signals in order to construct the features for the desired prediction, we present in this section the classification results we get using the time domain representation of the signals. As in Chapter 3, the signal of each electrode is truncated at the point corresponding to the subject's quickest reaction time. The difference here is that the starting point of the signal is chosen such that the length of the signal is 256 samples. This is in order to have a signal the length of which is a power of two and make the wavelet transform, used in the next section, easier. The number 256 is chosen because it is the power of two which makes the starting point of the signals being closest to the time point corresponding to the stimulus onset.

The features are constructed from the time signals in two ways. In the first way, each time sample is considered to be a feature, so in total 256 features are produced. In the second way, the classification capability $p(n)$ for each time sample n is computed as the percentage of correct classification produced using as only feature for the trials of the two classes the corresponding time sample. The classification task takes place using a Gaussian classifier and the trials in the training set. Then, the samples n_i for which $p(n_i)$ are maxima, are identified. For each sample n_i, the mean of the time signal over a region around it for which the values of $p(n)$ remain larger than $p(n_i) - 0.1$, is computed. The



(a)



(b)

Figure 4.4: Wavelet transform of the signal in Fig. 4.3 (a) CWT (b) DWT.

six largest means, under the constraint that the centres of their corresponding regions are at least ten samples away from each other, are kept as features.

For the classification task, a Gaussian classifier with a common diagonal covariance matrix for the two classes, see section 3.2 and Eq. (3.8), is used. The reason for assuming uncorrelated features is that in the first case, when 256 features are used, the small number of trials does not permit the estimation of an invertible covariance matrix. In the second way of feature construction (six features) where this is possible, the use of the covariance matrices in the classification (i.e. Eq. (3.6)) produced similar and in some cases worse results which are not reported. The same holds for the use of different variances for the

features in the two classes.

Let us also report here that the trials having an extreme value for a specific feature, are excluded from the computation of the statistics of the specific feature. This is done, for a given feature, by initially computing the mean m and the standard deviation s of each class, using all the trials in the training set. Then these statistics are computed again using only the trials for which the value of the feature is in the area $[m - s, m + s]$.

For the construction of the training and testing sets the "leave one out" method is used, which means that for the classification of a single trial, the training set is constituted from the rest of the trials. The procedure is repeated for all the trials and the classification rate is computed as the ratio of the correctly classified trials over their total number.

The classification results produced with the first way of feature construction are presented in Tables 4.1 and 4.2. In Table 4.1 the classification rates produced for each subject and channel are presented. In Table 4.2 the maximum classification rate and the channel for which it is produced is given for each subject. In the same table we also give the 95% confidence interval of the classification rates. This is the interval where the true classification rate lies with probability 95% and depends on the estimated classification rate and the size of the testing set. We explain in detail the way that this is computed in Appendix B.

In Tables 4.3 and 4.4 the classification results for the second way of feature construction are presented. The six features were sorted with respect to the corresponding classification capability and the classification process was repeated six times, where in the n^{th} time we use the n features with the largest classification capability. In Table 4.3 the maximum classification rate produced for a number of features for each channel and subject is given. In Table 4.4 we present the maximum classification rate produced for each subject along with the channel and the number of features for which this happens. Since we use the "leave one out method" to compute the classification rate, the training set changes in each round of classification. This means that there is a possibility that not all the times the same areas are chosen to construct a feature by averaging over the values of the samples contained in it. In order to quantify the consistency in choosing the same

Table 4.1: Time domain classification. The classification rates (%) for each channel and subject (1st way of feature production)

Subject	S1	S2	S3	S4	S5	S6	S8	S10	S11	S13	S14
Ch. Fz	72.41	62	52.63	68.09	60	63	77.78	54.69	57.81	43.94	66.67
Ch. Cz	72.41	64	52.63	62.77	53.75	70	75	50	50	56.06	68.75
Ch. Pz	56.9	66	68.42	51.06	56.25	75	66.67	65.63	62.5	53.03	77.08
Ch. F3	68.97	62	57.89	67.02	65	60	55.56	62.5	56.25	53.03	64.58
Ch. F4	72.41	54	57.89	63.83	65	66	77.78	53.13	57.81	53.03	62.5
Ch. T3	62.07	66	47.37	57.45	55	66	52.78	46.88	59.38	60.61	64.58
Ch. T4	56.9	60	55.26	56.38	57.5	53	58.33	43.75	68.75	57.58	60.42
Ch. C3	70.69	54	63.16	60.64	55	67	75	42.19	60.94	53.03	68.75
Ch. C4	72.41	66	71.05	46.81	56.25	66	72.22	51.56	62.5	59.09	72.92
Ch. T5	79.31	54	68.42	59.57	66.25	65	52.78	68.75	76.56	57.58	72.92
Ch. T6	58.62	60	68.42	63.83	66.25	46	41.67	50	70.31	54.55	68.75
Ch. P3	58.62	54	73.68	53.19	62.50	70	69.44	68.75	70.31	59.09	72.92
Ch. P4	55.17	64	60.53	51.06	63.75	65	58.33	54.69	67.19	59.09	66.67
Ch. O1	60.34	46	65.79	64.89	61.25	62	61.11	73.44	79.69	60.61	79.17
Ch. O2	60.34	64	65.79	69.15	68.75	47	61.11	65.63	75	57.58	70.83
Ch. Oz	51.72	62	65.79	61.70	56.25	54	58.33	68.75	73.44	53.03	79.17
Ch. F7	72.41	62	36.84	60.64	67.5	62	58.33	59.38	64.06	37.88	70.83
Ch. F8	75.86	64	36.84	64.89	67.5	66	63.89	56.25	67.19	39.39	64.58

Table 4.2: Time domain classification. The maximum classification rate (%) for each subject, the channel for which it is produced and the 95% confidence interval (1st way of feature production)

Subject	Classification rate (%)	Confidence interval (95%)	Channel
1	79.31	67.23 - 87.75	T5
2	66	52.15 - 77.56	Pz
3	73.68	57.99 - 85.03	P3
4	69.15	59.22 - 77.58	O2
5	68.75	57.93 - 77.85	O2
6	75	65.7 - 82.45	Pz
8	77.78	61.92 - 88.29	Fz
10	73.44	61.52 - 82.71	O1
11	79.69	68.29 - 87.73	O1
13	60.61	48.55 - 71.5	T3
14	79.17	65.74 - 88.27	O1

Table 4.3: Time domain classification. The maximum classification rates (%) produced for a certain number of features for each channel and subject (2^{nd} way of feature production)

Subject	S1	S2	S3	S4	S5	S6	S8	S10	S11	S13	S14
Ch. Fz	68.97	62	42.11	68.09	57.5	64	88.89	62.5	64.06	57.58	64.58
Ch. Cz	77.59	76	60.53	65.96	58.75	73	83.33	53.13	54.69	66.67	75
Ch. Pz	51.72	68	76.32	51.06	66.25	74	75	65.63	71.88	74.24	87.5
Ch. F3	68.97	70	60.53	65.96	58.75	66	66.67	75	60.94	66.67	70.83
Ch. F4	70.69	58	65.79	59.57	67.5	71	77.78	60.94	67.19	59.09	66.67
Ch. T3	63.79	66	39.47	50	65	71	58.33	67.19	70.31	60.61	70.83
Ch. T4	60.34	72	73.68	61.7	60	55	50	43.75	65.63	62.12	72.92
Ch. C3	65.52	62	65.79	63.83	60	70	61.11	57.81	64.06	60.61	77.08
Ch. C4	75.86	72	68.42	55.32	52.5	65	86.11	64.06	68.75	72.73	75
Ch. T5	79.31	54	81.58	54.26	70	66	58.33	64.06	78.13	68.18	87.5
Ch. T6	70.69	72	65.79	62.77	66.25	53	61.11	67.19	79.69	69.7	66.67
Ch. P3	58.62	68	71.05	58.51	63.75	71	72.22	70.31	76.56	72.73	77.08
Ch. P4	70.69	70	57.89	59.57	65	68	61.11	48.44	78.13	75.76	70.83
Ch. O1	68.97	52	65.79	68.09	68.75	68	52.78	71.88	78.13	63.64	81.25
Ch. O2	63.79	68	73.68	65.96	70	55	80.56	62.5	67.19	65.15	72.92
Ch. Oz	60.34	60	28.95	63.83	71.25	64	80.56	73.44	71.88	62.12	81.25
Ch. F7	70.69	68	71.05	65.96	66.25	68	69.44	62.50	62.5	46.97	79.17
Ch. F8	77.59	78	63.16	63.83	67.5	72	75	65.63	71.88	66.67	66.67

areas for different training sets, areas the centres of which are five or fewer samples away from each other, are considered to be instances of the same area. Then we compute the rate of a specific area being chosen for feature construction for the different training sets. The rates that concern the features producing the maximum classification rate for each subject are presented in Table 4.4.

Comparing the classification rates produced with the two ways of feature construction (using especially Tables 4.2 and 4.4) we see that the second one is superior. This is something we expected as isolating the samples with high distance and averaging in a neighbourhood to increase the robustness of the feature is better than simply considering as features all the time samples. Thus, in the next section the concept of the second way will be used for the construction of features from the NMF coefficients.

Before ending this section we also present the classification results produced using all the channels at the same time for the feature construction. The corresponding results

Table 4.4: Time domain classification. The maximum classification rate (%) for each subject along with the 95% confidence interval, the channel, the number and the appearance rate of features for which it is produced (2^{nd} way of feature production)

Subj.	Classification rate (%)	Confidence interval (95%)	Ch.	Number of features	Appearance rate
1	79.31	67.23 - 87.75	T5	4	1, 0.97, 0.97, 0.9
2	78	64.76 - 87.25	F8	3	1, 1, 1
3	81.58	66.58 - 90.78	T5	2	1, 1
4	68.09	58.12 - 76.64	Fz	6	1, 1, 0.95, 0.79, 0.76, 0.76
5	71.25	60.54 - 80.01	Oz	1	1
6	74	64.63 - 81.6	Pz	2	1, 1
8	88.89	74.69 - 95.59	Fz	2	1, 1
10	75	63.18 - 83.99	T5	2	1, 1
11	79.69	68.29 - 87.73	T6	4	1, 1 1, 0.47
13	75.76	64.19 - 84.49	P4	1	1
14	87.5	75.3 - 94.14	Pz	3	1, 0.98, 0.96

can be seen in Table 4.5. However, comparing Tables 4.4 and 4.5 we see that for all subjects, there is a channel that when used on its own, better rates are produced.

4.5 Classification analyzing the single-trial time-frequency representations with NMF

We present in this section the classification results we acquire producing a time - frequency representation of the EEG signals and then decomposing it with NMF. Our motivation for this is to study whether NMF can produce spectral components which are correlated with cognitive processes and thus differences in the time of their appearance can be used to enhance the classification rates in our problem.

The procedure we use can be described as follows. For each trial, the EEG signal of a selected electrode, truncated as explained in section 4.4 to contain 256 samples, is transformed in the time-frequency domain using the Continuous Wavelet Transform (CWT) (see sections 4.1-4.3). The second derivative of the Gaussian probability density

Table 4.5: Time domain classification. The maximum classification rate (%) for each subject when all channels are used, along with the 95% confidence interval, the number and the appearance rate of features for which it is produced.

Subj.	Classification rate (%)	Confidence interval (95%)	Number of features	Appearance rate
1	79.31	67.23 - 87.75	6	1, 0.86, 0.83, 0.83, 0.69, 0.66
2	54	40.4 - 67.03	1	0.32
3	73.68	57.99 - 85.03	2	0.74, 0.47
4	65.96	55.92 - 74.74	3	0.9, 0.59, 0.54
5	61.25	50.29 - 71.18	6	0.94, 0.74, 0.66, 0.64, 0.56, 0.55
6	69	59.37 - 77.22	3	0.96, 0.9, 0.7
8	88.89	74.69 - 95.59	4	1, 1, 0.94, 0.94
10	65.63	53.41 - 76.08	1	0.77
11	78.13	66.57 - 86.5	1	1
13	72.73	60.96 - 82	3	0.98, 0.74, 0.59
14	83.33	70.42 - 91.3	4	0.98, 0.92, 0.81, 0.6

function is used as the mother wavelet. The range of frequencies we look at is 1-40 Hz, since this is where the vast majority of the energy of the EEG signals is concentrated (see Table 3.1) and the resolution used is 1 Hz. Then we keep the absolute value of the transform, which is directly related to the evolution of the energy across time and frequency. This produces a 40×256 matrix $\boldsymbol{F}$ with the spectral information lying across its rows and the temporal information across its columns.

Then the matrix $\boldsymbol{F}$ is decomposed using NMF, i.e. $\boldsymbol{F} \approx \boldsymbol{BH}$. For more information concerning the repetitive algorithm of NMF and its general concept see section 3.5.1. In this case the columns of $\boldsymbol{B}$ carry spectral information with each one having its energy concentrated in different frequency bands. This is due to the sparsity imposed by NMF. On the other hand, the rows of $\boldsymbol{H}$ carry temporal information with each one showing the evolution of the importance of the corresponding column of $\boldsymbol{B}$ across time. The number of columns of $\boldsymbol{B}$ is chosen to be four to match the number of natural rhythms of the brain in the band 1-40 Hz (see section 1.2.3).

The motivation for the NMF analysis is that NMF may isolate spectral components

(columns of $\boldsymbol{B}$) that are correlated with different undergoing cognitive processes and thus the coefficients of the components connected with the identification of the target stimulus can be used to discriminate between the two classes. This idea has been used for music transcription in [79, 90]. In these cases the magnitude of the time-frequency transform of a musical piece is analyzed with NMF to produce a number of components carrying the spectral content of different notes. Then the corresponding coefficients (rows of $\boldsymbol{H}$), denoting the time points where the different notes appear, are used for the transcription of the musical piece.

The four columns of $\boldsymbol{B}$ are normalised and the two having maximum energy in the bands 8-13 Hz (alpha band) and 1-3 Hz (delta band) are identified. These two bands are chosen because their energy is known to be correlated with the activity preceding a subject's reaction in an oddball experiment [7,10,41,80]. Then the corresponding rows of $\boldsymbol{H}$, showing the temporal evolution of the usage of these components, are used for feature construction. Three features are constructed from each row using the method described in section 4.4, i.e. computing the classification capability for each coefficient, finding the three largest maxima and constructing the features as an average over a corresponding neighbourhood. The three features are sorted in descending order with respect to the corresponding classification capability and a feature vector of six features is constructed concatenating in turn the features of the two components. The classification task is repeated six times, where at the n^{th} time the n first features in the feature vector are used.

As in section 4.4, a Gaussian classifier with common diagonal covariance matrix for the two classes is used for the classification task. The "leave one out" method is again used for the construction of the training and testing sets. The classification task is repeated for all channels for each subject. The classification rates acquired can be seen in Table 4.6. In this table in correspondence with Table 4.3, the maximum classification rate produced for a certain number of features for each subject is presented. Finally in Table 4.7 we give the maximum classification rate for each subject, the channel and number of features for which it is produced and the appearance rate of the corresponding features. The 95% confidence interval for the classification rate is also presented. For details concerning the

Table 4.6: Single-trial time-frequency analysis with NMF. The maximum classification rates (%) produced for a certain number of features for each channel and subject.

Subject	S1	S2	S3	S4	S5	S6	S8	S10	S11	S13	S14
Ch. Fz	72.41	50	76.32	48.94	56.25	48	75	57.81	56.25	51.52	72.92
Ch. Cz	60.34	68	47.37	62.77	58.75	64	80.56	54.69	60.94	42.42	62.5
Ch. Pz	68.97	60	47.37	55.32	71.25	68	75	59.38	64.06	66.67	72.92
Ch. F3	68.97	52	55.26	55.32	67.5	42	75	53.13	46.88	66.67	79.17
Ch. F4	70.69	68	71.05	57.45	62.5	45	61.11	54.69	70.31	50	70.83
Ch. T3	48.28	68	50	56.38	63.75	62	61.11	65.63	39.06	59.09	66.67
Ch. T4	62.07	66	55.26	73.4	52.5	52	72.22	57.81	60.94	59.09	66.67
Ch. C3	53.45	66	39.47	50	46.25	64	75	70.31	51.56	56.06	68.75
Ch. C4	53.45	58	57.89	61.70	60	61	63.89	56.25	62.5	62.12	64.58
Ch. T5	55.17	62	73.68	60.64	67.5	66	77.78	65.63	50	68.18	77.08
Ch. T6	67.24	56	73.68	74.47	63.75	59	72.22	57.81	65.63	37.88	68.75
Ch. P3	72.41	44	50	44.68	75	65	61.11	67.19	59.38	72.73	72.92
Ch. P4	63.79	66	50	63.83	65	52	52.78	70.31	67.19	62.12	70.83
Ch. O1	62.07	64	65.79	68.09	56.25	53	80.56	67.19	70.31	57.58	56.25
Ch. O2	70.69	76	47.37	62.77	58.75	57	72.22	53.13	57.81	69.7	77.08
Ch. Oz	55.17	50	55.26	64.89	58.75	62	52.78	65.63	64.06	57.58	68.75
Ch. F7	72.41	56	44.74	56.38	45	56	47.22	68.75	68.75	62.12	68.75
Ch. F8	65.52	34	50	55.32	68.75	54	63.89	50	59.38	63.64	68.75

computation of the appearance rate we refer to the procedure described in section 4.4.

Comparing this method of feature construction with the equivalent one in the time domain (i.e. Tables 4.4 and 4.7) we see that in general the classification rates are not improved. In fact for 9 out of the 11 subjects the use of the raw time signals produced better results. We present a more detailed comparison of the methods in section 4.7.

4.6 Classification analysing multi-trial time-frequency representations with NMF

In the previous section the time-frequency representation of each single trial was analysed with NMF. In that case we considered as feature vector the spectral content of the EEG signal in a specific time point and we decomposed it using its different instantiations across time. In this section, for each trial, each time-frequency point is considered to be

Table 4.7: Single-trial time-frequency analysis with NMF. The maximum classification rate (%) for each subject along with the 95% confidence interval, the channel, the number and the appearance rate of features for which it is produced.

Subj.	Classif. rate (%)	Confidence interval (95%)	Ch.	Number of features		Appearance rate	
				8-13 Hz	1-3 Hz	8-13 Hz	1-3 Hz
1	72.41	59.79 - 82.24	Fz	3	2	0.97, 0.86, 0.52	1, 0.48
2	76	62.59 - 85.7	O2	3	2	1, 1, 0.96	1, 0.48
3	76.32	60.8 - 87.01	Fz	1	0	1	-
4	74.47	64.82 - 82.2	T6	2	2	1, 0.53	1, 0.71
5	75	64.52 - 83.19	P3	1	1	1	1
6	68	58.34 - 76.33	Pz	3	3	1, 1, 1	1, 0.85, 0.79
8	80.56	64.98 - 90.25	Cz	1	0	1	-
10	70.31	58.23 - 80.09	C3	1	0	1	-
11	70.31	58.23 - 80.09	F4	3	3	1, 0.97, 0.55	1, 1, 0.52
13	72.73	60.96 - 82	P3	1	0	1	-
14	79.17	65.74 - 88.27	F3	2	2	0.96, 0.79	0.96, 0.85

a different feature (variable) and the corresponding feature vector is analysed with NMF using its different instantiations across trials.

The procedure of feature construction can be described as follows. For a given channel, an initial feature vector is constructed for each trial, computing the time-frequency representation of the EEG signal, keeping its absolute value and concatenating over the columns of the matrix containing the latter representation. The details concerning the length of the EEG signal used and its time frequency representation are the same as in the previous section. Then a matrix $\boldsymbol{F}$, the columns of which are the feature vectors of the trials, is constructed and analysed with NMF as $\boldsymbol{F} \approx \boldsymbol{BH}$. The number of columns of $\boldsymbol{B}$ is chosen to be ten. As explained in section 3.5.1, due to the non negative constraints, the columns of $\boldsymbol{H}$ are sparse. This means that only a subset of the atoms (columns of $\boldsymbol{B}$) are mainly needed for the approximated reconstruction of the feature vector of a trial. Thus, if there exist atoms which are mainly needed for the reconstruction of the trials of one of the two classes, their corresponding coefficients would be suitable to be used as features

Table 4.8: Multi-trial time-frequency analysis with NMF. The maximum classification rates (%) produced for a certain number of features for each channel and subject

Subject	S1	S2	S3	S4	S5	S6	S8	S10	S11	S13	S14
Ch. Fz	79.31	76	71.05	61.7	47.5	64	66.67	53.13	60.94	60.61	68.75
Ch. Cz	75.86	64	57.89	64.89	55	69	83.33	60.94	70.31	57.58	64.58
Ch. Pz	62.07	62	63.16	62.77	65	70	80.56	64.06	73.44	66.67	79.17
Ch. F3	70.69	64	60.53	55.32	62.5	59	52.78	60.94	68.75	66.67	68.75
Ch. F4	72.41	74	55.26	60.64	67.5	66	66.67	64.06	68.75	69.7	72.92
Ch. T3	63.79	66	63.16	61.7	51.25	59	72.22	56.25	67.19	63.64	64.58
Ch. T4	62.07	66	60.53	60.64	68.75	70	66.67	45.31	60.94	62.12	64.58
Ch. C3	75.86	66	50	61.7	42.5	64	69.44	50	65.63	54.55	52.08
Ch. C4	74.14	72	63.16	64.89	68.75	60	86.11	65.63	62.5	65.15	70.83
Ch. T5	72.41	56	73.68	54.26	65	68	69.44	67.19	67.19	65.15	70.83
Ch. T6	70.69	78	68.42	63.83	66.25	59	75	53.13	70.31	60.61	66.67
Ch. P3	74.14	54	63.16	57.45	55	69	83.33	64.06	70.31	63.64	72.92
Ch. P4	68.97	74	44.74	60.64	66.25	66	72.22	71.88	70.31	68.18	70.83
Ch. O1	56.90	50	63.16	58.51	62.5	66	80.56	60.94	68.75	51.52	60.42
Ch. O2	58.62	68	55.26	63.83	63.75	66	77.78	64.06	68.75	65.15	66.67
Ch. Oz	62.07	64	65.79	53.19	66.25	59	77.78	65.63	57.81	60.61	68.75
Ch. F7	74.14	66	78.95	60.64	72.5	62	69.44	64.06	67.19	56.06	68.75
Ch. F8	70.69	60	60.53	56.38	66.25	66	75	54.69	70.31	57.58	77.08

for classification.

In consistency with the previous two sections, the classification capability is measured for each of the ten coefficients (features), using the trials in the training set. Then the classification task is repeated ten times where in the n^{th} time the n features with the highest classification capability are used. A Gaussian classifier, removing samples with extreme values and assuming uncorrelated features with common variance in the two classes, is again used. The "leave one out" method is used for the construction of the training and testing sets. The classification task was repeated for all 18 channels and in Table 4.8 we present the maximum classification rate produced for each channel and subject.

In Table 4.9 the maximum classification rate is given for each subject along with the channel and the number of features for which it was produced. Since we use the "leave one out method" to compute the classification rate, the training set changes in each round of classification. This means that there is a possibility that not all the times the coefficients of

Table 4.9: Multi-trial time-frequency analysis with NMF. The maximum classification rate (%) for each subject along with the 95% confidence interval, the channel, the number and the appearance rate of features for which it is produced.

Subj.	Classif. rate (%)	Confidence interval (95%)	Channel	Number of features	Appearance rate
1	79.31	67.23 - 87.75	Fz	3	1, 1, 0.81
2	78	64.76 - 87.25	T6	3	1, 1, 1
3	78.95	63.66 - 88.93	F7	1	1
4	64.89	54.83 - 73.78	Cz	5	1, 1, 1, 1, 1
5	72.5	61.86 - 81.08	F7	5	1, 1, 1, 1, 1
6	70	60.42 - 78.11	Pz	10	1, 1, 1, 1, 1, 1, 1, 1, 1, 1
8	86.11	71.34 - 93.92	C4	4	1, 1, 1, 0.81
10	71.88	59.87 - 81.41	P4	3	1, 1, 1
11	73.44	61.52 - 82.71	Pz	6	1, 1, 1, 1, 1, 0.7
13	69.7	57.78 - 79.45	F4	2	1, 0.92
14	79.17	65.74 - 88.27	Pz	3	1, 1, 0.94

the same components are used as features. Thus, in Table 4.9 we also give the appearance rate for the coefficients used for the most times. We observe that in the vast majority of cases the same components are chosen for the purpose of classification. In general, the classification rates are equivalent with those produced analysing single trial's time-frequency representations for the feature construction (section 4.5) and worse than those produced when the raw time signals are used (section 4.4). A more detailed comparison of the methods follows in section 4.7.

Finally, in consistency with what we did in section 4.4, we used all channels at the same time for feature construction. This is done by constructing the (magnitude) time-frequency representation for each channel of each trial and then constructing an initial feature vector for each channel by concatenating the columns of the matrix containing the latter representation. Then an initial feature vector is constructed for each trial concatenating over channels' feature vectors. Finally a matrix $\boldsymbol{F}$, the columns of which are the feature vectors of the trials, is constructed and analysed with NMF as previously. The difference is that each feature-element of $\boldsymbol{F}$ corresponds to a time-frequency-channel point instead of a time-frequency point.

Table 4.10: Multi-trial time-frequency analysis with NMF. The maximum classification rate (%) for each subject when all channels are used, the number and the appearance rate of features for which it is produced.

Subj.	Classif. rate (%)	Confidence interval (95%)	Number of features	Appearance rate
1	82.76	71.09 - 90.36	1	1
2	68	54.19 - 79.24	8	1, 1, 1, 1, 1, 1, 0.98, 0.88
3	42.11	27.85 - 57.81	3	0.89, 0.58, 0.55
4	62.77	52.67 - 71.86	1	1
5	66.25	55.36 - 75.65	3	1, 1, 1
6	67	57.31 - 75.44	1	1
8	83.33	68.11 - 92.13	4	1, 1, 1, 0.64
10	62.5	50.25 - 73.33	1	1
11	70.31	58.23 - 80.09	5	1, 1, 1, 0.98, 0.98
13	68.18	56.21 - 78.15	1	1
14	79.17	65.74 - 88.27	2	1, 0.79

The classification rates produced can be seen in Table 4.10. Comparing these with those produced when a single channel is used (see Table 4.9), we see that, apart from Subject 1, for all other subjects there is one channel, the use of which produces better results on its own.

4.7 Comparison of the proposed methods

In this section we compare the classification rates produced with the three algorithms presented in the previous sections, i.e. the one using time domain features, the one using NMF features analysing single trial time-frequency representations and the one using NMF features analysing a multi trial time-frequency representation. For this reason we compare the maximum classification rates produced for each subject with the three algorithms (see Tables 4.4, 4.7 and 4.9). Looking at these tables we see that there is not one algorithm that produces best classification rates for all the subjects universally. In Tables 4.11, 4.12 and 4.13 we present the classification rates for the subjects for which each of the algorithms produces the best classification results, respectively.

In order to quantify our confidence that a certain algorithm A produces a higher

classification rate than another algorithm B, we compute the observed level of significance of the estimated classification rates for the null hypothesis that A has the same or smaller classification rate than B. This level of significance is measured using the observed difference $r_A - r_B$ in the estimated classification rates of the two algorithms and the number of trials used to get these estimations. It is actually the (maximum) probability of observing a difference $r_A - r_B$ or larger given that the null hypothesis holds. If the observed level of significance is adequately small then we can reject the null hypothesis and be confident that A produces a better classification rate than B. For more details concerning the computation of the observed level of significance see Appendix C.

As can be seen from Tables 4.11, 4.12 and 4.13 the algorithm working in the time domain seems to be superior as it produces best classification rates for all subjects except for 4 and 5. For these subjects the best classification rates are produced from the algorithm using the NMF analysis of single trials' time-frequency representations. However, in all cases, the observed level of significance of the difference in the estimated classification rates is not sufficiently small to be confident that one algorithm is better than another. It is a common practice [21] that an observed level of significance smaller than 5% or 1% is needed to reject a null hypothesis and in our cases the levels of significance are well above these values. However, we have to notice here that the number of trials in the testing set is relatively small. This means that if two algorithms have a small difference in their correct classification rates, this difference will not be adequate to support with confidence the superiority of one of the two algorithms.

In any case, if we have to make a choice between the algorithms with the given number of trials, then the algorithm with the largest estimated classification rate should be used for each subject. The fact that the channel giving the largest classification rate is different for each subject, even when the same algorithm is used, means that a "tailor made" classifier should be designed for each subject.

Table 4.11: The subjects for which the time domain algorithm is superior than the other algorithms. The classification rates and the observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of time domain algorithm (%)	Single trial NMF algorithm		Multi trial NMF algorithm	
		Classif. rate (%)	Signif. level (%)	Classif. rate (%)	Signif. level (%)
1	79.31	72.41	19.39	79.31	50
2	78	76	40.7	78	50
3	81.58	76.32	28.91	78.95	38.81
6	74	68	17.56	70	26.52
8	88.89	80.56	16.47	86.11	36.24
10	75	70.31	27.71	71.88	34.58
11	79.69	70.31	11.07	73.44	20.32
13	75.76	72.73	34.63	69.7	21.84
14	87.5	79.17	13.78	79.17	13.78

Table 4.12: The subjects for which the single trial analysis with NMF algorithm is superior than the other algorithms. The classification rates and the observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of single trial NMF algorithm (%)	Time domain algorithm		Multi trial NMF algorithm	
		Classif. rate (%)	Signif. level (%)	Classif. rate (%)	Signif. level (%)
4	74.47	68.09	16.75	64.89	7.65
5	75	71.25	29.73	72.5	36.05

Table 4.13: The subjects for which the multi trial analysis with NMF algorithm is superior than the other algorithms. The classification rates and the observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of multi trial NMF algorithm (%)	Time domain algorithm		Single trial NMF algorithm	
		Classif. rate (%)	Signif. level (%)	Classif. rate (%)	Signif. level (%)
1	79.31	79.31	50	72.41	19.39
2	78	78	50	76	40.7

Chapter 5

Construction of trial-invariant characteristic signals

We present in this chapter a novel method to tackle the classification problem. The idea is to construct a characteristic signal for each of the two classes that remains as invariant as possible over all data of the same class and thus it may be thought of as characterising the class. This is done by constructing a weighted signal for each trial, linearly combining the EEG signals of the various channels and then computing a mean signal for each class using the trials in the training set. The weights of the above combination are chosen so that the variance of the EEG samples over the trials in the training set belonging to the same class is minimised. In order to classify an unknown trial, two characteristic signals are constructed for it, using the weights of the two classes. The distance of each one from the corresponding class characteristic signal is computed. Then the trial is classified to the class producing the smallest distance.

Since for this algorithm to be effective it is very important which EEG channels are used, we also propose a novel algorithm for channel selection. The algorithm is based on stretching and averaging the EEG signals of the available channels, to identify the ones that show the least interference from background activity. This is based on the idea that when we average signals time locked to the stimulus and the response, interfering processes will be averaged out of phase and thus high frequency components of the useful channels

will be significantly reduced, while the processes of interest will be averaged in phase and dominate the signal appearance.

We first describe the method in detail in section 5.1. An expression for the variance of each sample is constructed and is minimised subject to two different constraints. The algorithm for channel selection is described in section 5.2. We then present the classification results obtained with the two different constraints in sections 5.3 and 5.4 respectively. Both, results using all 18 channels as well as a selected subset of channels for the construction of the characteristic signals, are presented. Then in sections 5.5 and 5.6 we present the classification rates produced when features are constructed from the characteristic signals and used for the classification task. Finally, in section 5.7 we compare the algorithms proposed in this chapter.

5.1 Description of the method

In this section we describe the method we intend to use for the classification of the two classes. As mentioned earlier, we are going to construct a characteristic signal from the trials of each class, linearly combining the EEG signals of the various channels.

Let us assume that we have the EEG signals of one person's trials during an oddball experiment recorded with M channels. Half of these trials belong to class "success" and the other half to class "failure". We denote with $c_{i,j}(n)$, with $i = 1, ..., L$ and $j = 1, ..., M$, the signal of the i^{th} trial of class "success", recorded on the j^{th} channel at time n. Equivalently, we denote with $c'_{i,j}(n)$ the signals of class "failure". Let us note here that the mean value of these signals is removed and no other normalisation is applied. Our purpose is to find the set of coefficients $w_j(n)$ for class "success" that can be used to construct, for each trial, a linear combination of the channels' EEG signals, which has, at each time n, minimum variance across trials. An equivalent set of coefficients $w'_j(n)$ has to be obtained for class "failure".

Thus restricting the analysis to class "success" we have to obtain $w_j(n)$ that minimise:

$$s^2(n) = E[x^2(n)] - E[x(n)]^2 = \frac{1}{L}\sum_{i=1}^{L} x_i(n)^2 - \frac{1}{L^2}\left(\sum_{i=1}^{L} x_i(n)\right)^2 \tag{5.1}$$

for each n, where

$$x_i(n) = w_1(n)c_{i,1}(n) + ... + w_M(n)c_{i,M}(n), \quad i = 1, ..., L. \tag{5.2}$$

Since the minimisation has to take place for each time point n separately, we assume a fixed value for n and continue the analysis with $s^2 = s^2(n)$, $x_i = x_i(n)$, $w_j = w_j(n)$ and $c_{i,j} = c_{i,j}(n)$. The variance s^2 which has to be minimised can be written as:

$$\begin{aligned} s^2 =& \frac{1}{L}\sum_{i=1}^{L}\left(\sum_{j=1}^{M} w_j c_{i,j}\right)^2 - \frac{1}{L^2}\left(\sum_{j=1}^{M} w_j \sum_{i=1}^{L} c_{i,j}\right)^2 \\ =& \frac{1}{L}(\boldsymbol{Cw})^T(\boldsymbol{Cw}) - \frac{1}{L^2}(\mathbf{1}_L^T\boldsymbol{Cw})^T(\mathbf{1}_L^T\boldsymbol{Cw}) \\ =& \frac{1}{L}\boldsymbol{w}^T\boldsymbol{C}^T\boldsymbol{Cw} - \frac{1}{L^2}\boldsymbol{w}^T\boldsymbol{C}^T\mathbf{1}_L\mathbf{1}_L^T\boldsymbol{Cw} \\ =& \boldsymbol{w}^T\left(\frac{1}{L}\boldsymbol{R_{cc}} - \frac{1}{L^2}\boldsymbol{A}\right)\boldsymbol{w} \\ =& \boldsymbol{w}^T\boldsymbol{Rw} \end{aligned} \tag{5.3}$$

where

$$\boldsymbol{C}^{(L\times M)} = \begin{pmatrix} c_{1,1} & \dots & c_{1,M} \\ \vdots & & \vdots \\ c_{L,1} & \dots & c_{L,M} \end{pmatrix}, \quad \boldsymbol{w}^{(M\times 1)} = \begin{pmatrix} w_1 \\ \vdots \\ w_M \end{pmatrix}, \quad \mathbf{1}_L^{(L\times 1)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{5.4}$$

$$\boldsymbol{R_{cc}} = \boldsymbol{C}^T\boldsymbol{C}, \quad \boldsymbol{A} = \boldsymbol{C}^T\mathbf{1}_L\mathbf{1}_L^T\boldsymbol{C}, \quad \boldsymbol{R} = \frac{1}{L}\boldsymbol{R_{cc}} - \frac{1}{L^2}\boldsymbol{A}$$

For the minimisation of Eq.(5.3) a constraint should be introduced for $\boldsymbol{w}$, as for the unconstrained case and a non singular matrix $\boldsymbol{R}$, it is obvious that we obtain the useless solution $\boldsymbol{w} = \mathbf{0}$.

We consider two different constraints for the minimisation of s^2. In the first case

we choose to minimise s^2 subject to:

$$\sum_{j=1}^{M} w_j = 1 \Leftrightarrow \boldsymbol{w}^T \mathbf{1}_M - 1 = 0 \tag{5.5}$$

The minimisation of Eq.(5.3) with respect to $\boldsymbol{w}$ subject to Eq.(5.5) can be solved using the method of Lagrange multipliers [3]. According to this method the wanted w_j are given by the minimum point of the Lagrangian function which is defined as:

$$\Lambda(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T \boldsymbol{R} \boldsymbol{w} + \lambda(\boldsymbol{w}^T \mathbf{1}_M - 1) \tag{5.6}$$

where λ is an additional variable called Lagrange multiplier.

Thus taking the first derivatives of $\Lambda(\boldsymbol{w}, \lambda)$ with respect to all its arguments and setting them equal to zero, we have to solve the system of the following $M + 1$ equations:

$$\frac{\partial \Lambda}{\partial \boldsymbol{w}} = 2\boldsymbol{R}\boldsymbol{w} + \lambda \mathbf{1}_M = \mathbf{0} \tag{5.7}$$

$$\frac{\partial \Lambda}{\partial \lambda} = \boldsymbol{w}^T \mathbf{1}_M - 1 = 0 \tag{5.8}$$

Solving Eq.(5.7) with respect to $\boldsymbol{w}$ we find $\boldsymbol{w} = -\frac{\lambda}{2}\boldsymbol{R}^{-1}\mathbf{1}_M$ and substituting into Eq.(5.8) we get $\lambda = -\frac{2}{\mathbf{1}_M^T (\boldsymbol{R}^{-1})^T \mathbf{1}_M}$. Therefore the wanted vector $\boldsymbol{w}$ is:

$$\boldsymbol{w} = \frac{\boldsymbol{R}^{-1}}{\mathbf{1}_M^T (\boldsymbol{R}^{-1})^T \mathbf{1}_M} \mathbf{1}_M \tag{5.9}$$

Note that this solution corresponds to the case in which we consider the elements of $\boldsymbol{w}$ as weights that have to sum up to 1.

In the second case we minimise s^2 subject to:

$$\sum_{j=1}^{M} w_j^2 = 1 \Leftrightarrow \boldsymbol{w}^T \boldsymbol{w} = 1 \tag{5.10}$$

When this constraint is used, vector $\boldsymbol{w}$ is considered as a direction on which the projections of all data vectors have as invariant value as possible. In this case, since matrix $\boldsymbol{R}$ is

symmetric, see its definition in Eq.(5.4), s^2 is minimised when $\boldsymbol{w}$ equals the eigenvector corresponding to the minimum eigenvalue of $\boldsymbol{R}$ [82]. Thus, the wanted vector $\boldsymbol{w}$ can be found computing the eigenvalue decomposition of $\boldsymbol{R}$.

Let us mention here that $\boldsymbol{R}$ is actually the covariance matrix of the random variables r_j, $j = 1, ..., M$, where r_j denotes the value of the potential recorded on channel j and has different instantiations for the trials of class "success". This means that the procedure is actually a PCA on the random variables r_j, keeping only the smallest eigenvector. This is natural as we know that this is the vector minimising the variance of the projected variables. Thus, subject to the constraint of Eq.(5.10), x_i for $i = 1, ..., L$ are instances of the smallest principal component of variables r_j.

As mentioned earlier, this procedure is followed for every time point n, so we finally end up with the set of coefficients $w_j(n)$ for class "success". The same procedure can be followed using the trials of class "failure" to get the equivalent coefficients $w'_j(n)$. Then we can compute the mean signals $\overline{x}(n)$, $\overline{x}'(n)$ for classes "success" and "failure" respectively, as:

$$\overline{x}(n) = w_1(n)\overline{c}_1(n) + ... + w_M(n)\overline{c}_M(n)$$

$$\overline{x}'(n) = w'_1(n)\overline{c}'_1(n) + ... + w'_M(n)\overline{c}'_M(n)$$

where $\overline{c}_j(n)$, $\overline{c}'_j(n)$, $j = 1, ..., M$ are the mean values of the EEG signals of the two classes, at channel j and time point n, computed over the trials of each class, respectively.

In order to classify an unknown trial, we compute the equivalent signals $\widetilde{x}(n)$, $\widetilde{x}'(n)$ for each of the two classes using the EEG signals of the unknown trial and the coefficients of each class. Then we compute the Euclidean distance between $\widetilde{x}(n)$ and $\overline{x}(n)$ and between $\widetilde{x}'(n)$ and $\overline{x}'(n)$ and the trial is classified to the class producing the smallest distance.

5.2 A proposed algorithm for channel selection

One can use all the available EEG channels to tackle the classification problem. However, it is likely that selecting a subset of channels, the signals of which are particularly correlated with the undergoing process of recognising the target, will boost the classification results.

In the literature of Human Performance Monitoring [45,70,71,86] the subset of the midline frontal, central and parietal channels (Fz, Cz, Pz) is used as the most suitable for feature construction. Moreover physiological studies such as [40] also report that the signal of these channels exhibit correlations with the subjects' time of response in an oddball experiment. However, in all these studies the whole length of EEG signals is used, whereas in our cases the signals are truncated at an early stage before the subjects' response has taken place. Thus, due to the difficulty of our problem, a channel selection algorithm applied on each subject separately maybe more appropriate. The need of identifying the "useful" channels on a single subject basis, has also been reported in [48], in a study concerning channel selection algorithms for BCI.

We propose here an automated method for evaluating the usefulness of a channel for studies concerning oddball experiments. In order to do that, we consider that there must be a characteristic process that takes place in the brain of the subject recorded by each channel. This process sometimes happens faster and sometimes more slowly. In all cases, however, it is superimposed with many other processes taking place in the brain of the subject at the same time. All these other processes may happen in a variety of phases, so if we average many trials these processes may be averaged out. On the other hand, the process of interest will dominate, as it will be always averaged in phase.

The difficulty arises from the fact that in each trial we have two fixed points: the point the stimulus was shown and the point the action was taken, and that the time span between these two points is of a different duration for each trial and thus is sampled by a different number of samples. To identify the process that actually is repeated between these two fixed points from trial to trial, we must have the same number of samples over all trials and average them. To perform this averaging we consider the trial for each subject with the longest reaction time and upsample all other trials to have the same number of samples. In a sense, we stretch the time of short trials in order to make all trials last the same "time", in the hope that the process we are seeking to view will become apparent by the out of phase averaging of all other interfering processes. Note that now we are using all trials we have for a subject, including those with the average time response which we

had omitted from either class "success" or class "failure".

In order to perform the upsampling of the trials we use Lagrange interpolation. According to Lagrange interpolation the value y of an interpolated point at time x is computed using the polynomial:

$$y = \frac{(x-x_2)(x-x_3)...(x-x_N)}{(x_1-x_2)(x_1-x_3)...(x_1-x_N)}y_1 + \frac{(x-x_1)(x-x_3)...(x-x_N)}{(x_2-x_1)(x_2-x_3)...(x_2-x_N)}y_2 + ...+ \\ + \frac{(x-x_1)(x-x_2)...(x-x_{N-1})}{(x_N-x_1)(x_N-x_2)...(x_N-x_{N-1})}y_N, \tag{5.11}$$

where $y_1, ..., y_N$ are the values of N neighbours at time points $x_1, ..., x_N$. For the computation of each interpolated point we used six neighbouring samples, three preceding and three following the point. This approach has been used in [76] for the upsampling of heart signals.

After all signals of each channel have been truncated at the two points of interest, stretched and averaged, we have to identify the channels that exhibit some indication of useful information. Mean signals that have high degree of fluctuation, i.e. high frequency components, may be thought of as rather noisy: the interfering components have not really been removed effectively. Channels exhibiting some degree of smoothness are most likely to contain the sought out recording of the process of interest. In order to quantify the smoothness of a mean signal f, we perform a low pass filtering to construct a signal f_{LP}, using an elliptic IIR low pass filter with a cut off frequency of 13Hz. This frequency is chosen because the band 0-13 Hz contains the Delta, Theta and Alpha rhythms which are known to be related to the cognitive processes occurring in an oddball experiment [7] and contribute in the formation of the main components of the ERP signal [10, 41]. We then remove the low pass signal from the original one and compute the ratio of the energy of the resulting signal over the energy of the original one, i.e. $IS \equiv \frac{\|f-f_{LP}\|^2}{\|f\|^2}$. We call this the "index of smoothness" and is considered as a quantity characterising the usefulness of a channel. The closer to zero its value is, the more smooth the signal is and thus the more useful the corresponding channel.

In table 5.1 the channels sorted in ascending order with respect to the index of smoothness, are presented for each subject separately. Let us note here that the compu-

Table 5.1: The EEG channels presented in decreasing order with respect to their importance for each subject as indicated by the proposed algorithm. The channels Fz, Cz, Pz, the signals of which are considered by the literature to be correlated with the reaction to stimulus cognitive processes are presented in bold.

S1	S2	S3	S4	S5	S6	S8	S10	S11	S13	S14
Cz	**Pz**	**Pz**	**Cz**	**Cz**	Oz	**Cz**	O1	**Fz**	O2	**Pz**
Fz	**Cz**	P4	**Fz**	**Fz**	**Cz**	**Pz**	**Pz**	T5	**Pz**	P4
T5	**Fz**	**Cz**	O1	**Pz**	C4	P3	T5	**Cz**	Oz	P3
F4	P4	P3	C4	T6	**Pz**	**Fz**	**Fz**	C3	**Cz**	**Cz**
F3	C4	C4	P3	T5	**Fz**	P4	P3	P3	P4	T6
C3	F4	T6	**Pz**	P4	C3	C4	C4	T6	P3	O1
O1	P3	**Fz**	F4	Oz	P4	O1	C3	**Pz**	**Fz**	Oz
Oz	F8	Oz	T5	P3	F4	C3	P4	P4	C4	O2
Pz	T4	O1	P4	O2	O2	F4	Oz	T3	T6	C4
P3	F7	O2	Oz	C4	P3	O2	T6	F7	F4	**Fz**
C4	O2	F4	C3	O1	O1	F3	F4	F3	C3	T5
F8	Oz	F3	F3	C3	T5	Oz	F8	F4	O1	F4
P4	C3	T5	F8	F4	F3	T6	F3	O2	T5	F3
F7	O1	T4	T4	F3	T6	T5	**Cz**	C4	T3	C3
O2	T5	C3	F7	T4	F8	T4	F7	F8	T4	T3
T6	F3	F8	T6	F8	F7	T3	O2	Oz	F3	F7
T3	T3	F7	O2	T3	T4	F7	T3	T4	F8	T4
T4	T6	T3	T3	F7	T3	F8	T4	O1	F7	F8

tation of the index of smoothness of each channel was done using all subject's available trials, including the 50% of the trials having medium reaction times. One can see that the ranking of channels is different among subjects, as expected. However, channels Fz, Cz, Pz, the signals of which, as we said above, are considered in the literature to be correlated with the cognitive processes of recognising a target stimulus, are ranked in high positions for the vast majority of subjects. This is a strong indication that the proposed algorithm succeeds in indicating the channels, the signals of which exhibit correlations with the underlying cognitive processes.

5.3 Classification results using the whole characteristic signals with weights that sum up to 1

We present in this section the classification results we get, using the method described in section 5.1. As far as the construction of the training and testing sets is concerned, we use the "leave one out" method, as in the previous chapters. We construct here the

characteristic signal for each class, finding the vector of coefficients $\boldsymbol{w}$ which minimises the variances subject to the first constraint, i.e.: $\sum_{j=1}^{M} w_j = 1$.

5.3.1 A universal set of channels used

We first use a common set of channels for all the subjects to construct the signals for the classification. We consider two types of sets. The first one is constituted from all the available 18 channels and the second one from channels Fz, Cz, Pz, which are considered in the literature as suitable for studies concerning an oddball experiment. The correct classification rates produced are presented in Table 5.2. In the same table we give the 95% confidence interval for the correct classification rate of each subject. This is the interval where the true correct classification rate lies with a probability of 95%. For more details concerning the way that this interval is computed from the estimated correct classification rate produced by the classification task, see Appendix B.

Subjects 3 and 8 constitute special cases as for these subjects there are not enough trials in the training set to construct an invertible matrix $\boldsymbol{R}$ (see its definition in Eq. (5.4)). In order to permit the inversion of $\boldsymbol{R}$, we discard channel F8 for subject 3 and channels F7 and F8 for subject 8. The selection of these channels is done because they are the last in the order in which the data was given. The results using a more sophisticated way for the selection of channels are presented in section 5.3.2.

As we can see in Table 5.2, the correct classification rates produced, when all channels are used, are very low. It is worth indicating that the lower bound of the 95% confidence interval is below 50%, i.e. random performance, for the majority of subjects. The same observation holds when channels Fz, Cz, Pz, although there is a slight improvement for some subjects.

5.3.2 Channel selection applied on a single subject basis

We then perform the classification experiments using the channel selection algorithm presented in section 5.2. The classification procedure takes place as follows. First the "index of smoothness" is computed for every channel and the channels are sorted with respect to

Table 5.2: The correct classification rates when all channels or channels Fz, Cz, Pz are used (when $\sum_{j=1}^{M} w_j = 1$) and the corresponding 95% confidence intervals. These are the intervals where the true correct classification rates lie, with a probability of 95%.

Subj.	All channels		Channels: Fz, Cz, Pz	
	Classification rate (%)	Confidence Interval	Classification rate (%)	Confidence Interval
1	53.45	40.80 - 65.67	56.9	44.12 - 68.82
2	54	40.40 - 67.03	66	52.15 - 77.56
3	36.84	23.38 - 52.71	55.26	39.71 - 69.85
4	52.13	42.15 - 61.94	55.32	45.26 - 64.96
5	56.25	45.34 - 66.59	70	59.23 - 78.94
6	60	50.20 - 69.06	43	33.73 - 52.78
8	72.22	56.01 - 84.15	75	58.93 - 86.25
10	53.13	41.07 - 64.82	51.56	39.58 - 63.37
11	40.63	29.46 - 52.85	43.75	32.29 - 55.91
13	51.52	39.71 - 63.15	40.91	29.87 - 52.95
14	47.92	34.47 - 61.67	62.5	48.36 - 74.78

their importance (see Table 5.1. We then perform the classification task 17 times where in the i^{th} time we use the $i+1$ most important channels for classification. As usual, the "leave one out" method is used for the classification task. Let us note here that the computation of the "index of smoothness" takes place once, using all the available trials. This means that in each round of the classification task the testing trial has been considered for the above computation. However, during this computation the trials are not divided into classes, so there is no use of the knowledge of the class where the testing trial belongs to, in the training process.

The classification rates produced across the number of "important" channels used can be seen in Figure 5.1(a). The equivalent rates when the set of channels Fz, Cz, Pz is used are denoted with a dotted straight line. Observing the classification rate across the number of channels used we see that there is not a tendency common for all subjects and methods of classification. However, in the majority of cases the rates are maximised for a small/moderate number of "important" channels. This means that it is usually better to reduce the number of channels used for feature construction instead of using all the available ones. Moreover, one can see that the set of channels Fz, Cz, Pz is not the

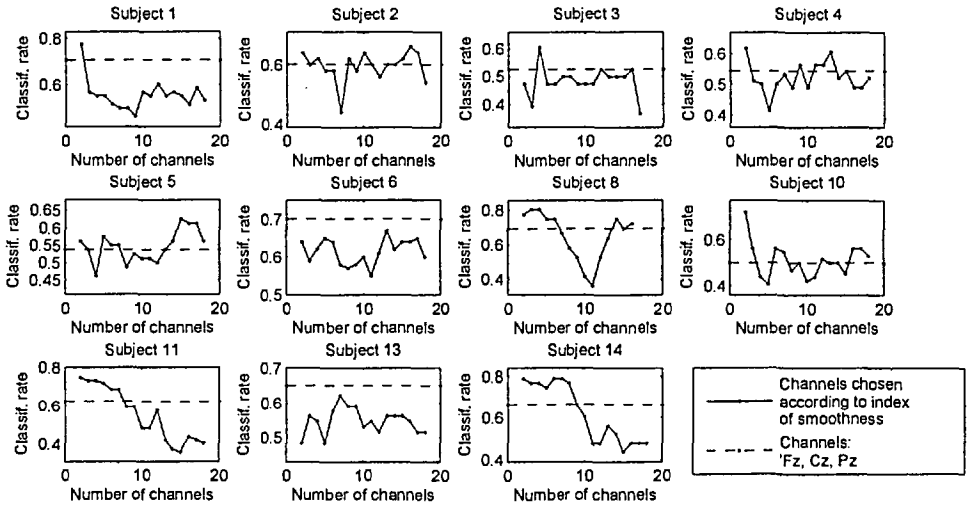
Table 5.3: The maximum correct classification rates for a selected subset of channels (when $\sum_{j=1}^{M} w_j = 1$) the names of these channels and the corresponding 95% confidence intervals. These are the intervals where the true correct classification rates lie, with a probability of 95%

Subject	Classification rate (%)	Confidence interval (95%)	Channels
1	77.59	65.34 - 86.41	Cz, Fz
2	66	52.15 - 77.56	Pz, Cz, Fz, P4, C4, F4, P3, F8, T4, F7, O2, Oz, C3, O1, T5, F3
3	60.53	44.72 - 74.4	Pz, P4, Cz, P3
4	61.7	51.6 - 70.89	Cz, Fz
5	62.5	51.55 - 72.31	Cz, Fz, Pz, T6, T5, P4, Oz, P3, O2, C4, O1, C3, F4, F3, T4
6	67	57.31 - 75.44	Oz, Cz, C4, Pz, Fz, C3, P4, F4, O2, P3, O1, T5, F3
8	80.56	64.97 - 90.25	Cz, Pz, P3
10	71.88	59.87 - 81.41	O1, Pz
11	75	63.18 - 83.99	Fz, T5
13	62.12	50.06 - 72.85	O2, Pz, Oz, Cz, P4, P3, Fz
14	79.17	65.74 - 88.27	Pz, P4

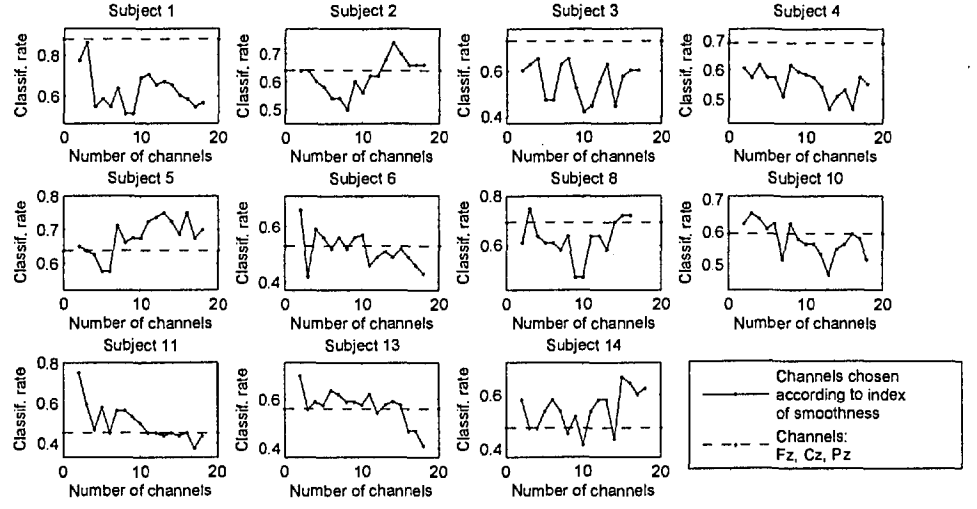
ideal one as in 9 out of 11 cases there is a subset of channels producing better results. This again supports the need of a channel ranking algorithm, which will indicate for each subject separately, the channels correlated with the ongoing cognitive processes.

We present in Table 5.3 the maximum correct classification rates we get for a specific number of channels used for each subject, the corresponding 95% confidence intervals, as well as the names of the channels producing these results. We also present in Figure 5.2 the mean characteristic signal for each of the two classes, using for each subject the channels producing the best classification rates. Let us note here that all available trials have been used for the computation of these signals, so the actual mean characteristic signals constructed in each round of the "leave one out" method in the classification process may have slightly different shapes.

Comparing Tables 5.2 and 5.3 one can make the same observations as looking



(a)



(b)

Figure 5.1: The correct classification rates across the number of "important" channels used for each subject. The dotted lines indicate the correct classification rates produced when the subset of channels Fz, Cz, Pz, is used. (a) Constraint: $\sum_{j=1}^{M} w_j = 1$. (b) Constraint: $\sum_{j=1}^{M} w_j^2 = 1$.

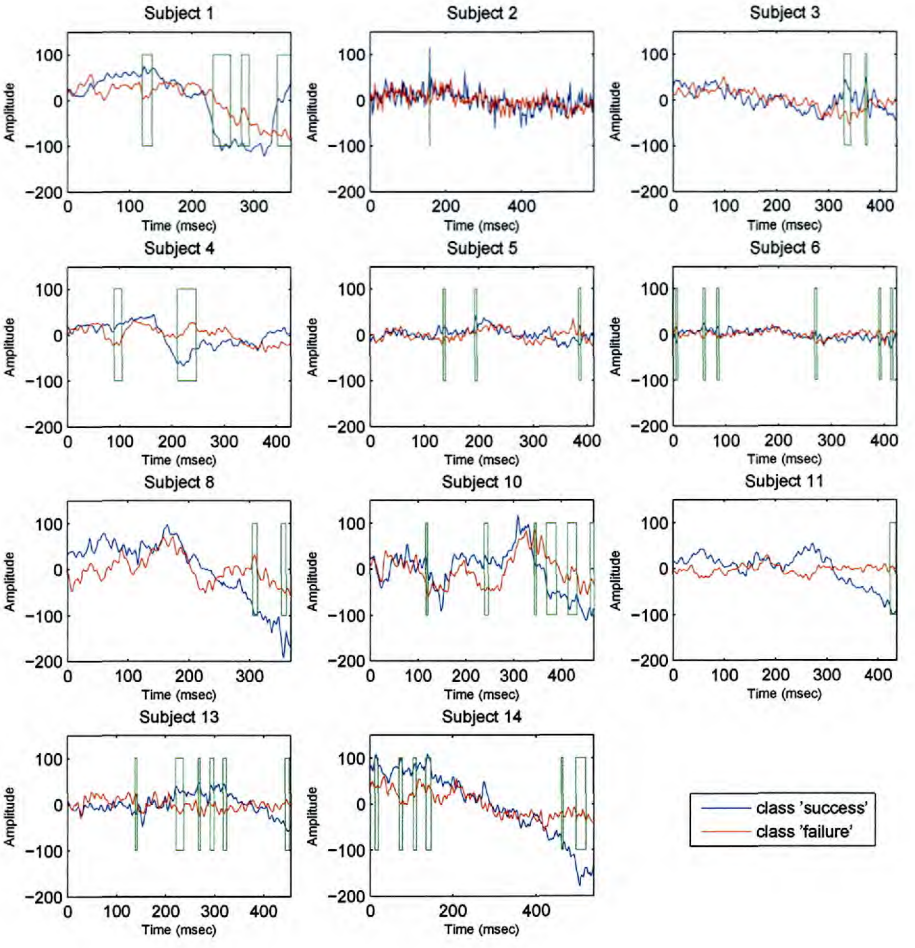


Figure 5.2: The mean characteristic signal for class "success" (blue) and class "failure" (red) for the 11 subjects, using all available trials. For each subject the channels producing the best classification rates (see Table 5.3) are used. The areas used for feature construction (see section 5.5) are marked with green rectangles. Constraint used: $\sum_{j=1}^{M} w_j = 1$

at Figure 5.1, i.e. the selection of a subset of channels prior to the application of our method increases the correct classification rates in general. Another observation is that the channels producing best results are different for each subject. This supports our belief that a "tailor-made" system has to be designed for each subject.

5.4 Classification results using the whole characteristic signals with squared weights that sum up to 1

We present in this section the classification results we get, constructing the characteristic signal for each class, finding the vector of coefficients $\boldsymbol{w}$ which minimises the variance subject to the second constraint, i.e.: $\sum_{j=1}^{M} w_j^2 = 1$.

5.4.1 A universal set of channels used

As we did in section 5.3 we first construct the characteristic signals using all 18 channels or the subset of channels Fz, Cz, Pz. The classification rates acquired as well as the corresponding 95% confidence intervals are presented in Table 5.4. As explained in section 5.3, we discarded channel F8 for subject 3 and channels F7 and F8 for subject 8 in order to have a non singular matrix $\boldsymbol{R}$.

In consistency with what we observed when the first constraint was used (see Table 5.2) the classification rates we acquire when all channels are used are very low. When the set of channels Fz, Cz, Pz is used there is a considerable improvement for Subjects 1, 3 and 4 but for the rest of them the rates remain low.

5.4.2 Channel selection applied on a single subject basis

We then apply the method described in section 5.3.2 to sort the channels according to their "usefulness" and apply the proposed method using a selected subset of channels. We perform the classification task 17 times where in the i^{th} time the $i+1$ most important channels are used. The produced correct classification rates across the number of "important" channels used are presented in Figure 5.1(b). The classification rates acquired with

Table 5.4: The correct classification rates when all channels or channels Fz, Cz, Pz are used (when $\sum_{j=1}^{M} w_j^2 = 1$) and the corresponding 95% confidence intervals. These are the intervals where the true correct classification rates lie, with a probability of 95%.

Subj.	All channels		Channels: Fz, Cz, Pz	
	Classification rate (%)	Confidence Interval	Classification rate (%)	Confidence Interval
1	56.9	44.12 - 68.82	87.93	77.12 - 94.03
2	66	52.15 - 77.56	64	50.14 - 75.86
3	60.53	44.72 - 74.41	73.68	57.99 - 85.03
4	55.32	45.26 - 64.96	69.15	59.22 - 77.58
5	70	59.23 - 78.94	63.75	52.81 - 73.43
6	43	33.73 - 52.78	53	43.29 - 62.49
8	72.22	56.01 - 84.15	69.44	53.14 - 81.99
10	51.56	39.58 - 63.37	59.38	47.15 - 70.55
11	43.75	32.29 - 55.91	45.31	33.73 - 57.42
13	40.91	29.87 - 52.95	56.06	44.08 - 67.37
14	62.5	48.36 - 74.78	47.92	34.47 - 61.67

the set of channels Cz, Pz, Fz are indicated with a dotted line. We again observe that there is always a subset of channels producing better results than the ones produced when all channels are used. The same holds for the comparison with the case when channels Cz, Pz, Fz, with the exception of Subjects 3 and 4.

In Table 5.5 we present the maximum correct classification rates produced for a certain number of channels, the names of these channels and the corresponding 95% confidence intervals. One can see that the channels producing the best results are again different across subjects. Moreover as we see from Tables 5.3 and 5.5 the subset of channels producing best results for the same subject is generally different for the two constraints. The same also stands for the shape of the mean characteristic signals (see Figure 5.3). A comparison between the classification rates produced with the two constraints is presented in section 5.7.

Table 5.5: The maximum correct classification rates for a selected subset of channels (when $\sum_{j=1}^{M} w_j^2 = 1$) the names of these channels and the corresponding 95% confidence intervals. These are the intervals where the true correct classification rates lie with a probability of 95%

Subject	Classification rate (%)	Confidence interval (95%)	Channels
1	86.21	75.07 - 92.84	Cz, Fz, T5
2	74	60.45 - 84.13	Pz, Cz, Fz, P4, C4, F4, P3, F8, T4, F7, O2, Oz, C3, O1
3	65.79	49.89 - 78.79	Pz, P4, Cz, P3
4	61.7	51.6 - 70.89	Cz, Fz, O1, C4
5	75	64.52 - 83.19	Cz, Fz, Pz, T6, T5, P4, Oz, P3, O2, C4, O1, C3, F4
6	66	56.28 - 74.54	Oz, Cz
8	75	58.93 - 86.25	Cz, Pz, P3
10	65.63	53.4 - 76.08	O1, Pz, T5
11	75	63.18 - 83.99	Fz, T5
13	69.7	57.78 - 79.45	O2, Pz
14	66.67	52.54 - 78.32	Pz, P4, P3, Cz, T6, O1, Oz, O2, C4, Fz, T5, F4, F3, C3, T3

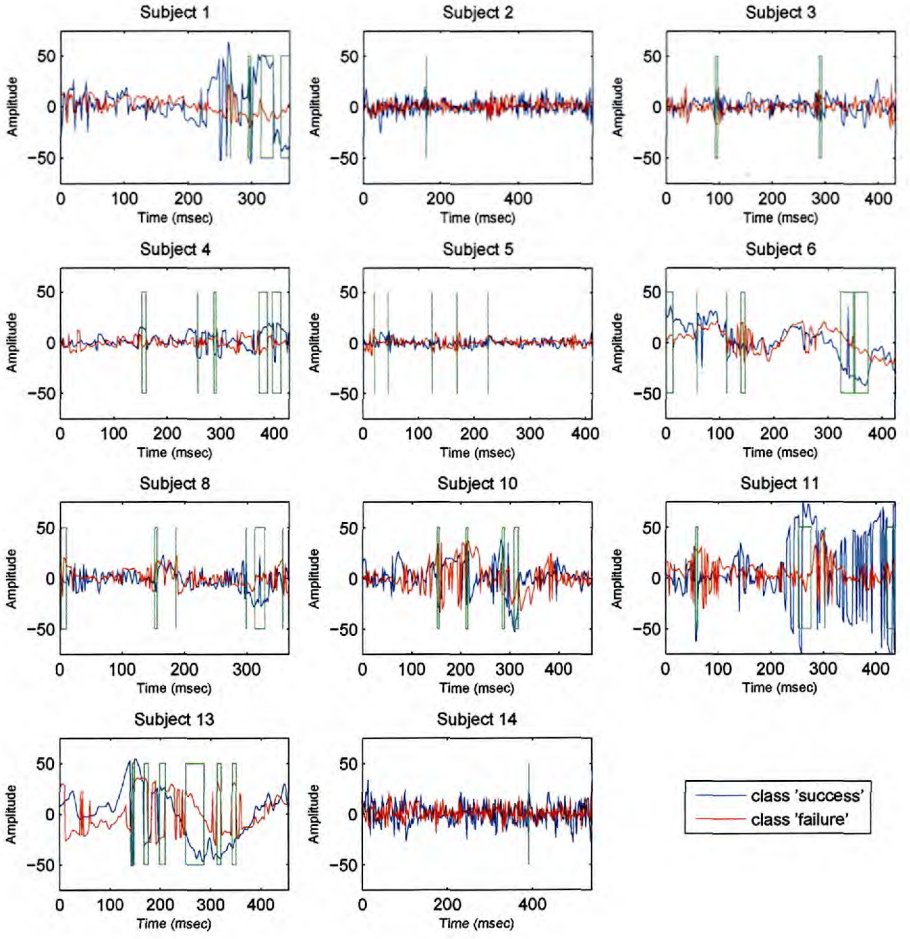


Figure 5.3: The mean characteristic signal for class "success" (blue) and class "failure" (red) for the 11 subjects, using all available trials. For each subject the channels producing the best classification rates (see Table 5.5) are used. The areas used for feature construction (see section 5.6) are marked with green rectangles. Constraint used: $\sum_{j=1}^{M} w_j^2 = 1$

5.5 Classification results constructing features from characteristic signals with weights that sum up to 1

In the previous sections we used the whole length of the characteristic signals to perform the desired classification which means that we considered each sample of the signals as a distinct feature. We present here the classification results we get when we construct a number of features from the characteristic signals and use these features for classification.

The feature construction can be described as follows. We first construct the mean characteristic signals for the two classes as described in section 5.1. We then compute the absolute difference between the two signals and we find its maxima. For each maximum m, an area around it the samples of which remain larger than $0.9m$, is identified. Then the mean over this area in the characteristic signals of the two classes is computed. The six means corresponding to the six largest maxima, under the constraint that the centres of their corresponding areas are at least ten samples away from each other, are selected as features. This procedure is equivalent to the one used in sections 4.4 and 4.5. Thus, we have finally constructed two feature vectors $\overline{f}$ and $\overline{f}'$ of size six, for classes "success" and "failure", respectively.

The procedure described above takes place using only the data in the training set, as it only involves the trial-invariant signals, which are produced from these data. Once the starting and ending points of the areas that are used for the feature construction have been identified, they are used to construct features for the trials in the testing set. For each trial in the testing set two feature vectors $\widetilde{f}$ and $\widetilde{f}'$ are constructed through averaging over the selected areas of its characteristic signals. $\widetilde{f}$ is produced from the characteristic signal constructed using the coefficients of class "success" and $\widetilde{f}'$ from the one using the coefficients of class "failure". In order to classify the trial we compute the Euclidean distance between $\widetilde{f}$ and $\overline{f}$ and between $\widetilde{f}'$ and $\overline{f}'$ and the trial is classified to the class producing the smallest distance.

In order to perform the classification task we used the "leave one out" method as usual to construct the training and testing sets. For the construction of the characteristic

signals the first constraint is used, i.e. $\sum_{j=1}^{M} w_j = 1$. The subset of channels used for each subject is the one producing maximum classification rates when the whole signals are used (section 5.4.2) and can be seen in Table 5.3. During the feature production the six features were sorted in descending order with respect to the value of their corresponding maxima in the difference between the two mean characteristic signals, which is assumed to denote their importance. We performed the classification task six times where in the i^{th} time the i most important features were used. The maximum classification rates we got for a certain number of features each time are presented in Table 5.6. The areas used for the construction of these features, when all available trials in classes "success" and "failure" are taken into account, are presented in Figure 5.2.

As we have noticed in sections 4.4 and 4.5 the use of the "leave one out method" means that the training set changes in each round of classification. In order to quantify the consistency in choosing the same areas for the feature construction for different training sets, areas the centres of which are five or fewer samples away from each other, are considered to be instantiations of the same area. We then compute the ratio of selection of each area for feature construction over the different training sets. The rates that concern the features producing the maximum classification rate for each subject are presented in Table 5.6.

Comparing the classification rates we get here with the ones when the whole characteristic signals are used (see Table 5.3) we see that the rates are improved for all subjects with the exception of subjects 5, 10, and 14. Another observation is that the ratios of selection of the areas used for the feature construction are high. This is encouraging as it implies that the process of feature construction for each subject generally does not depend on the constitution of the training set.

Table 5.6: The maximum correct classification rates and the corresponding 95% confidence intervals when features are constructed from the trial-invariant signals (when $\sum_{j=1}^{M} w_j = 1$), the number of features producing these rates and the ratios of selection of these features.

Subject	Classification rate (%)	Confidence interval (95%)	Number of features	Ratio of selection
1	81.03	69.15 - 89.07	4	1, 1, 1, 1
2	66	52.15 - 77.56	1	0.88
3	68.42	52.54 - 80.92	2	1, 1
4	67.02	57.01 - 75.69	2	1, 0.98
5	57.5	46.57 - 67.74	3	1, 0.93, 0.88
6	70	60.42 - 78.11	6	1, 1, 1, 0.9, 0.88, 0.76
8	91.67	78.17 - 97.13	2	1, 1
10	54.69	42.57 - 66.27	6	1, 1, 1, 0.97, 0.91, 0.72
11	78.13	66.57 - 86.5	1	1
13	69.7	57.78 - 79.45	6	1, 1, 1, 1, 0.94, 0.55
14	83.33	70.42 - 91.3	6	1, 1, 0.96, 0.96, 0.92, 0.67

5.6 Classification results constructing features from characteristic signals with squared weights that sum up to 1

In this section we present the results we get when the characteristic signals, which are used to produce the features, are constructed using the second constraint, i.e. $\sum_{j=1}^{M} w_j^2 = 1$. The procedure of constructing the features from the characteristic signals as well as the details of performing the classification task is exactly the same as the one described in section 5.5. The channels used for the construction of the characteristic signals for each channel are those producing the best results when the whole signal is used in section 5.4.2 and can be seen in Table 5.5.

We therefore give the maximum correct classification rates as well as the number of features for each subject used to produced them, in Table 5.7. We also present the ratio of selection of the areas used for the feature construction. These areas, when all trials of the two classes are taken into account for the construction of the class characteristic signals, are marked with green rectangles in Figure 5.3.

Table 5.7: The maximum correct classification rates and the corresponding 95% confidence intervals when features are constructed from the trial-invariant signals (when $\sum_{j=1}^{M} w_j^2 = 1$), the number of features producing these rates and the ratios of selection of these features.

Subject	Classification rate (%)	Confidence interval (95%)	Number of features	Ratio of selection
1	81.03	69.15 - 89.07	4	1, 1, 0.83, 0.79
2	54	40.4 - 67.03	1	0.64
3	55.26	39.71 - 69.85	2	0.74, 0.53
4	64.89	54.83 - 73.78	5	1, 1, 0.98, 0.96, 0.89
5	66.25	55.36 - 75.65	5	1, 1, 0.98, 0.98, 0.83
6	65	55.25 - 73.64	6	1, 1, 1, 0.9, 0.74, 0.68
8	66.67	50.33 - 79.79	6	1, 0.94, 0.89, 0.89, 0.83, 0.78
10	64.06	51.82 - 74.71	4	1, 0.97, 0.97, 0.91
11	75	63.18 - 83.99	4	1, 1, 0.97, 0.78
13	62.12	50.06 - 72.85	6	1, 1, 1, 1, 0.97, 0.91
14	64.58	50.44 - 76.57	1	0.33

Comparing the results produced here, with those acquired when the whole characteristic signals were used (see Table 5.5) we see that the classification rates produced here are inferior for most of the subjects. Moreover the ratio of selection of the areas for the feature construction is considerably low in several cases. This indicates that the algorithm is not reliable as the feature construction process varies with the changes in the training set.

5.7 Comparison of the proposed methods

In this section we compare the algorithms proposed in this chapter in an attempt to conclude which one is superior. These algorithms are: the algorithm using the whole characteristic signal to perform classification using the 1st constraint ($\sum_{j=1}^{M} w_j = 1$, section 5.3.2) , the equivalent algorithm using the 2nd constraint ($\sum_{j=1}^{M} w_j^2 = 1$, section 5.4.2), the algorithm using features constructed from the characteristic signal to perform classification using the 1st constraint (section 5.5) and the equivalent algorithm using the 2nd constraint (section 5.6).

Table 5.8: The subjects for which the algorithm using the whole signal (when $\sum_{j=1}^{M} w_j = 1$) produces better classification rates than the other algorithms. The classification rates and the observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Algorithm using whole signal $\sum_{j=1}^{M} w_j = 1$	Algorithm using whole signal $\sum_{j=1}^{M} w_j^2 = 1$		Algorithm using features $\sum_{j=1}^{M} w_j = 1$		Algorithm using features $\sum_{j=1}^{M} w_j^2 = 1$	
	Classif. rate (%)	Classif. rate (%)	Signif. level (%)	Classif. rate (%)	Signif. level (%)	Signif. level (%)	Classif. rate (%)
10	**71.88**	65.63	22.41	**54.69**	**2.1**	64.06	17.26

Table 5.9: The subjects for which the algorithm using the whole signal (when $\sum_{j=1}^{M} w_j^2 = 1$) produces better classification rates than the other algorithms. The classification rates and the observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Algorithm using whole signal $\sum_{j=1}^{M} w_j^2 = 1$	Algorithm using whole signal $\sum_{j=1}^{M} w_j = 1$		Algorithm using features $\sum_{j=1}^{M} w_j = 1$		Algorithm using features $\sum_{j=1}^{M} w_j^2 = 1$	
	Classif. rate (%)	Classif. rate (%)	Signif. level (%)	Classif. rate (%)	Signif. level (%)	Signif. level (%)	Classif. rate (%)
1	86.21	77.59	11.45	81.03	22.69	81.03	22.69
2	**74**	66	19.29	66	19.29	**54**	**1.75**
5	**75**	**62.5**	**4.36**	**57.5**	**0.9**	66.25	11.26
13	69.7	62.12	18.02	69.7	50	62.12	18.02

In Tables 5.8 - 5.10 the subjects for which each of the above algorithms produces the best classification rates are presented respectively (the algorithm producing features using the second constraint does not produce better results than the other algorithms for any subject). In each of these Tables we present the corresponding classification rates as well as the observed levels of significance of the hypothesis that the other algorithms are equivalent or superior than the one that produced the best results. As we explained in section 4.7 (and in more detail in Appendix C), when the level of significance is below 5% we can reject the null hypothesis and we can be confident (at a 95% level) that the algorithm which produced better results is actually superior. In the opposite case we lack such confidence and the algorithms should be considered equivalent.

Table 5.10: The subjects for which the algorithm uses features constructed from the characteristic signal (when $\sum_{j=1}^{M} w_j = 1$) produces better classification rates than the other algorithms. The classification rates and the observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Algorithm using features $\sum_{j=1}^{M} w_j = 1$	Algorithm using whole signal $\sum_{j=1}^{M} w_j = 1$		Algorithm using whole signal $\sum_{j=1}^{M} w_j^2 = 1$		Algorithm using features $\sum_{j=1}^{M} w_j^2 = 1$	
	Classif. rate (%)	Classif. rate (%)	Signif. level (%)	Classif. rate (%)	Signif. level (%)	Signif. level (%)	Classif. rate (%)
3	68.42	60.53	23.84	65.79	40.48	55.26	11.98
4	67.02	61.7	22.4	61.7	22.4	64.89	37.96
6	70	67	32.47	66	27.3	65	22.6
8	**91.67**	80.56	8.66	**75**	**2.74**	**66.67**	**0.34**
11	78.13	75	33.91	75	33.91	75	33.91
13	69.7	62.12	18.02	69.7	50	62.12	18.02
14	**83.33**	79.17	30.24	**66.67**	**2.87**	**64.58**	**1.7**

As can be seen from Tables 5.8 - 5.10 there is not one algorithm producing best results for all subjects. The algorithm producing best results for most of the subjects (7 out of 11) is the one constructing features from the characteristic signal using the 1^{st} constraint. However, the 95% level of confidence for the superiority of one of two algorithms holds only for few cases (they are indicated with bold letters). Although this fact does not permit us to be confident for the superiority of the latter algorithm, we should keep in mind that the number of trials in the testing set is relatively small. As commented in section 4.7 this means that if two algorithms have a small difference in their true classification rates, then it is likely that this difference will not be adequate to support with confidence the superiority of one over the other.

Chapter 6

A comparison with methods from the field of Human Performance Monitoring

In this chapter, we apply to the prediction problem a number of methods proposed in the closely related field of Human Performance Monitoring (HPM). In section 6.1, we briefly describe the HPM problem, focusing on the differences with the one in our case. Let us note here that the relevant literature has already been presented in Chapter 2. We then describe the methods we apply to the prediction problem and present the acquired results. These results are compared in section 6.2 with the ones acquired with some of the main methods proposed in this thesis.

6.1 Methods from the Human Performance Monitoring Field

The most relevant, to the problem we are tacking, field in the literature is the one of Human Performance Monitoring during critical tasks. These tasks, e.g. air traffic control or military applications, usually involve an operator detecting various signals on a screen, evaluating them and proceeding into appropriate actions. The errors are very rare so it

is not possible to evaluate the operator's performance through them, however just one error could have very serious consequences. HPM aims to use the operator's ongoing EEG signals to detect when their performance related to factors such as the reaction time, the accuracy and the confidence falls beyond an acceptable level. In such cases the operator could be replaced by someone else and a potential error could be avoided.

As in our case, studies in HPM use data acquired from an oddball experiment, in which a subject monitors a screen and has to respond upon the presentation of a specific stimulus. Parameters concerning the performance of their response are recorded and then combined to produce a performance factor. Then this factor has to be estimated using the subject's recorded EEG signals. However, on contrast to the prediction problem we are studying in this thesis, there is no restriction on the length of the EEG signals used. This permits the use of the most important components of the Event Related Potential, such as the P300, which are known to be correlated with the reaction time [40]. Moreover, since the estimation of the performance has to take place for a time duration in the near past and not on a single trial basis, averaging of the EEG signals over a window of a number of trials can be used. This is again important as averaging can enhance the ERP which is heavily buried into the background EEG activity (averaging of several trials has actually been used in all studies of HPM referenced in section 2.2).

Although the absence of the two restrictions presented above make the HPM problem much easier than the one of prediction, the methods proposed for HPM can still be used in our case. The relevant literature of HPM has been presented in section 2.2. Since approaches using PCA [45] and wavelets [86] have already been used in this thesis, we decided to apply to the prediction problem the Kernel PCA (KPCA) method in combination with Support Vector Regression, proposed in [70,71]. Since our problem is a classification and not a regression one, Support Vector Classification is used here. Moreover, the Gaussian classifier, used throughout this study, is used here as well for comparative purposes. Finally, apart from KPCA, linear PCA (LPCA) and time features are also used.

In the rest of this section, we first give a brief description of the new methods used (Kernel Principal Component Analysis and Support Vector Classification) and then

present the acquired results.

6.1.1 Kernel PCA

Kernel Principal Component Analysis (KPCA) [40] uses a non linear function to map the original vectors of variables (features) $\boldsymbol{x}_i$, belonging to a space $\mathbb{R}^M$, to a new space $\mathcal{F}$ of a higher dimensionality M' ($M < M' \leq \infty$). PCA is then applied on the mapped data $\boldsymbol{\phi}(\boldsymbol{x})$. The key point of the algorithm is that there is no need to explicitly compute the mapped data as KPCA can be computed knowing only the inner products between them. The inner product between two instances $\boldsymbol{\phi}(\boldsymbol{x}_i)$, $\boldsymbol{\phi}(\boldsymbol{x}_j)$ of mapped data can be computed using a function k, named *kernel* function, which takes as input vectors $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ in their original form, i.e.:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{\phi}(\boldsymbol{x}_i).\boldsymbol{\phi}(\boldsymbol{x}_j)) \tag{6.1}$$

The advantage of KPCA is that the elements of the vectors belonging to the higher dimensionality space $\mathcal{F}$, are usually constructed by taking higher order correlations between the input variables. This is likely to enhance the classification capability of the (non linear) principal components extracted from them compared with that of the linear case. The type of correlations used depends on the chosen mapping, i.e. the choice of the kernel function. We use here the Gaussian kernel function defined as:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{L}} \tag{6.2}$$

This kernel is proposed in [71] to address the HPM problem due to its smoothness properties. In the rest of this section we describe how KPCA works.

Let us assume we have N feature vectors $\boldsymbol{x}_i$, $i = 1, ..., N$, on the initial space $\mathbb{R}^M$. The covariance matrix of their mapped version can be written as $\boldsymbol{C} = \frac{1}{N}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$, where $\boldsymbol{\Phi}$ is an $M \times N$ matrix the i^{th} column of which is the vector $\boldsymbol{\phi}(\boldsymbol{x}_i)$. The PCA on space $\mathcal{F}$ lies in finding the eigenvectors $\boldsymbol{v}_k$ of $\boldsymbol{C}$ corresponding to eigenvalues $l \geq 0$ (we are interested

only in the positive ones), i.e. solve:

$$lv = Cv \Leftrightarrow lv = \frac{1}{N}\Phi\Phi^{\mathrm{T}}v \tag{6.3}$$

In the case of positive eigenvalues, the eigenvectors lie in the space spanned by the columns of Φ, i.e. $v_k = \Phi\alpha_k$. Based on this notice it can be proven [71] that as far as the positive eigenvalues and corresponding eigenvectors are concerned, the eigenvalue problem of Eq.(6.3) is equivalent to the one of:

$$Nl\alpha = \Phi^{\mathrm{T}}\Phi\alpha \Leftrightarrow \lambda\alpha = K\alpha \tag{6.4}$$

For the computation of the $N \times N$ matrix K the explicit knowledge of Φ is not needed as its elements can be computed through:

$$K^{(i,j)} = (\phi(x_i).\phi(x_j)) = k(x_i, x_j) \tag{6.5}$$

After solving the eigenvalue problem of Eq.(6.4), we find the eigenvectors α_k corresponding to non zero eigenvalues λ_k. The maximum number of these vectors is N. Since the norm of the non linear eigenvectors v_k should be equal with one, the vectors α_k should be normalised to satisfy:

$$v_k^{\mathrm{T}}v_k = 1 \Leftrightarrow \alpha_k^{\mathrm{T}}\Phi^{\mathrm{T}}\Phi\alpha_k = 1 \Leftrightarrow \lambda_k\alpha_k^{\mathrm{T}}\alpha_k = 1 \tag{6.6}$$

After the normalisation of the vector α_k, the k^{th} non linear principal component of a feature vector x can be computed as:

$$b(x)^{(k)} = v_k^{\mathrm{T}}\phi(x) = \alpha^{\mathrm{T}}\Phi^{\mathrm{T}}\phi(x) = \sum_{i=1}^{N}\alpha_k^{(i)}k(x_i, x) \tag{6.7}$$

Thus it is obvious from Eq.(6.7) that the non linear principal components are computed without the explicit knowledge of the mapped vectors $\phi(x)$ in the higher dimensional space. Their maximum number is N, i.e. equal with the number of instances. This means

that, since usually $N < M'$, we only "see" a subspace of $\mathcal{F}$. In practice we select the first $p < N$ components describing a certain amount of the data variance and use them for the classification task.

6.1.2 Support vector classification

The Support Vector Machine (SVM) [87] is a tool performing classification in the following way. Let us suppose we have a two class problem and a training set $\boldsymbol{x}_i, i = 1, ..., N$ of N instances of feature vectors $\boldsymbol{x} \in \mathbb{R}^M$. Let us also associate a scalar $y_i \in \{1, -1\}$ with each instance $\boldsymbol{x}_i$ denoting the class where the latter belongs to. SVM performs the classification task using a hyperplane, i.e. a subspace of dimension $M - 1$, defined as:

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0 \tag{6.8}$$

The hyperplane divides the feature space into two halves and an unknown feature vector $\widetilde{\boldsymbol{x}}$ is classified according to which half it belongs to, i.e.

$$y(\widetilde{\boldsymbol{x}}) = sgn(\boldsymbol{w}^{\mathrm{T}}\widetilde{\boldsymbol{x}} + b) \tag{6.9}$$

In the case that the instances in the training set are linearly separable, the separating hyperplane is found as the one which classifies the data without error and maximises the distance from the closest feature vector. Since the same hyperplane can be described through Eq.(6.8) with many different $\boldsymbol{w}$ pointing to the same direction, the norm of $\boldsymbol{w}$ is constrained to be equal to the inverse of the distance of the nearest feature vector in the training set to the hyperplane. Such hyperplanes are called canonical and they satisfy:

$$\min_i |\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b| = 1 \tag{6.10}$$

A canonical hyperplane which separates the data without errors should satisfy the following relation:

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1, \quad i = 1, ..., N \tag{6.11}$$

Moreover in order for such a hyperplane to be optimal it has to maximise the margin m between the two classes. Noticing that the distance of a vector $\boldsymbol{x}_i$ from a hyperplane $(\boldsymbol{w}, b)$ is computed as $d_{(\boldsymbol{w},b)}(\boldsymbol{x}_i) = \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|}$ and that the distances of the hyperplane from the closest vector of each class should be the same, the margin m is given by:

$$\begin{aligned} m(\boldsymbol{w}, b) &= \min_{\boldsymbol{x}_i:y_i=-1} d_{(\boldsymbol{w},b)}(\boldsymbol{x}_i) + \min_{\boldsymbol{x}_i:y_i=1} d_{(\boldsymbol{w},b)}(\boldsymbol{x}_i) \\ &= \min_{\boldsymbol{x}_i:y_i=-1} \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|} + \min_{\boldsymbol{x}_i:y_i=1} \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|} \\ &= \frac{1}{\|\boldsymbol{w}\|}\left(\min_{\boldsymbol{x}_i:y_i=-1} |\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b| + \min_{\boldsymbol{x}_i:y_i=1} |\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b|\right) \\ &= \frac{2}{\|\boldsymbol{w}\|} \end{aligned} \tag{6.12}$$

Maximising Eq.(6.12) is equivalent to minimising:

$$\phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\| \tag{6.13}$$

Thus the separating hyperplane that SVM uses is the one occurring from the minimisation of Eq.(6.13) subject to the constraint of Eq.(6.11). This optimisation problem is solved using Lagrange multipliers and for more details we refer to [87].

In the case that the data in the training set is not linearly separable, a set of parameters $\xi_i \geq 0, i = 1, ..., N$ is introduced to account for the classification errors. Thus Eq.(6.11) that the separating hyperplane should satisfy, is modified to:

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, ..., N \tag{6.14}$$

At the same time the cost function of Eq.(6.13) is modified to:

$$\phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{N} \xi_i \tag{6.15}$$

in order to account for the minimisation of the classification error measured by $\sum_{i=1}^{N} \xi_i$. C is a scalar parameter adjusting the trade-off between maximising the margin between the two classes and minimising the classification error. It is seen from the optimisation

process of Eq.(6.15) subject to Eq.(6.14) [87] that the larger the value of C is, the closer the produced hyperplane is to the one of the separable case.

6.1.3 Classification Results

Before performing any classification task we computed the index of smoothness for each channel with respect to each subject separately, as described in section 5.2. This was done using all the available trials, including the 50% of the trials having medium reaction times. Then the channels were sorted in ascending order with respect to their index of smoothness as seen in Table 5.1. We performed the classification experiments 18 times, where in the i^{th} time the feature vectors, for the classes "success" and "failure", were constructed concatenating the signals of the subject's i most important channels. The EEG signals were first filtered with a lowpass filter at a cut off frequency of 25Hz and then decimated to correspond to a sampling rate of 50Hz. This was done in order to reduce the amount of features used and be consistent with [71] where the use of KPCA is proposed. After the downsampling the dc component was removed in order to have signals with a zero mean value.

The use of three types of feature vectors is studied. In the first case the original feature vectors containing the raw time samples, constructed as described above are used. In the other two cases LPCA and KPCA are applied on the time features and the new feature vectors are constructed using the first largest principal components accounting for the 99% of the data variance. As far as the width L of the Gaussian kernel is concerned the experiments are repeated 10 times, choosing L uniformly from the range $(0.2\sigma^2 M, 5\sigma^2 M)$, where σ^2 is the variance of the EEG samples over all time points and channels used and M the length of the original time feature vectors.

As far as the construction of the training and testing set is concerned we use the "leave one out" method. The support vector and the Gaussian classifier are used for the classification task. The experiments were repeated five times when the support vector classifier was used with the parameter C taking values from the set $\{0.1, 1, 10, 100, \infty\}$. The results reported here are the best ones, produced when $C = 0.1$ for the time and

LPCA features and when $C = \infty$ for the KPCA features. When the Gaussian classifier was used, a common diagonal covariance matrix was constructed for the two classes (see Eq.(3.8).

The correct classification rates produced across the number of "important" channels used can be seen in Figure 6.1 for SVM and in Figure 6.2 for Gaussian classification. We also computed the correct classification rates when the set of channels Fz, Cz, Pz is used. We repeat here that these channels were used in the HPM study in [71], as the signals of them are considered in the literature to exhibit correlations with the cognitive processes of recognising a target. The classification rates when these channels are used are denoted in the figures with a straight dotted line. Observing the classification rate across the number of channels used we see that in the majority of cases the rates are maximised for a small/moderate number of "important" channels. Moreover, for all subjects, this set of channels produces better results than the ones when Fz, Cz, Pz are used. These observations are qualitatively similar to the ones of chapter, 5 when classification is performed using the class characteristic signals, and imply the need of an individual channel selection.

In order to compare the six methods resulting from the different combinations of feature construction and classification method, we present in Table 6.1 the maximum correct classification rates produced per subject for each methods combination. The rate which is maximum for each subject is indicated in bold. Although the differences are not very big, we observe a consistent superiority of the KPCA features for 9 out of 11 subjects. This is in consistency with the work in [71] where the superiority of KPCA over LPCA and raw time features for the human performance monitoring problem is reported. Concerning the comparison between the two types of classifiers we see that remarkably the Gaussian classifier is better for seven subjects, for three of them the two classifiers produce the same maximum rates (for different type of features) and for only one subject the SVM produces best results. This is not however the case if we restrict the comparison in time features for which SVM is superior to Gaussian classification with 7 over 4 subjects. Based on these two observations we conclude that it is worth using the more complicated SVM algorithm

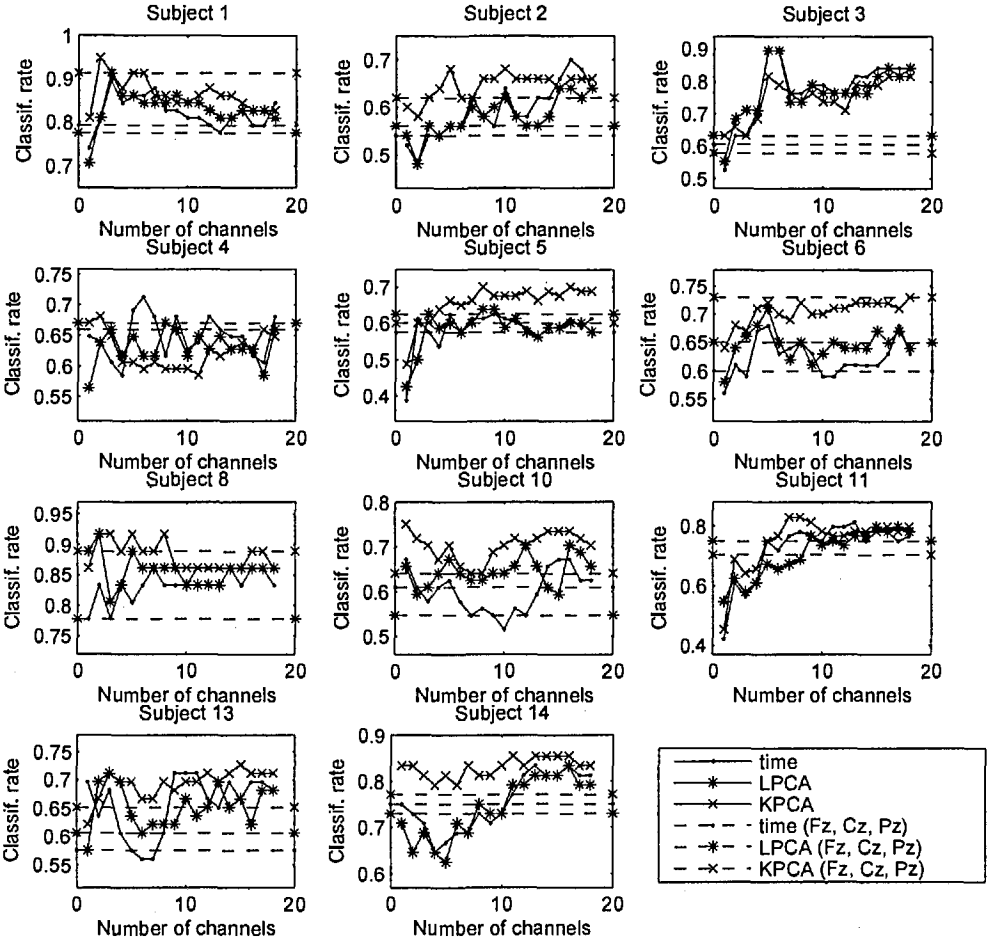


Figure 6.1: Correct classification rates vs number of channels used. Time, LPCA, KPCA features with Support Vector classification. The dotted horizontal lines denote the classification rates produced when channels Fz, Cz, Pz are used.

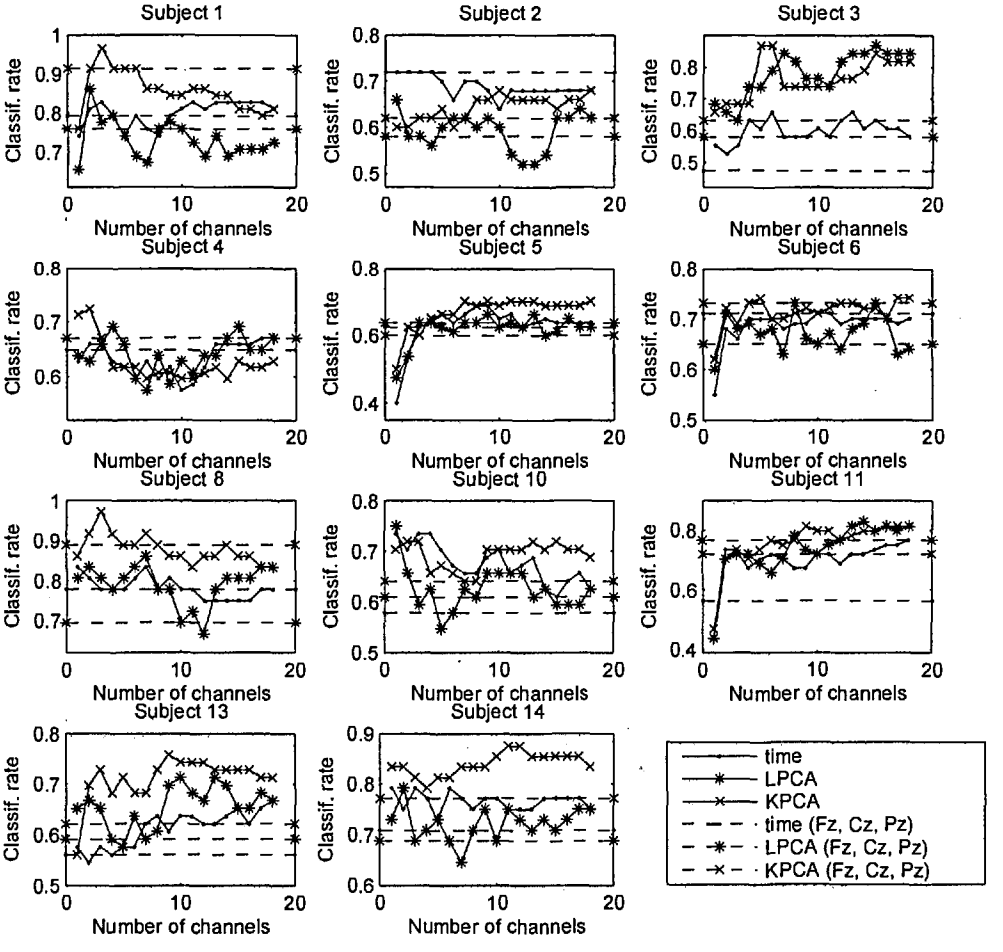


Figure 6.2: Correct classification rates vs number of channels used. Time, LPCA, KPCA features with Gaussian Classification. The dotted horizontal lines denote the classification rates produced when channels Fz, Cz, Pz are used.

Table 6.1: The correct classification rates for time, LPCA, KPCA features and Support Vector and Gaussian type of classification.

Subj.	SVM			Gaussian Classif.		
	Time	LPCA	KPCA	Time	LPCA	KPCA
S1	89.66	91.38	94.83	82.76	86.21	**96.55**
S2	70.00	64.00	68.00	**72.00**	66.00	68.00
S3	**89.47**	**89.47**	81.58	65.79	86.84	86.84
S4	71.28	67.02	68.09	67.02	69.15	**72.34**
S5	62.50	63.75	**70.00**	68.75	66.25	**70.00**
S6	68.00	71.00	73.00	71.00	73.00	**74.00**
S8	86.11	91.67	91.67	83.33	86.11	**97.22**
S10	67.19	70.31	**75.00**	73.44	**75.00**	71.88
S11	81.25	79.69	**82.81**	76.56	**82.81**	81.25
S13	71.21	71.21	72.73	66.67	71.21	**75.76**
S14	83.33	83.33	85.42	79.17	79.17	**87.50**

for classification when raw time features are used, but the simpler Gaussian classification algorithm is equivalent or even better when the time features have been processed with LPCA and especially with KPCA.

6.2 Comparison of various methods

Throughout this study a variety of methods were used to tackle the classification problem of predicting a person's performance in an oddball experiment. Since none of them appears to produce best classification rates for all subject universally we are restricted here in comparing the real novel ones. These methods are:

- NMF for analysis of single trial's time-frequency representation for feature construction combined with Gaussian Classification (section 4.5).
- NMF for analysis of multi trial's time-frequency representation for feature construction combined with Gaussian Classification (section 4.6).
- Classification using class characteristic signals with squared weights that sum up to 1 (section 5.4).
- Classification using features constructed from class characteristic signals with weights that sum up to 1 (section 5.5).

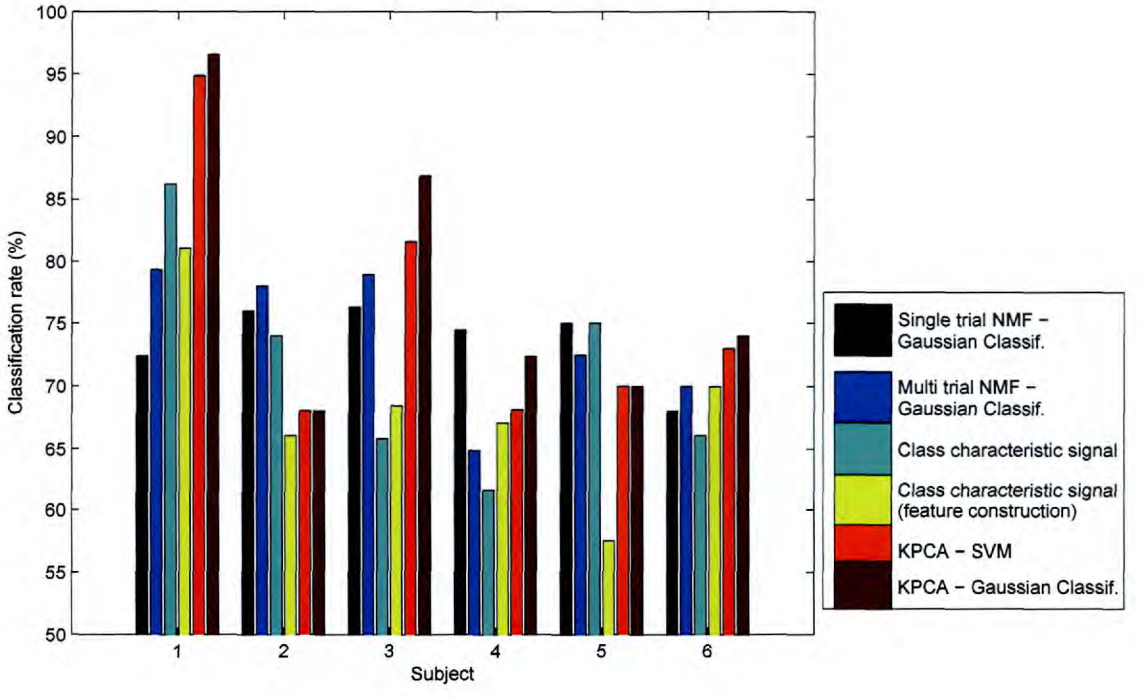


Figure 6.3: The correct classification rates for Subjects S1, S2, S3, S4, S5, S6 produced by various methods.

Together with these methods we also consider, for comparison purposes, the classification results produced with the Kernel PCA methods proposed for HPM, i.e. use of KPCA features with SVM or Gaussian Classification (section 6.1). The classification results produced with all these methods are presented in the bar Figures 6.3 and 6.4 for subjects 1-6 and 8-14, respectively.

Observing Figure 6.3 and 6.4 we see that there is not one algorithm producing best classification rates for all subjects. However the KPCA features seem to be superior as for 6 out of the 11 subjects, KPCA features combined with the Gaussian classifier seems to be optimal. The same holds for two more subjects when KPCA features are combined with Support Vector Classification. For the rest three subjects some of the other other methods produces the best classification rates. However, the differences in the estimated classification rates are generally small and as can be seen in Tables 6.2- 6.6, for few cases (they are indicated in bold) we can be confident at a level of 95% that the algorithm

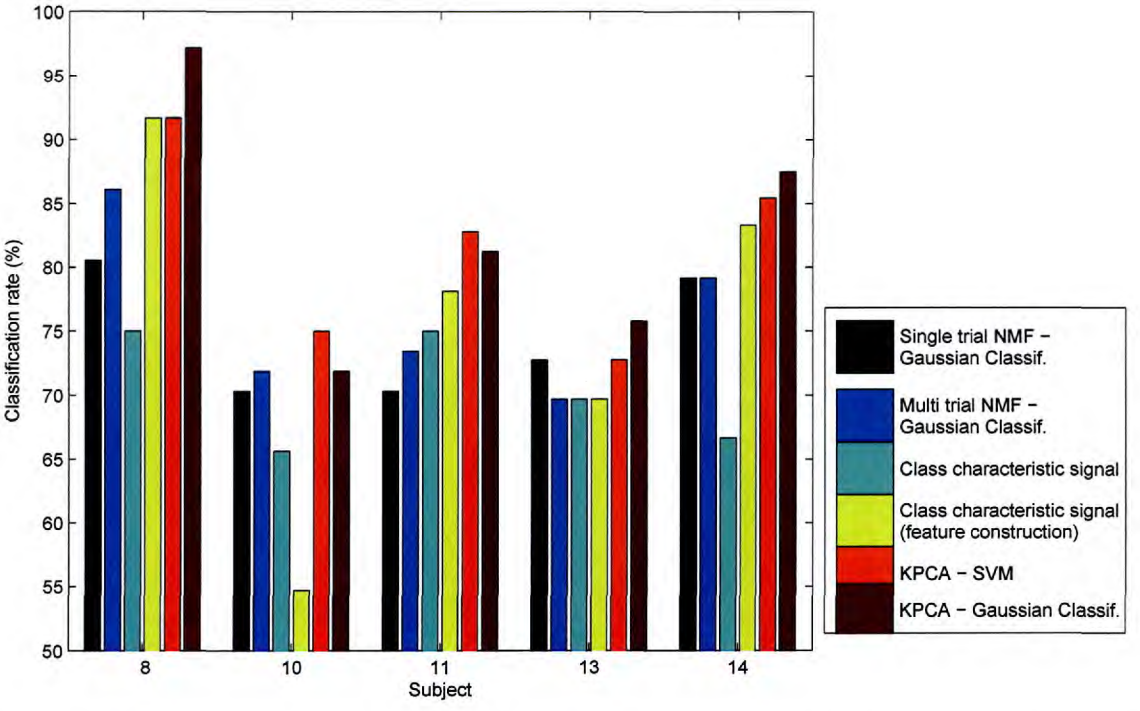


Figure 6.4: The correct classification rates for Subjects S8, S10, S11, S13, S14 produced by various methods.

Table 6.2: The subjects for which the method using NMF of single trial's time-frequency representations and Gaussian classification produces better rates than the other algorithms. The observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of single trial NMF - Gauss. (%)	Significance Level (%)				
		Multi trial NMF-Gauss.	Class charact. signal	Class charact. signal (feat. constr.)	Time-SVM	Time-Gauss.
4	74.47	7.66	**2.97**	13.14	16.74	37.13
5	75	36.05	50	**0.9**	24.04	24.04

Table 6.3: The subjects for which the method using NMF of multi-trial's time-frequency representations and Gaussian classification produces better rates than the other algorithms. The observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of multi-trial NMF - Gauss. (%)	Significance Level (%)				
		Single-trial NMF-Gauss.	Class charact. signal	Class charact. signal (feat. constr.)	Time-SVM	Time-Gauss.
2	78	40.7	32.13	9.1	13.09	13.09

estimating the best results is actually better than the others. (For more information concerning the meaning and the computation of the level of significance see Appendix C.)

As a final remark, let us repeat here, that the observed variability in the algorithm producing best classification rates across the subjects, probably indicate that in a real application, a separate system should be designed for each person.

Table 6.4: The subjects for which the method using class characteristic signals produces better rates than the other algorithms. The observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of class charact. signal (%)	Significance Level (%)				
		Single-trial NMF-Gauss.	Multi-trial NMF-Gauss.	Class charact. signal (feat. constr.)	Time-SVM	Time-Gauss.
5	75	50	36.05	**0.9**	24.04	24.04

Table 6.5: The subjects for which the method using KPCA features and Support Vector Classification produces better rates than the other algorithms. The observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of KPCA - SVM (%)	Significance Level (%)				
		Single-trial NMF-Gauss.	Multi-trial NMF-Gauss.	Class charact. signal	Class charact. signal (feat. constr.)	Time-Gauss.
10	75	27.72	34.56	12.35	**0.73**	34.56
11	82.81	**4.7**	10.01	14.01	25.31	40.96

Table 6.6: The subjects for which the method using KPCA features and Gaussian classification produces better rates than the other algorithms. The observed level of significance when each of the other algorithms is considered equivalent or superior.

Subj.	Classif. rate of KPCA - Gauss. (%)	Significance Level (%)				
		Single-trial NMF-Gauss.	Multi-trial NMF-Gauss.	Class charact. signal	Class charact. signal (feat. constr.)	Time-SVM
1	96.55	**0.01**	**0.17**	**2.26**	**0.34**	32.5
3	86.84	11.92	18.23	**1.4**	**2.56**	26.68
6	74	17.56	26.52	10.88	26.52	43.67
8	97.22	**1.07**	**4.3**	**0.23**	15.34	15.34
13	75.76	34.63	21.84	21.84	21.84	34.63
14	87.5	13.77	13.77	**0.66**	28.32	38.39

Chapter 7

Conclusions and future perspectives

We proposed in this thesis a number of methods aiming to discriminate between a person's quick and slow responses in an oddball experiment, using only an early part of their EEG signals, i.e. a part always preceding the person's response. In Chapter 3 we focused on using the magnitude of the spectrum of the signals as features. These features were then processed with a number of techniques such as: their Euclidean distance, PCA, LDA and NMF. In all cases a Gaussian classifier was used for the classification task. Moreover, PCA and NMF were also used for the construction of a subspace of the original feature space where the classification task could be easier. In Chapter 4 the Time-Frequency representations of the EEG signals, acquired using the Continuous Wavelet Transform, were analyzed using NMF to construct features for the desired classification. The analysis was done in a single-trial as well as in a multi-trial basis. In Chapter 5 the use of a characteristic signal for each class, constructed by combining the signals of the various channels, was proposed to perform the classification task. A channel selection algorithm was also developed for choosing the appropriate channels. Finally, in Chapter 6 the KPCA method proposed for tackling the HPM problem was applied to our problem. Taking into account the results acquired by all these approaches we can draw the following conclusions:

- There is a large variability in the classification rates both across subjects and meth-

ods used. There is no method appearing to be optimal for all the subjects. This observation supports our belief that in a real application a "tailor made" system should be designed for each user separately.

- The right selection of EEG channels can considerably improve the classification results. The channel selection algorithm we propose, indicates a subset of channels which produces better classification rates compared with the ones acquired when all channels are used. This stands for all subjects. However, the subset of channels is not identical across subjects. This again implies the need of designing a "tailor made" system for each user.

- Channels Fz, Cz, Pz, the signals of which are considered in the literature to exhibit correlations with the cognitive processes of recognising a target, are generally ranked high by our channel selection algorithm. Although in most cases there is a different subset of channels producing better results (usually containing them), these channels are a good choice if a universal set has to be used for all subjects.

- KPCA features performed quite well for most subjects. Although they were not the best in all cases, they seem to be the best choice if there is a need of a universal approach to the problem.

- Although the classification rates produced were up to 97.22%, they are usually between 70% and 80%. This is certainly encouraging if we consider the difficulty of the problem. However, unless there is a remarkable improvement in the classification performance, real applications should only consider the generation of notification signals to increase the attention of operators and not involve any critical, automatic decision making process.

A number of issues concerning the work presented here should be investigated in the future. First, the presented methods should be applied to more subjects and with more trials to be able to acquire more accurate classification rates and verify their validity. Moreover, we intend to compare the channel selection method we proposed here with other methods in the literature such as the ones presented in [48], which are based on SVM

techniques for the identification of the useful channels. Towards this direction, we are also going to investigate the possibility of analysing the data in a cross-subjects manner, in order to produce a set of channels for the classification task, being robust across subjects. Such an approach was followed in [78] using the methods of [48], producing moderate results.

Another issue that is worth looking at, is the application of the proposed methods in the field of Human Performance Monitoring. As explained in section 6.1 this area is closely related to the one of our problem but due to the lack of the need of prediction the problem is easier. This is because we are able to use the preprocessing technique of averaging as well as take advantage of components in the EEG signals appearing after the person's reaction to the stimulus. These two facts are likely to enable our methods to capture more characteristic features correlated with the subjects' performance so we expect that the classification rates should be higher. The results should then be compared with the ones presented in the HPM literature.

Finally, methods from the field of Neuroimaging, such as functional Magnetic Resonance Imaging (fMRI), are interesting to be investigated whether they can be applied to the performance prediction problem. There is the possibility that such methods could produce good results as they provide much better insight to the human brain's states due to their excellent spatial resolution. On the other hand, problems arising from the poor time resolution of fMRI should be addressed. Moreover, fMRI is quite expensive for the time being and of course could not be applied in a real application due to the "heavy" equipment needed to be acquired. However, fMRI has been used in the literature to study human cognition, so its use to tackle the prediction problem is certainly of academic interest.

Bibliography

[1] P. Abry. *Ondelettes et turbulence. Multirsolutions, algorithmes de dcomposition, invariance d'chelles.* Diderot Editeur, Paris, 1997.

[2] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[3] G. Arfken. *Mathematical methods for physicists.* Academic Press, 4th edition, 1995.

[4] J. I. Aunon and C. D. McGillem. Techniques for processing single evoked potentials. *Proc. San Diego Biomed. Symp.*, pages 211–218, 1975.

[5] J. I. Aunon, C. D. McGillem, and D. G. Childers. Signal processing in evoked potential research: Averaging and modeling. *CRC Crit. Rev. Bioeng.*, 5:323–367, 1981.

[6] E. A. Bartnik, K. J. Blinowska, and P. J. Durka. Single evoked potential reconstruction by means of wavelet transform. *Biological Cybernetics*, 67:175–181, 1992.

[7] E. Basar, C. Basar-Eroglu, S. Karakas, and M. Schürmann. Oscillatory brain theory: A new trend in Neuroscience. *IEEE Engineering in Medicine and Biology*, 18(3):56–66, May/June 1999.

[8] M. F. Bear, B. W. Connors, and M. A. Paradiso. *Neuroscience: exploring the brain.* Lippincott Williams and Wilkins, Philadelphia, 2nd edition, 2001.

[9] M. M. Beigi, A. Erfanian, and M. Elkhani. Multiresolution adaptive filter for estimating brainstem auditory evoked potential. *Proc. of the 20th Annual International Conference of the IEEE Eng. in Medicine and Biology Society*, 20(3):1474–1477, 1998.

[10] E.M. Bernat, S.M. Malone, W.J. Williams, C.J. Patrick, and W.G. Iacono. Decomposing delta, theta, and alpha time-frequency ERP activity from a visual oddball task using PCA. *International Journal of Psychophysiology*, 64:62–74, 2007.

[11] O. Bertrand, J. Bohorquez, and J. Pernier. Time-frequency digital filtering based on an invertible wavelet transform: an application to evoked potentials. *IEEE Transactions on Biomedical Engineering*, 41:77–88, 1994.

[12] G. Box. *Time series analysis, forecasting and control.* Holden-Day, San Francisco, 1976.

[13] M. A. B. Brazier. Evoked responses recorded from the depths of the human brain. *Ann. N.Y. Acad. Sci*, 112:33–59, 1964.

[14] S.L. Bressler. Event Related Potentials. In *The handbook of brain theory and neural networks*, pages 412–415. The MIT Press, 2003.

[15] J.T. Cacioppo, L.G. Tassinary, and G.G. Berntson. *Handbook of Psychophysiology.* Cambridge University Press, Second Edition, 2000.

[16] S. Cerutti, G. Baselli, D. Liberati, and G. Pavesi. Single sweep analysis of visual evoked potentials through a model of parametric identification. *Biological Cybernetics*, 56:111–120, 1987.

[17] S. Cerutti, V. Bersani, A. Carrara, and D. Liberati. Analysis of visual evoked potentials through wiener filtering applied to a small number of sweeps. *Journal of Biomedical Engineering*, 9:3–12, 1987.

[18] S. Cerutti, G. Chiarenza, P. Mascellani D. Liberati, and G Pavesi. A parametric method of identification of single-trial event-related potentials in the brain. *IEEE Transactions on Biomedical Engineering*, 35(9):701–711, 1988.

[19] F. H. Y. Chan, F. K. Lam, P. W. F. Poon, and W. Qiu. Detection of brainstem auditory evoked potential by adaptive filtering. *Medical and Biological Engineering and Computing*, 33:69–75, 1995.

[20] J. O. Chapa and M. R. Raghuveer. Optimal matched wavelet construction and its application to image pattern recognition. In *Wavelet applications: II. Proceedings of the Society for Photo Instrumentation Engineering*, volume 2491, pages 518–529. H. H. Szu (Ed.), 1995.

[21] C. Chatfield. *Statistics for Technology, A course in applied statistics (third edition).* Chapman and Hall, New York, third edition, 1983.

[22] I. Daubechies. *Ten lectures on wavelets.* Capital city press, Montpelier, Vermont, 1992.

[23] C. E. Davila, A. J. Welch, and H. G. Rylander. Eigenvector decomposition of single-trial evoked potentials. *Proc. 9th Int. Conf. IEEE Eng. Med. Biol. Society*, pages 602–603, Boston, MA, 1987.

[24] J. P. C. deWeerd. A posteriori time-varying filtering of averaged evoked potentials - i. introduction and conceptual basis. *Biological cybernetics*, 41:211–222, 1981.

[25] K.I. Diamantaras and S.Y. Kung. *Principal Components Neural Networks: Theory and Applications.* Wiley, 1996.

[26] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis.* New York: Wiley, 1973.

[27] P. Eykhoff. *System identification - parameter and state estimation.* Wiley, New York, 1971.

[28] L. A. Farwell, J. M. Martinerie, T. R. Bashore, P. E. Rapp, and P. H. Goddard. Optimal digital filters for long-latency components of the event-related brain potential. *Psychophysiology*, 30:306–315, 1993.

[29] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, San Diego, California, 1990.

[30] M.S. Gazzaniga. *Cognitive Neuroscience: A Reader.* Blackwell Publishing, 2000.

[31] R. R. Gharieb and A. Cichocki. Noise reduction in brain evoked potentials based on third-order correlations. *IEEE Transactions on Biomedical Engineering*, 48:501–512, 2001.

[32] R. Grieve, P. A. Parker, B. Hudgins, and K. Englehart. Nonlinear adaptive filtering of stimulus artifact. *IEEE Transactions on Biomedical Engineering*, 47:389–395, 2000.

[33] C. W. Groetsch. *Inverse Problems in the Mathematical Sciences.* Vieweg, 1993.

[34] Grossman and Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.*, 15:723–736, 1984.

[35] D. Guillamet, B. Schiele, and J. Vitrià. Analyzing non-negative matrix factorization for image classification. *16th International Conference on Pattern Recognition. Proceedings IEEE.*, 2:116–119, 2002.

[36] D. Guillamet and J. Vitrià. Discriminant basis for object classification. *11th International Conference on Image Analysis and Processing. Proceedings IEEE.*, pages 256–261, 2001.

[37] M. Hanke and P. C. Hansen. Regularization methods for large-scale problems. *Surveys on Mathematics for Industry*, 3:253–315, 1993.

[38] S. Haykin. *Adaptive filter theory.* 2nd edition, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[39] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley and Sons, Inc, 2001.

[40] A. Imai and K. Tsuji. Event-related potential correlates of judgment categories and detection sensitivity in a visual detection task. *Vision Research 44*, pages 763–773, 2004.

[41] S. Karakas, U. Erzengin, and E. Basar. The genesis of human event-related responses explained through the theory of oscillatory neural assemblies. *Neuroscience Letters*, 285:45–48, 2000.

[42] P. A. Karjalainen. Regularization and bayesian methods for evoked potential estimation. Ph.D. thesis, University of Kuopio, Department of Applied Physics, Kuopio, Kuopio, Finland, 1997.

[43] P. A. Karjalainen, J. P. Kaipio, and A. S. Koistinen. Pca based bayesian estimation of single trial evoked potentials. *Proc 1st International Conference Biolectromagnetism*, pages 195–196, Tampere, Finland, 1996.

[44] P. A. Karjalainen, J. P. Kaipio, A. S. Koistinen, and M. Vauhkonen. Subspace regularization method for the single-trial estimation of evoked potentials. *IEEE Transactions on Biomedical Engineering*, 46(7):849–860, 1999.

[45] M. Koska, R. Rosipal, A. König, and L. J. Trejo. Estimation of human signal detection performance from erps using feed-forward network model. In *Computer Intensive Methods in Control and Signal Processing, The Curse of Dimensionality*, Birkhauser, Boston, 1997.

[46] S. Krieger, J. Timmer, S. Lis, and H. M. Olbrich. Some considerations on estimating event-related brain signals. *Journal of Neural Transmission*, 99:103–129, 1995.

[47] P. Laguna, J. Raimon, O. Meste, P. W. Pooh, P. Caminal, H. Rix, and N. V. Thakor. Adaptive filter for event-related bioelectric signals using an impulse correlated reference input: Comparison with signal averaging techniques. *IEEE Transactions on Biomedical Engineering*, 39:1032–1044, 1992.

[48] T.N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf. Support vector channel selection in BCI. *IEEE Transactions on Biomedical Engineering*, 51(6):1003–1010, June 2004.

[49] D. H. Lange and G. F. Inbar. A robust parametric estimator for single-trial movement related brain potentials. *IEEE Transactions on Biomedical Engineering*, 43(4):341–347, 1996.

[50] D. H. Lange, H. Pratt, and G. F. Inbar. Segmented matched filtering of single event related potentials. *IEEE Transactions on Biomedical Engineering*, 42(3):317–321, March, 1995.

[51] D.H. Lange, H. Pratt, and G.F. Inbar. Segmented matched filtering of single event related evoked potentials. *IEEE Transactions on Biomedical Engineering*, 42(3):317–321, March 1995.

[52] D.D. Lee and H.S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 180:788–791, 1999.

[53] D.D. Lee and H.S. Seung. Algorithms for non negative matrix factorization. *Advances in Neural Information Processing Systems 13. MIT Press, Cambridge, MA*, pages 556–562, 2001.

[54] H. Lee and S. Choi. PCA-based linear dynamical systems for multichannel EEG classification. *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02)*, 2:745–749, 2002.

[55] H. Lee and S. Choi. PCA + HMM + SVM for EEG pattern classification. *Seventh International Symposium on Signal Processing and Its Applications. Proceedings IEEE.*, 1:541–544, 2003.

[56] B.-S. Lin, B.-S. Lin, F.-C. Chong, and F. Lai. Adaptive filtering of evoked potentials using higher-order adaptive signal enhancer with genetic-type variable step-size prefilter. *Medical and Biological Engineering and Computing*, 43:638–647, 2005.

[57] W. Liu, N. Zheung, and X. Li. Nonnegative matrix factorization for EEG signal classification. *International Symposium on Neural Networks. Proceedings Springer.*, 2:470–475, 2004.

[58] K. Lugger, D. Flotzinger, A. Schlogl, M. Pregenzer, and G. Pfurtscheller. Feature extraction for on-line EEG classification using principal components and linear discriminants. *Medical and Biological Engineering and Computing*, 36(3):309–314, May 1998.

[59] P. J. Maccabee, E. I. Pinkhasov, and R. Q. Cracco. Short latency somatosensory evoked potentials to median nerve stimulation: effect of low frequency filter. *Electroencephalogr. Clin. Neurophysiol.*, 55(1):34–44, 1983.

[60] P. G. Madhavan. Minimal repetition evoked potentials by modified adaptive line enhancement. *IEEE Transactions on Biomedical Engineering*, 39:760–764, 1992.

[61] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(7):674–693, 1989.

[62] S. Mallat. *A wavelet tour of signal processing.* Elsevier, Second Edition, 1999.

[63] S. Nishida, M. Nakamura, and H. Shibasaki. Method for single-trial recording of somatosensory evoked potentials. *Journal of Biomedical Engineering*, 15(3):257–262, May, 1993.

[64] P. Nunez. *Neocortical Dynamics and Human EEG Rhythms.* New York OUP, New York, 1995.

[65] M. Nuwer, G. Comi, F. Emerson, A. Fuglsang, J. Guerit, H. Hinrichs, A. Ikeda, F. Luccas, and P. Rapperlsburger. IFCN standards for digital recording of clinical EEG. *Electroencephalography and clinical Neurophysiology*, 106:259–261, Jan 1998.

[66] V. Parsa and P. A. Parker. Multireference adaptive noise cancellation applied to somatosensory evoked potentials. *IEEE Transactions on Biomedical Engineering*, 41:792–800, 1994.

[67] S.H. Patel and P.N. Azzam. Characterization of N200 and P300: Selected studies of the event-related potential. *International Journal of Medical Sciences*, 2(4):147–154, 2005.

[68] R. Quian Quiroga. Obtaining single stimulus evoked potentials with wavelet denoising. *Physica D*, 145:278–292, 2000.

[69] R. Quian Quiroga and H. Garcia. Single-trial event-related potentials with wavelet denoising. *Clinical Neurophysiology*, 114:376–390, 2003.

[70] R. Rosipal, M. Girolami, and L.J. Trejo. Kernel PCA feature extraction of Event-Related Potentials for human signal detection performance. *Artificial Neural Networks in Medicine and Biology, (Proc. of the ANNIMAB-1 Conference), Springer*, pages 321–326, 2000.

[71] R. Rosipal, M. Girolami, L.J. Trejo, and A. Cichocki. Kernel PCA for feature extraction and de-noising in nonlinear regression. *Neural Computing and Applications*, 10(3):231–243, 2001.

[72] P. M. Rossini, R. Q. Cracco, J. B. Cracco, and W. J. House. Short latency somatosensory evoked potentials to peroneal nerve stimulation: scalp topography and the effect of different frequency filters. *Electroencephalogr. Clin. Neurophysiol.*, 52(6):540–552, 1981.

[73] D. S. Ruchkin and E. M. Glaser. Simple digital filters for examining cnv and p300 on a single trial basis. In *Multidisciplinary Perspectives on Event-Related Brain Potential Research*, pages 579–581. D.A. Otto, Ed. Washington, DC: US Government Printing Office, 1978.

[74] M. Rugg and M. Coles. *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition.* Oxford Psychology Series. New York, 1997.

[75] V. J. Samar, H. Begleiter, J. O. Chapa, M. R. Raghuveer, M. Orlando, and D. Chorlian. Matched meyer neural wavelets for clinical and experimental analysis of auditory and visual evoked potentials. In *Signal processing: VIII. Theories and applications, Proceedings of EUSIPCO-96*, pages 387–390. G. Ramponi, G. I. Sicuranza, S. Carrato, S. Marsi (Eds.). Trieste: Edizioni LINT., 1996.

[76] M. Samonas, M. Petrou, and A.A. Ioannides. Identification and elimination of cardiac contribution in single-trial magnetoencephalographic signals. *IEEE Transactions on Biomedical Engineering*, 44(5), 1997.

[77] B. Schölkopf, A. J. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[78] M. Schröder, T.N. Lal, T. Hinterberger, M. Bogdan, N. Jeremy Hill, N. Birbaumer, W. Rosenstiel, and Bernhard Schölkopf. Robust eeg channel selection across subjects for brain-computer interfaces. *2005 Hindawi Publishing Corporation*, 19:31033112.

[79] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003.

[80] K.M. Spencer and J. Polich. Poststimulus EEG spectral analysis and P300: Attention, task, and probability. *Psychophysiology*, 36:220–232, 1999.

[81] G. H. Steeger, O. Herrmann, and M. Spreng. Some improvements in the measurements of variable latency acoustically evoked potentials in human eeg. *IEEE Transactions on Biomedical Engineering*, BME-30:295–303, 1983.

[82] G.W. Stewart. *Introduction to Matrix Computations.* Academic Press, Inc., 1973.

[83] N. V. Thakor. Adaptive filtering of evoked potentials. *IEEE Transactions on Biomedical Engineering*, 34:6–12, 1987.

[84] R. F. Thompson. *The brain: a neuroscience primer.* Worth Publishers, New York, 3rd edition, 2000.

[85] L. J. Trejo, A. F. Kramer, and J. A. Arnold. Event-related potentials as indices of display-monitoring performance. *Biological Psychology*, 40:33–71, 1995.

[86] L. J. Trejo and M. J. Shensa. Feature extraction of event-related potentials using wavelets: An application to human performance monitoring. *Brain and Language*, 66:89–107, 1999.

[87] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, N.Y., 1995.

[88] C. A. Vaz and N. V. Thakor. Adaptive fourier estimation of time-varying evoked potentials. *IEEE Transactions on Biomedical Engineering*, 36:448–455, 1989.

[89] D. O. Walter. A posteriori Wiener filtering of average evoked responses. In *Advances in EEG analysis*, pages 61–70. D. O. Walter and M. A. Brazier, Eds., Electroenceph. Clin. Neurophysiol., Suppl. 27, 1969.

[90] B. Wang and M.D. Plumbley. Musical audio stream separation by non negative matrix factorization. *Proc. UK Digital Music Research Network (DMRN) Summer Conf.*, 2005.

[91] J. J. Westerkamp and J. I. Aunon. Optimum multielectrode a posteriori estimates of single-response evoked potentials. *IEEE Transactions on Biomedical Engineering*, BME-34:13–22, 1987.

[92] E.B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212, 1927.

[93] K.-B. Yu and C. D. McGillem. Optimum filters for estimating evoked potential waveforms. *IEEE Transactions on Biomedical Engineering*, BME-30:730–737, 1983.

[94] X-H Yu, Z-Y He, and Y-S Zhang. Time-varying adaptive filters for evoked potential estimation. *IEEE Transactions on Biomedical Engineering*, 41:1062–1071, 1994.

[95] X. H. Yu, Z. Y. He, and Y. S. Zhang. Time-varying adaptive filters for evoked potential estimation. *IEEE Transactions on Biomedical Engineering*, 41:1062–1071, 1994.

Appendix A

Averages of EEG signals

We present here (Figures A.1-A.10) the averages of the EEG signals, in all 18 channels, of subjects S2, S3, S4, S5, S6, S8, S10, S11, S13, S14, for classes "success" and "failure". The mean value of each signal has been subtracted from its samples.

The averages of the magnitude of the spectrum over the EEG signals of valid trials for the two classes are presented in Figures A.11-A.20

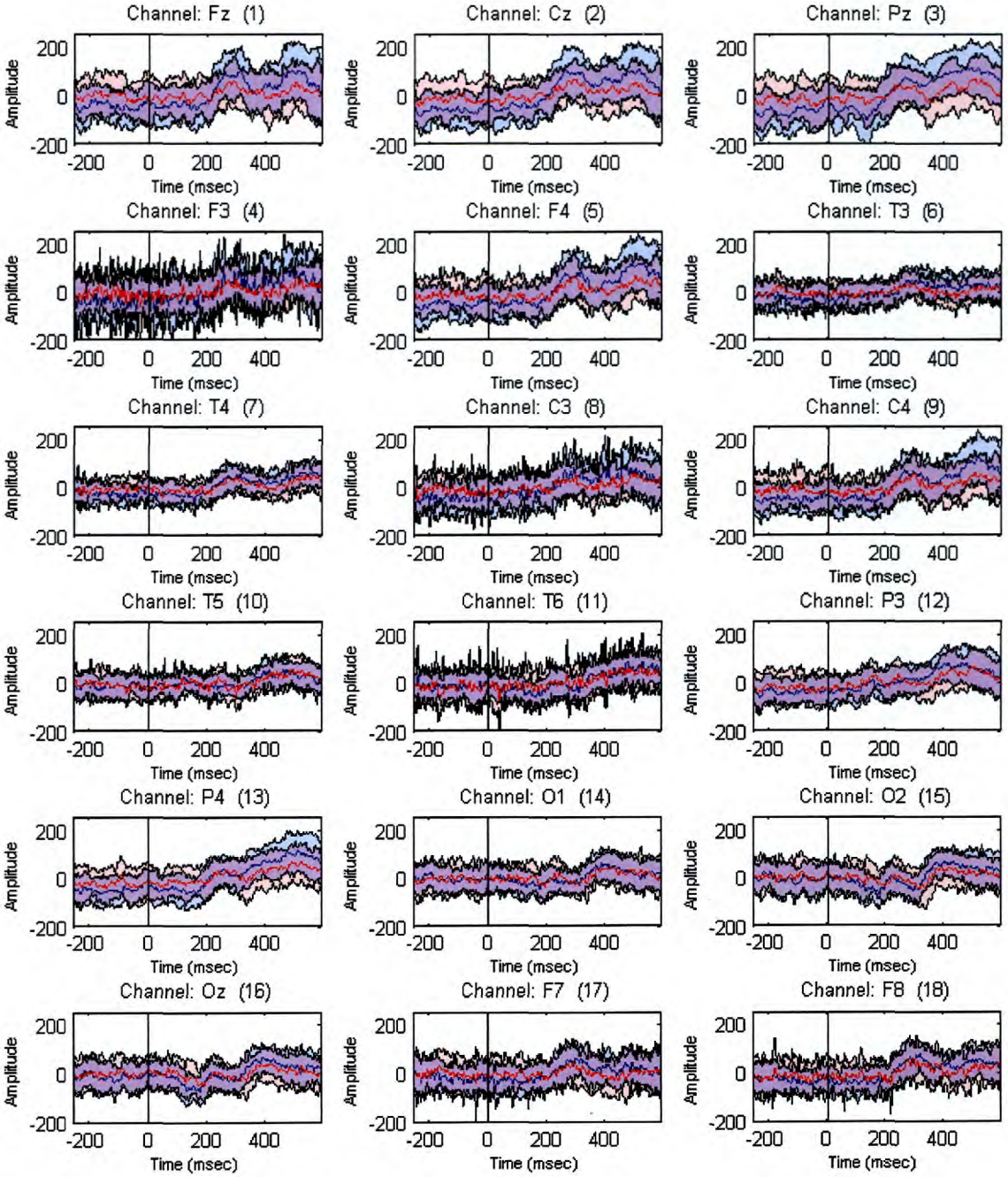


Figure A.1: Subject 2. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

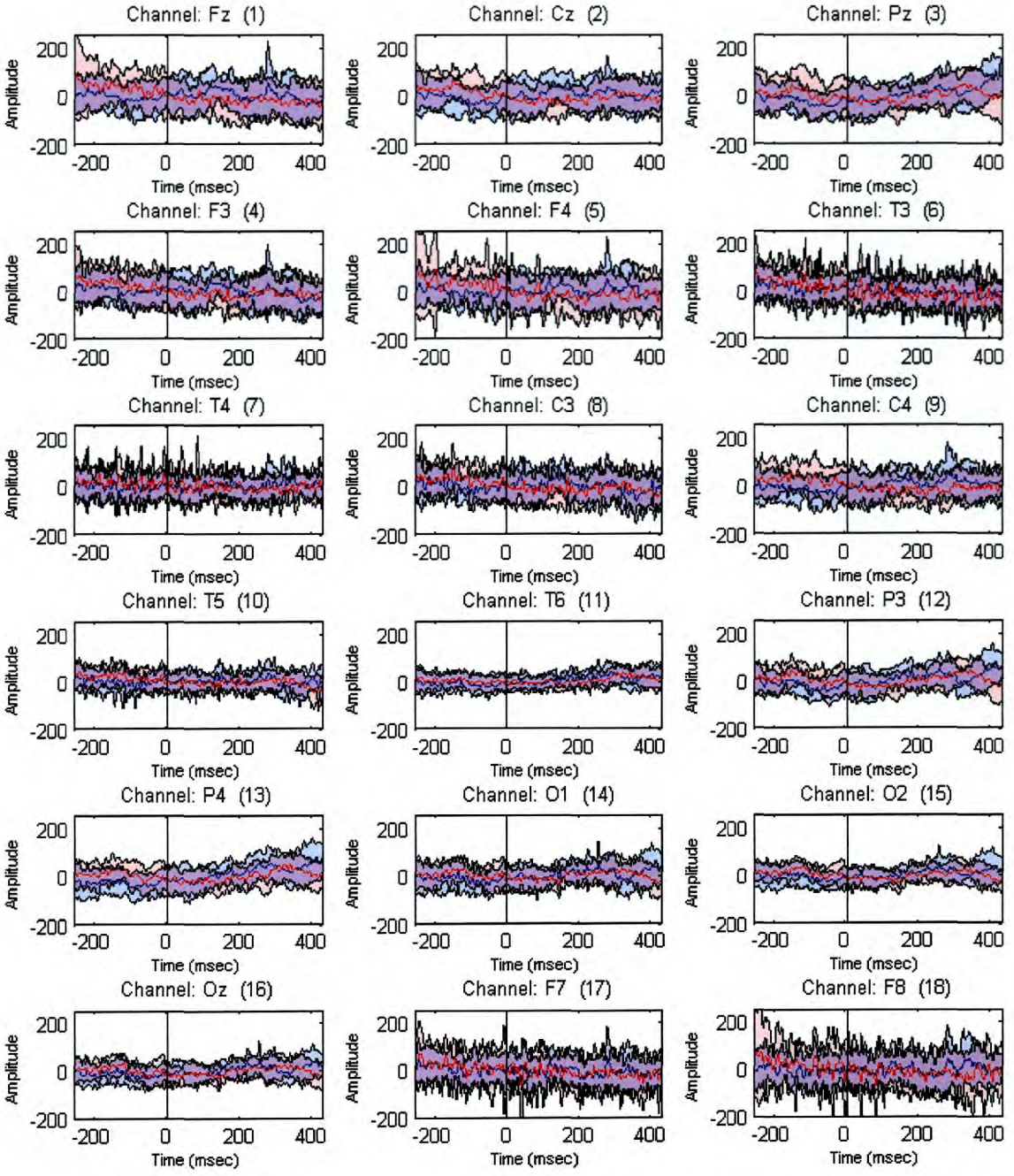


Figure A.2: Subject 3. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

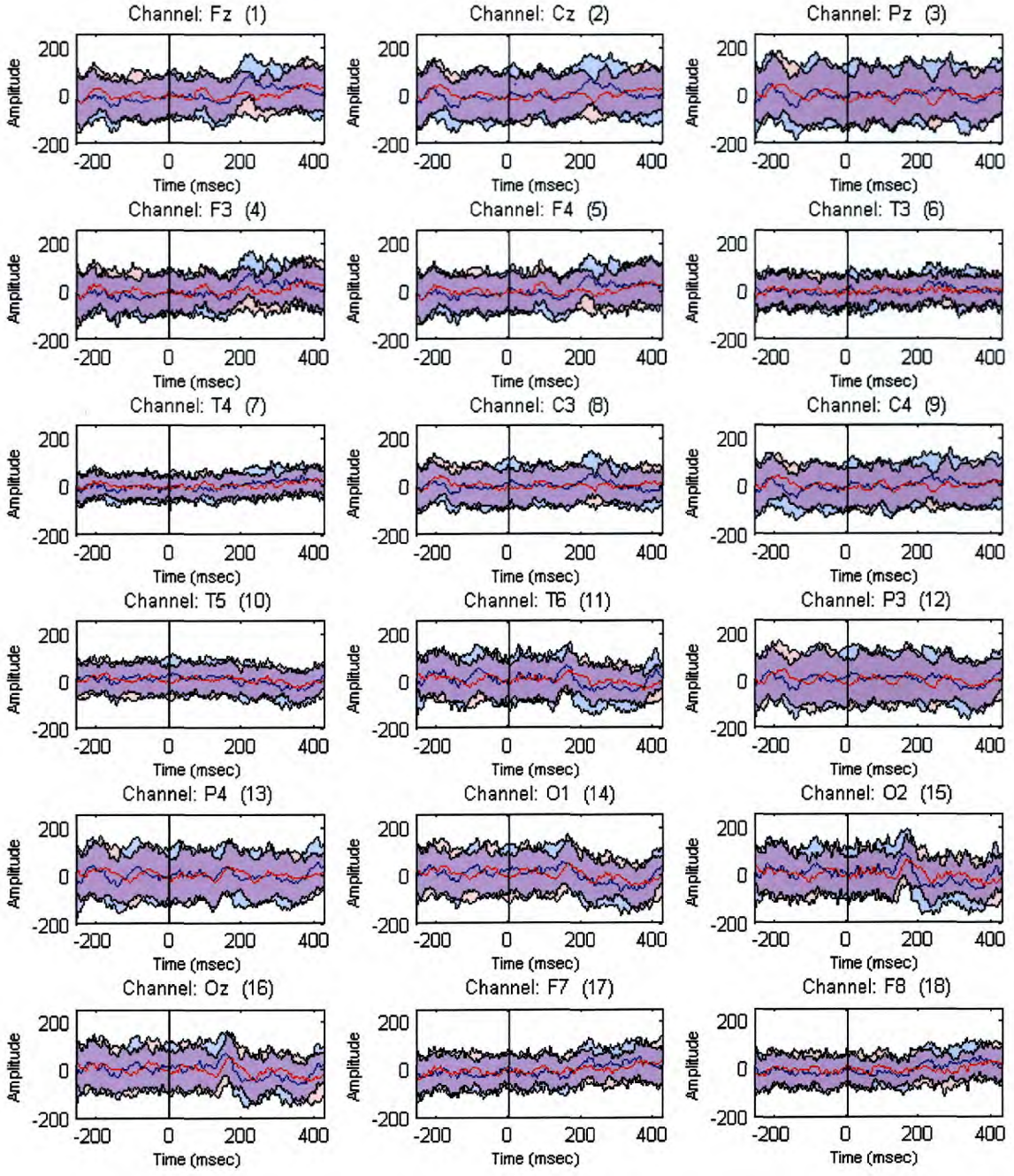


Figure A.3: Subject 4. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

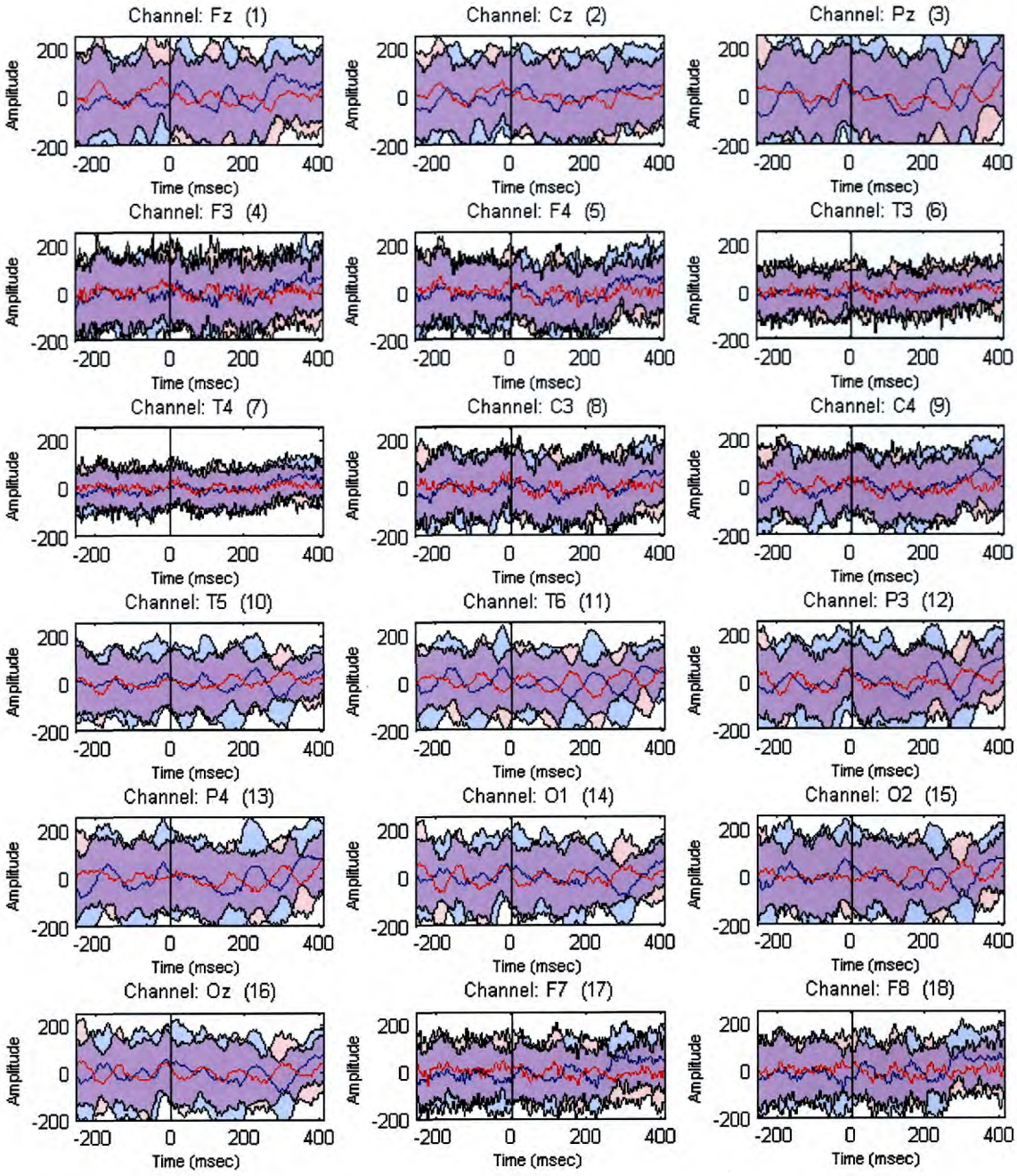


Figure A.4: Subject 5. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

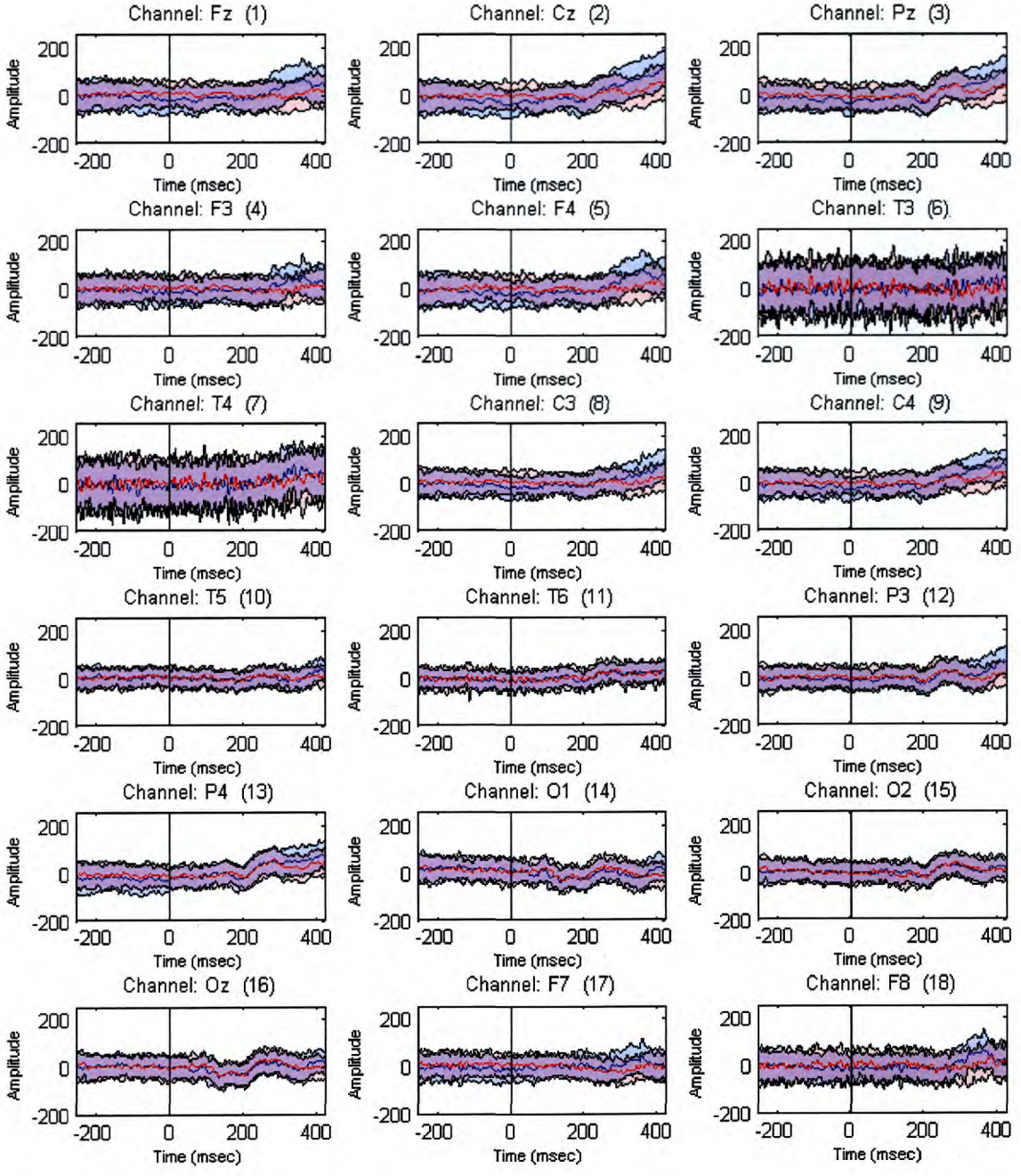


Figure A.5: Subject 6. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

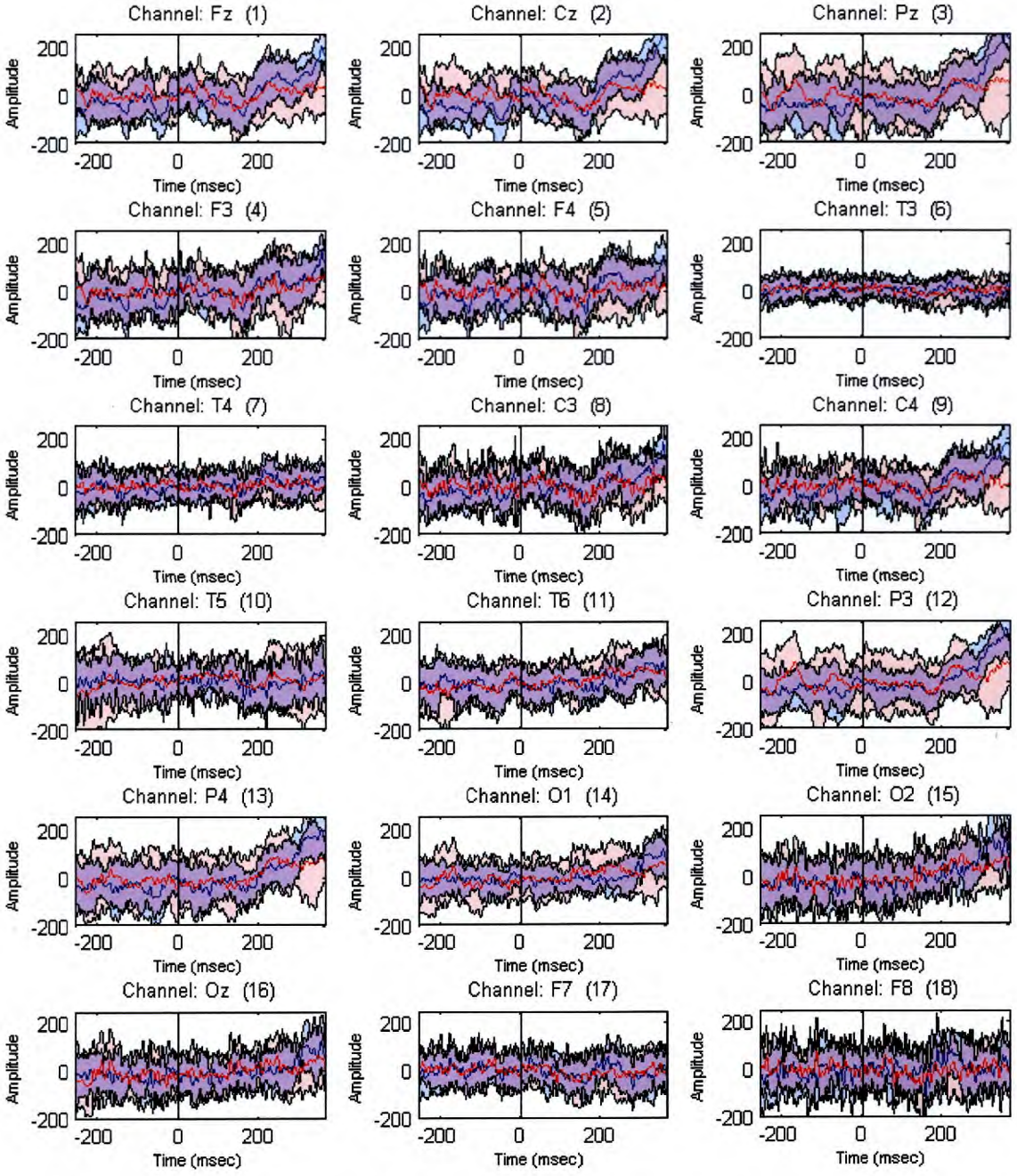


Figure A.6: Subject 8. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

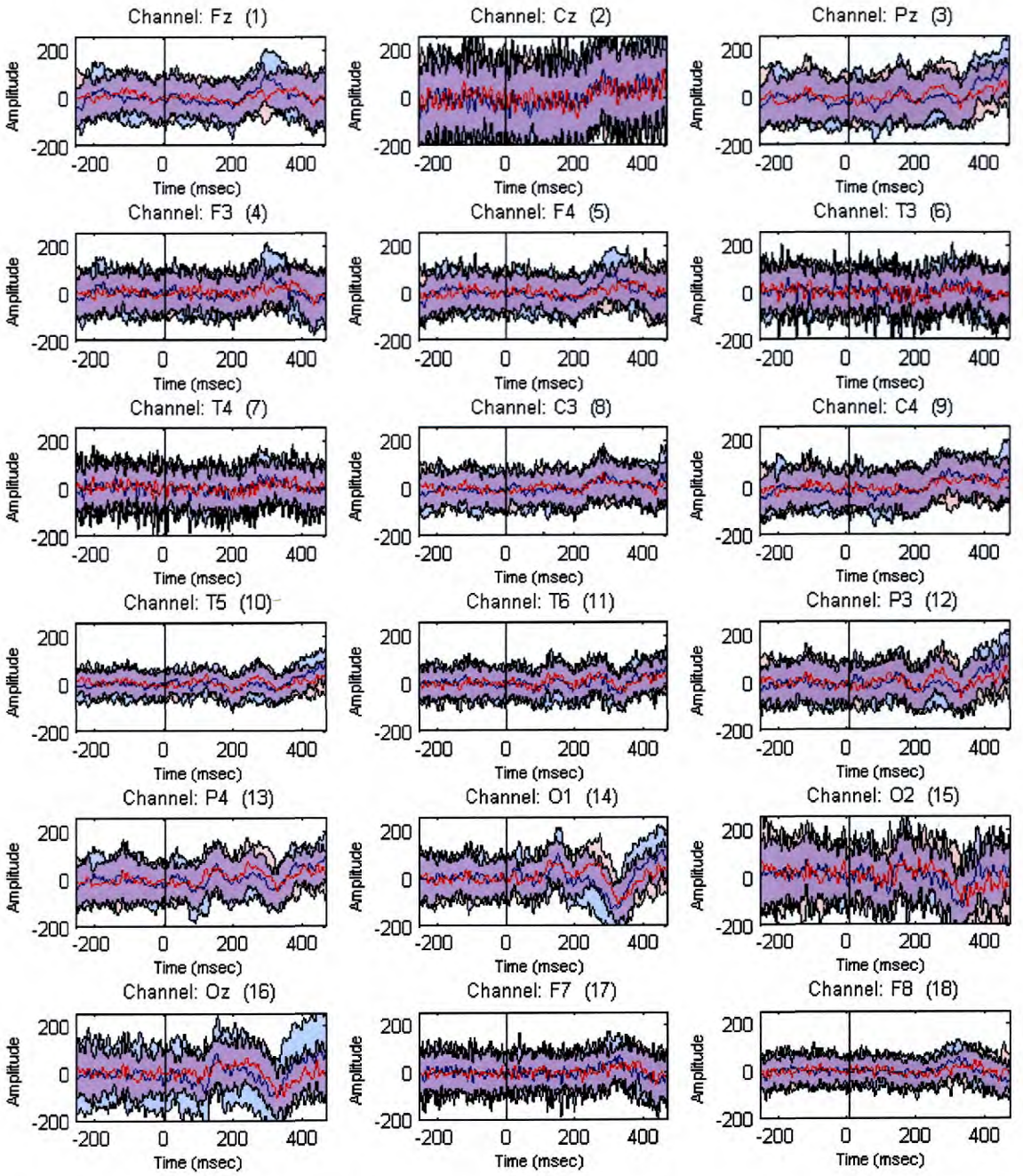


Figure A.7: Subject 10. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

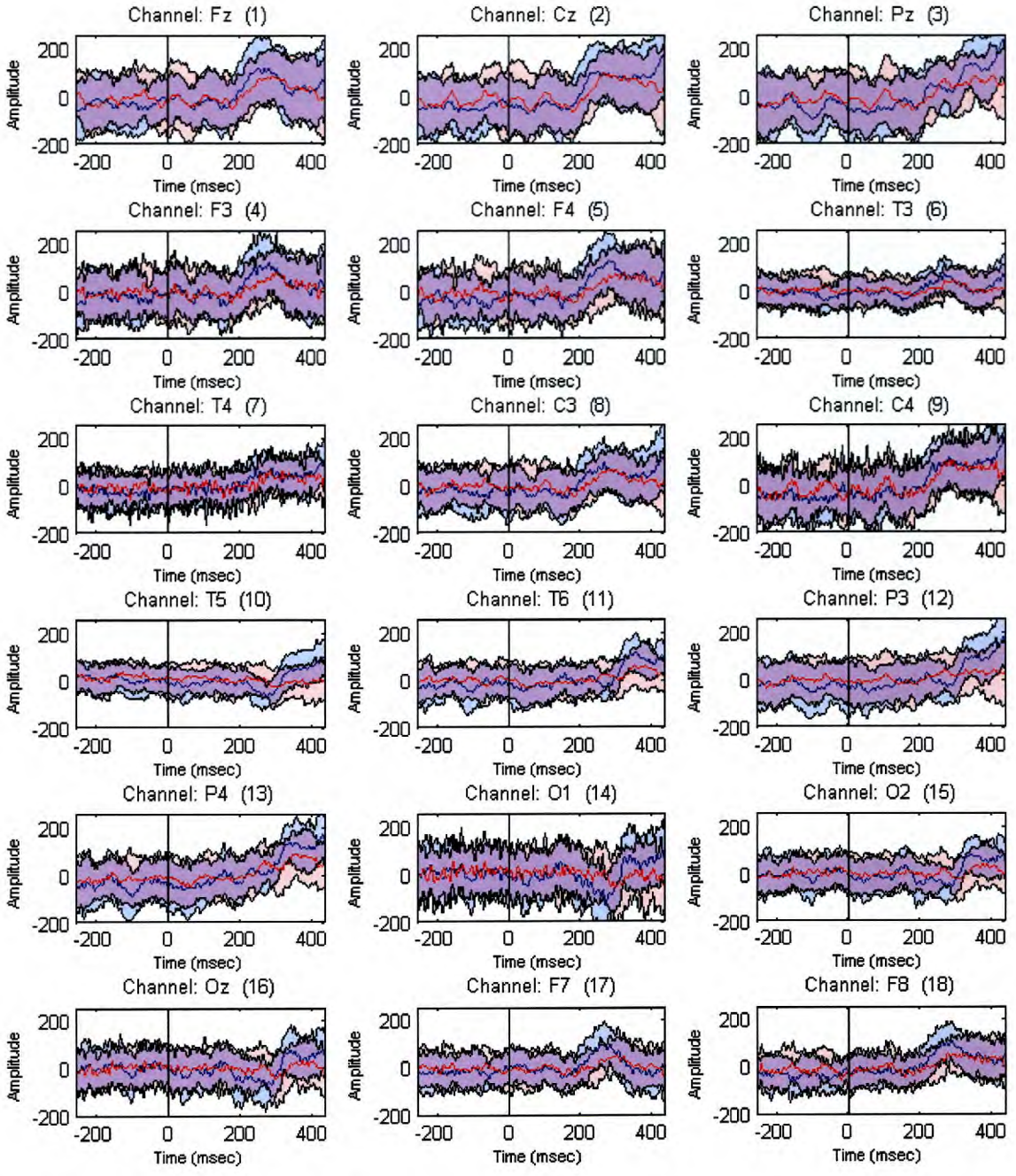


Figure A.8: Subject 11. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

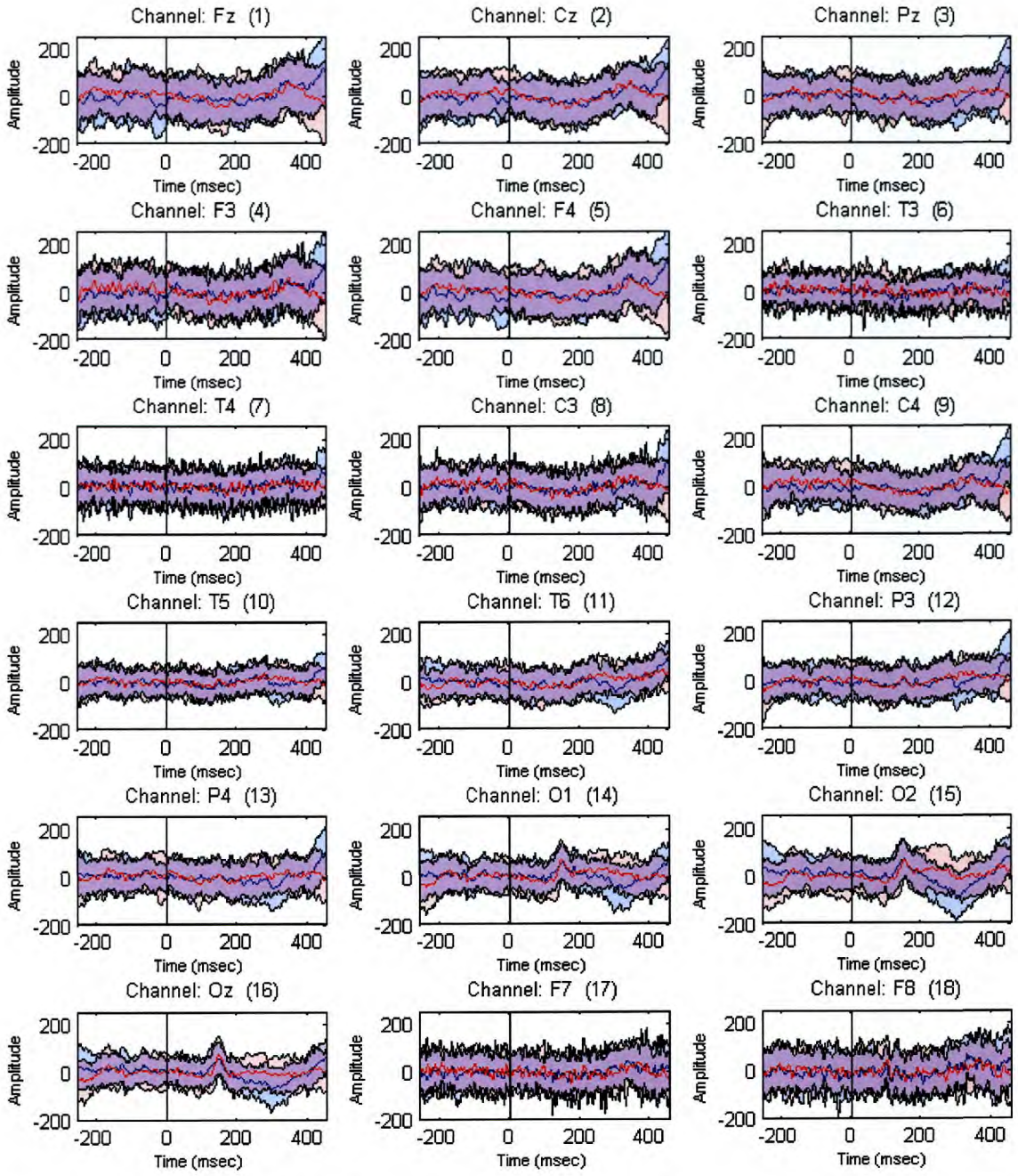


Figure A.9: Subject 13. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

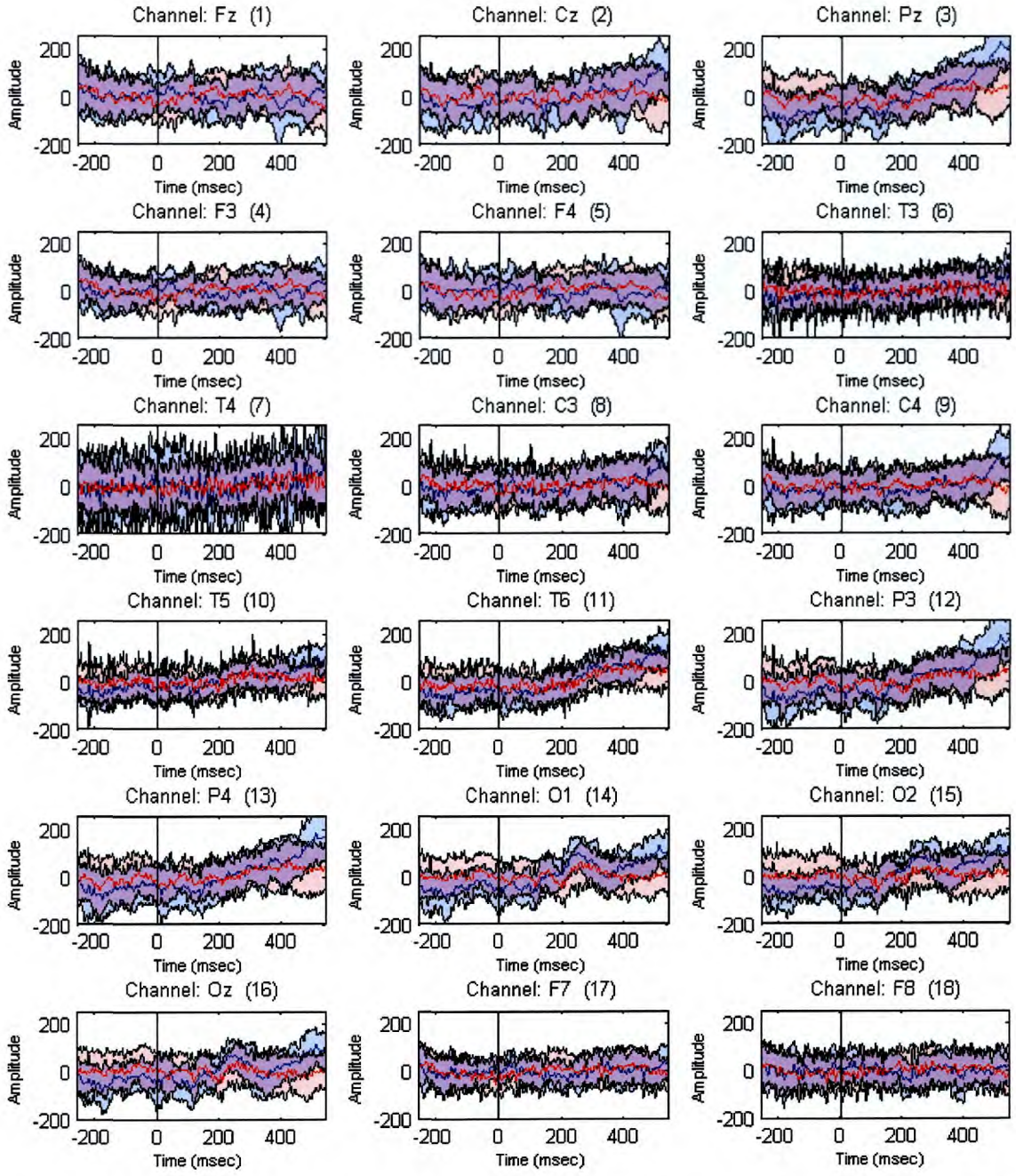


Figure A.10: Subject 14. The blue and red lines are the average signals over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The vertical line indicates the time of stimulus onset. The signals are truncated at the time point of the subject's quickest reaction.

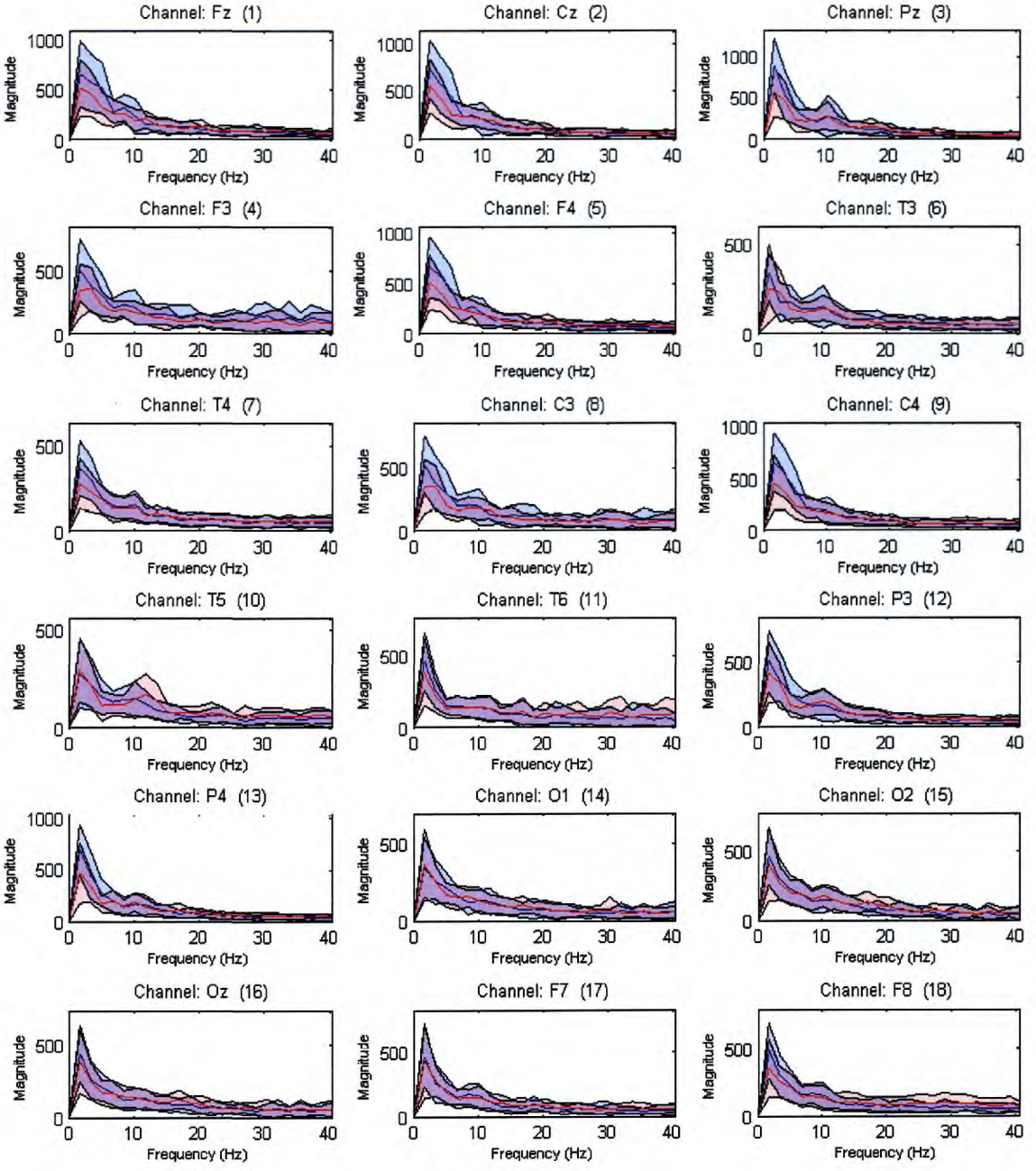


Figure A.11: Subject 2. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

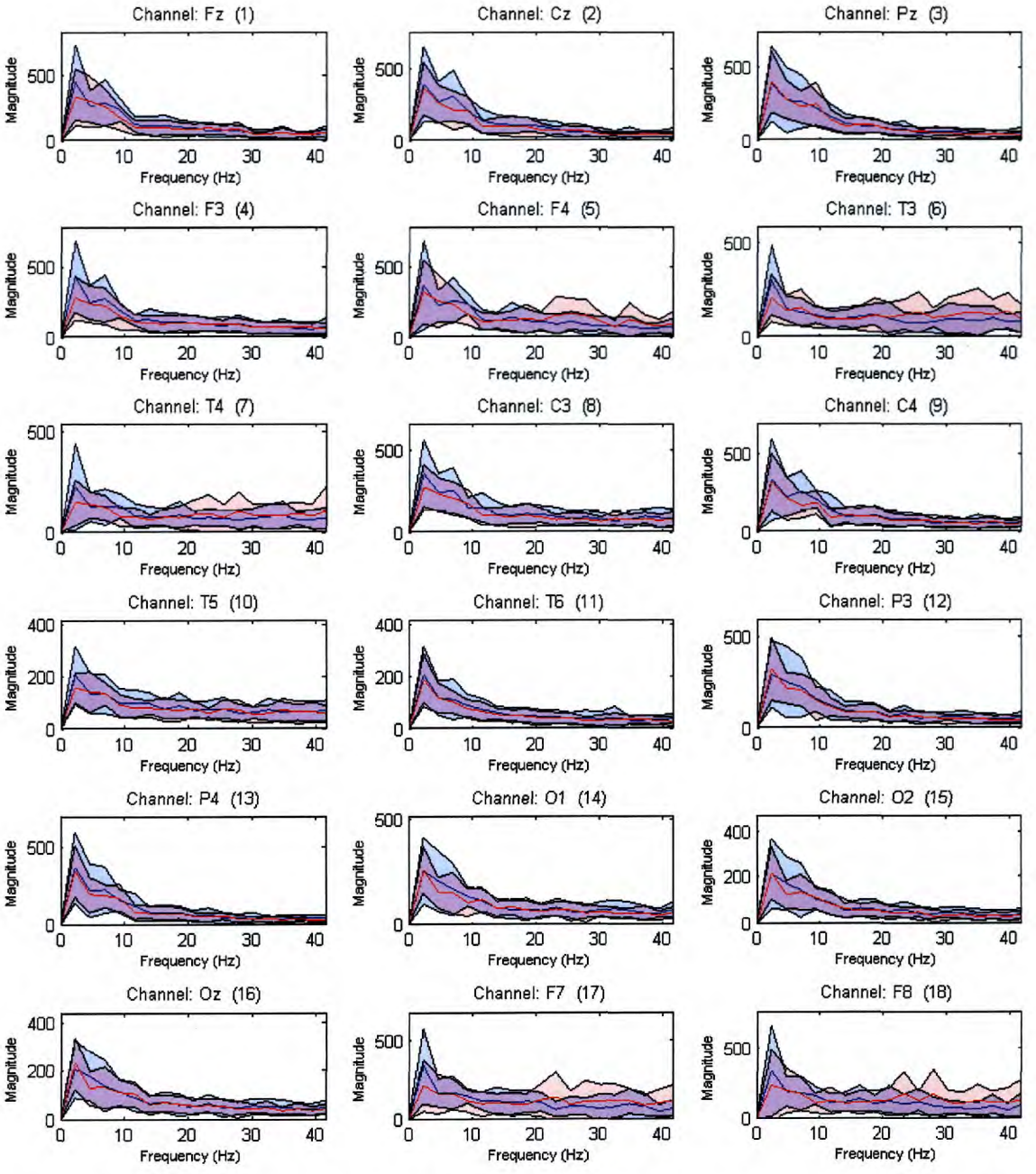


Figure A.12: Subject 3. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

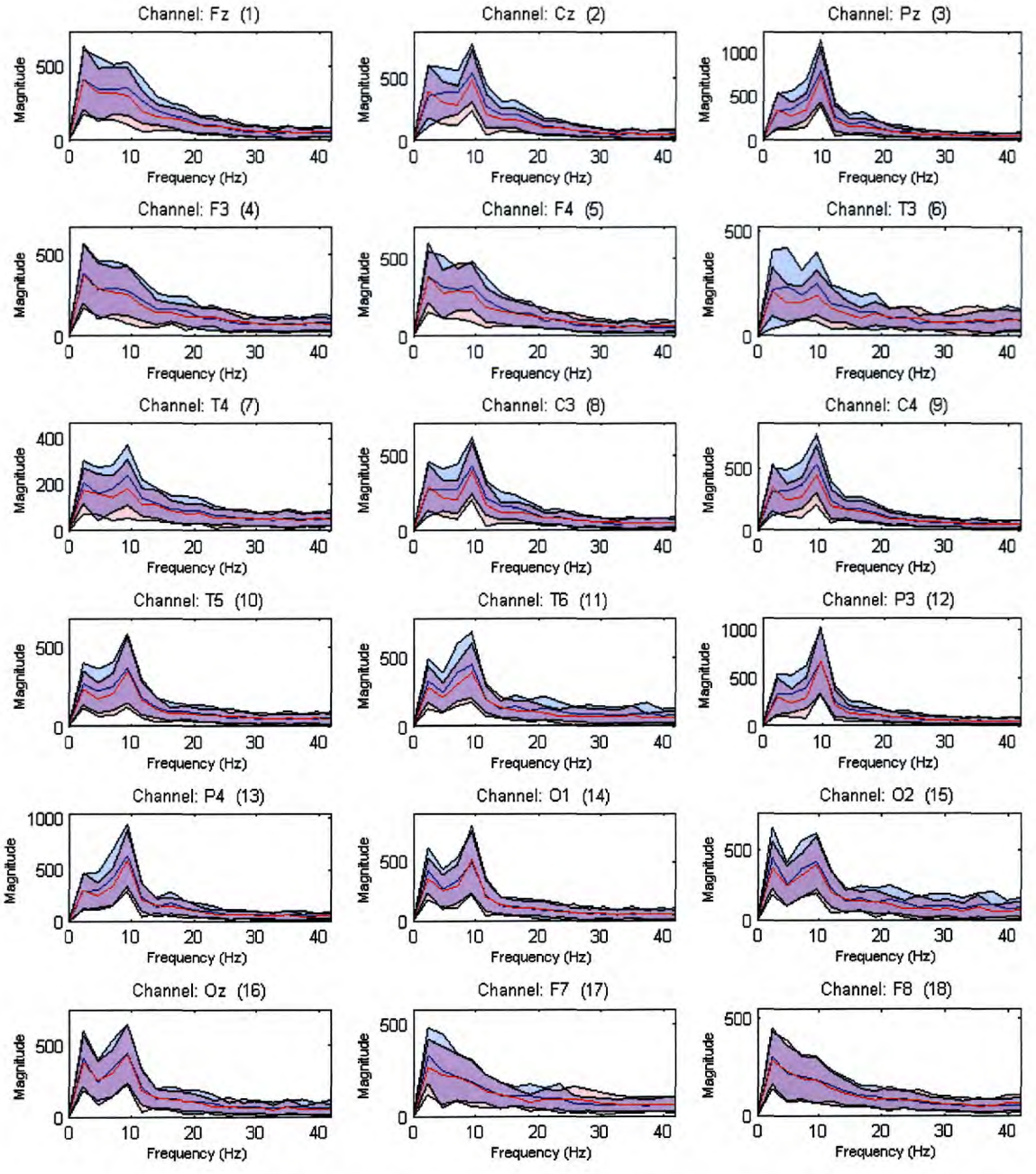


Figure A.13: Subject 4. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

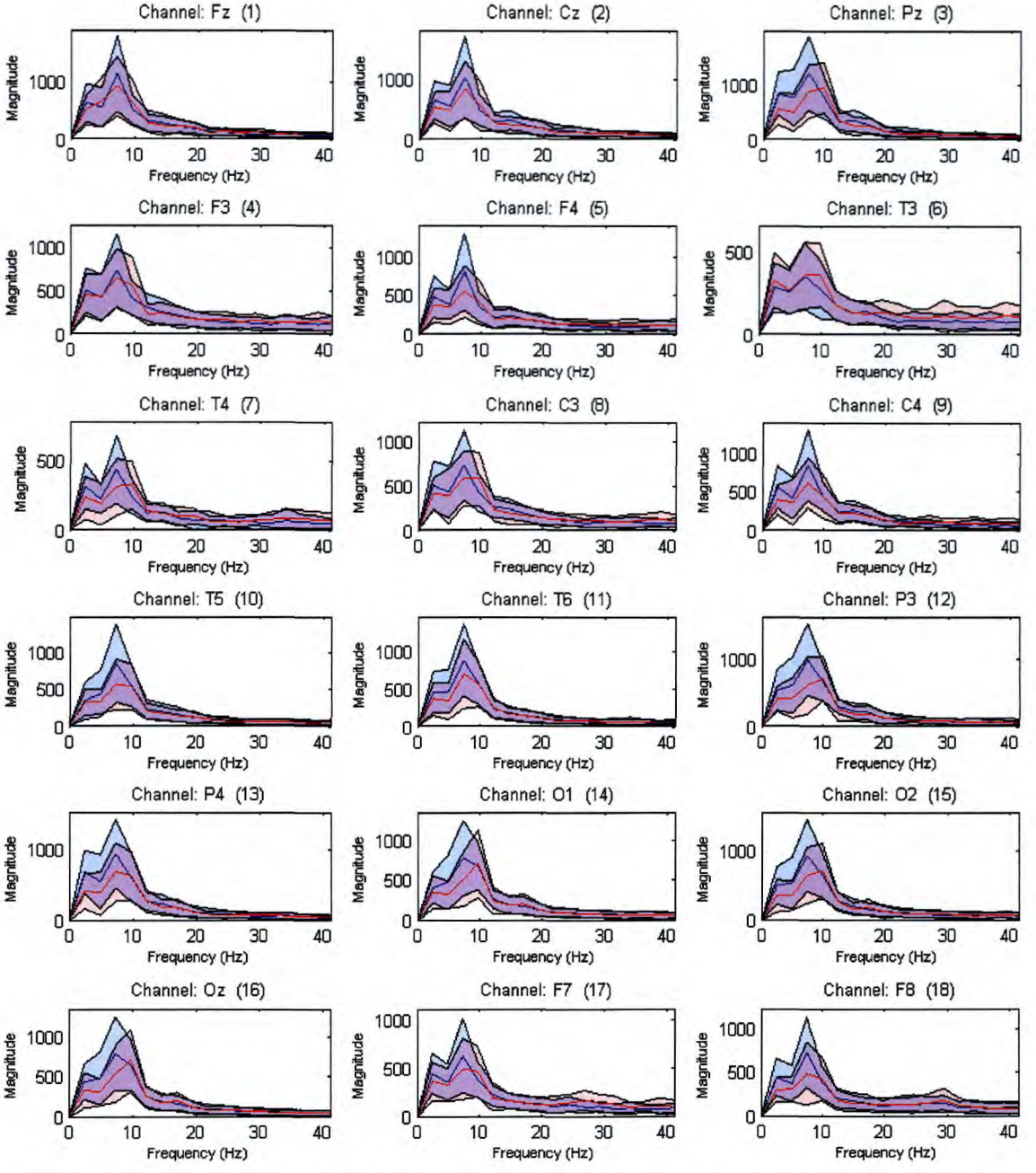


Figure A.14: Subject 5. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

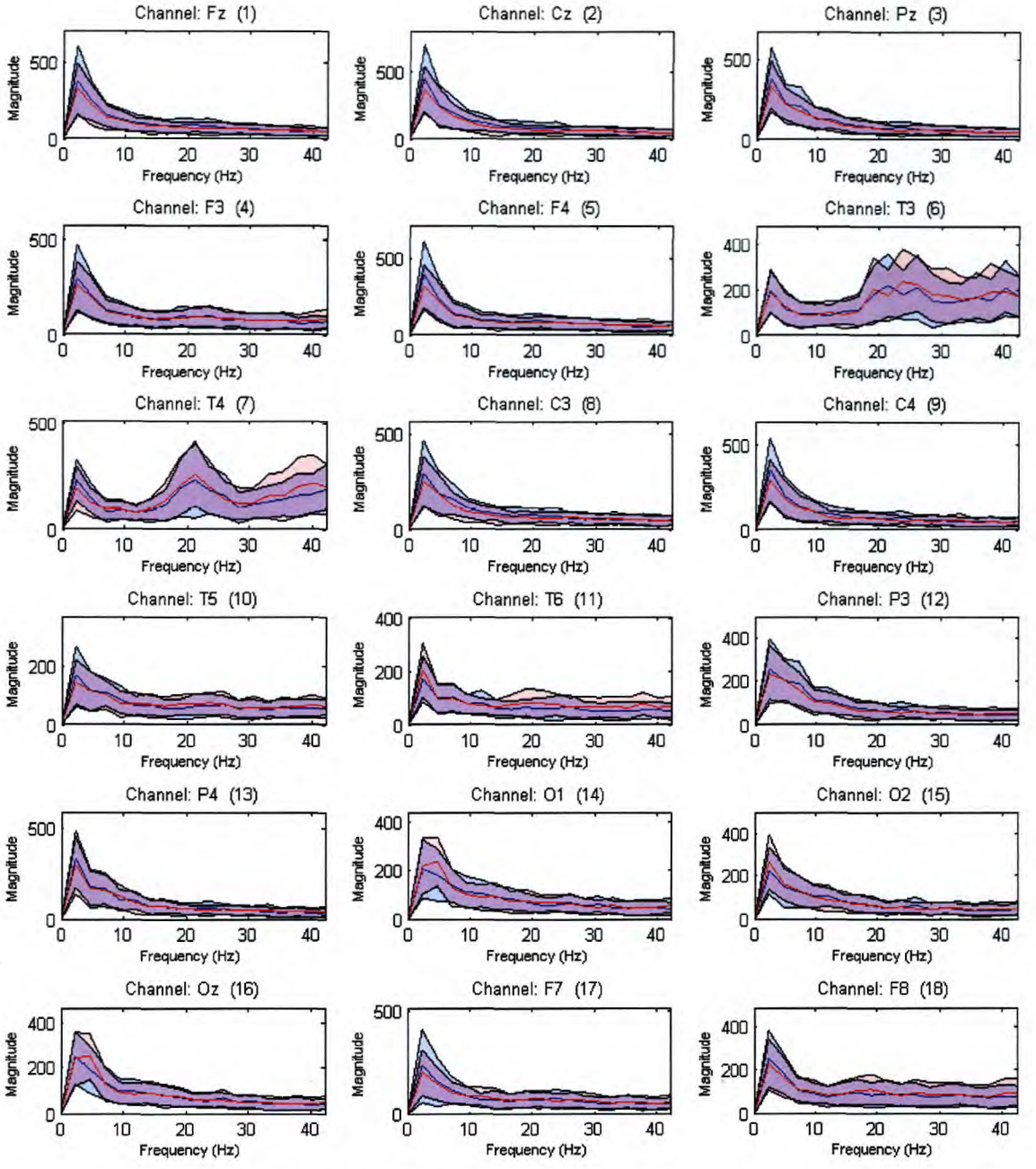


Figure A.15: Subject 6. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

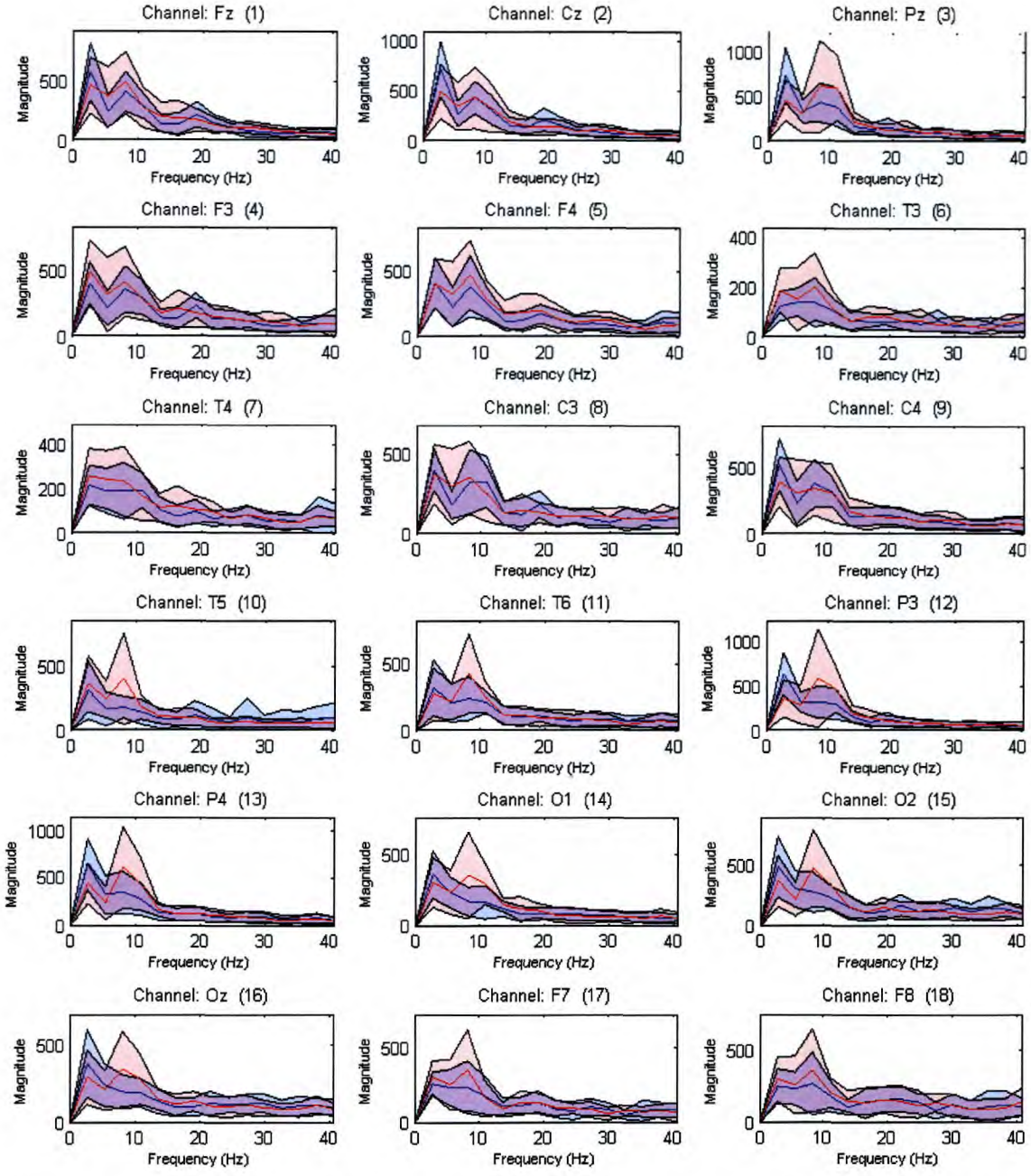


Figure A.16: Subject 8. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

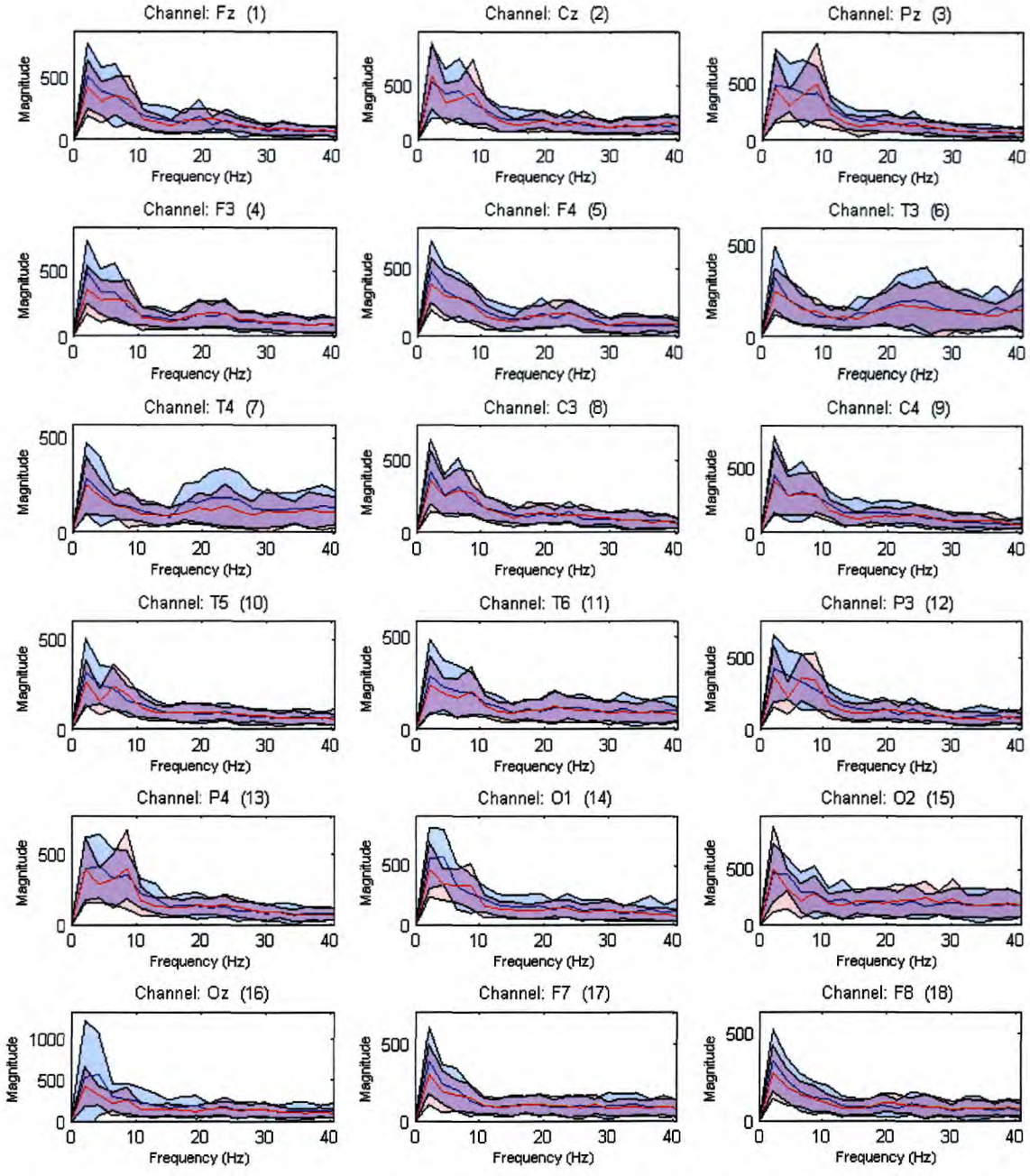


Figure A.17: Subject 10. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

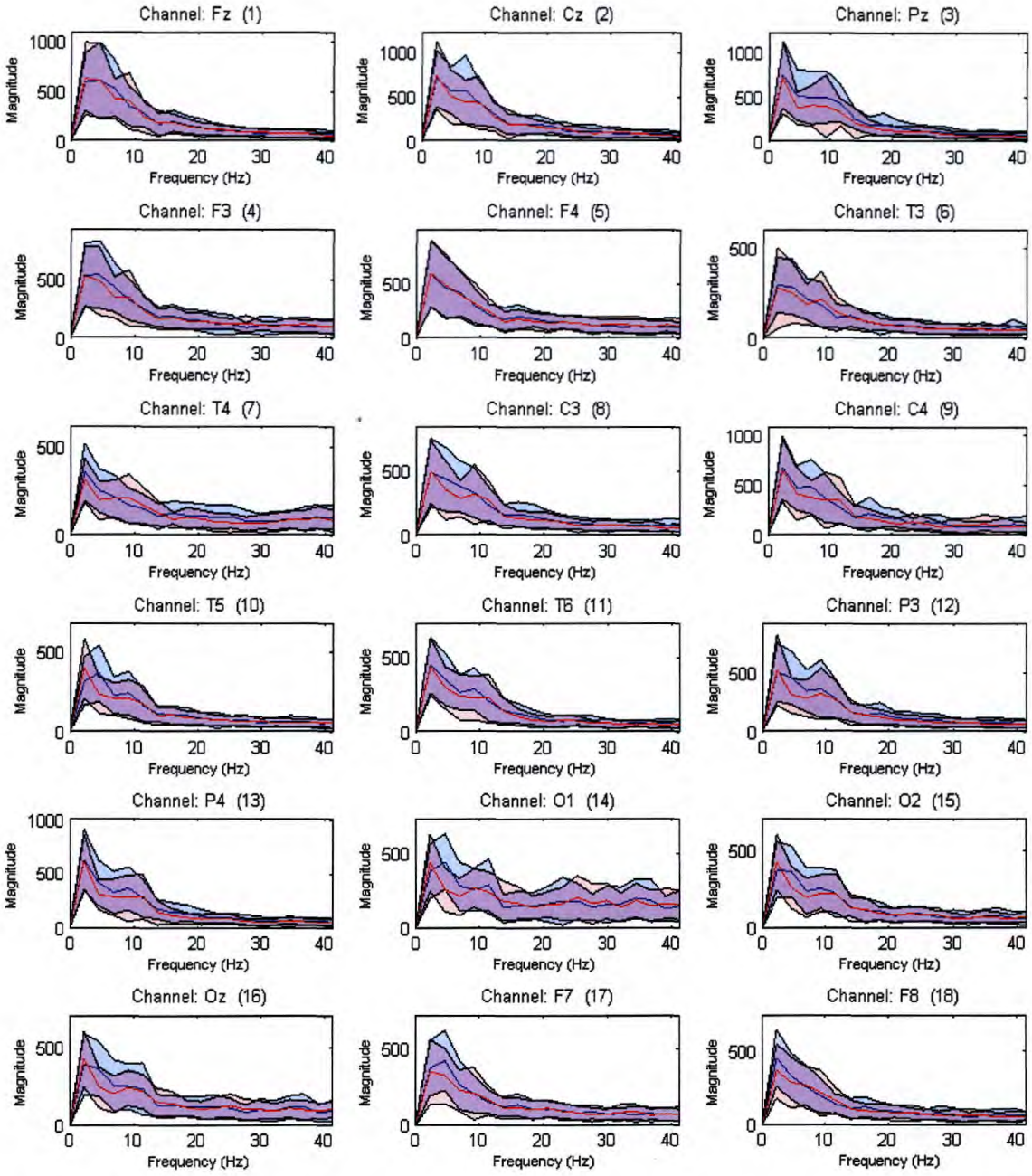


Figure A.18: Subject 11. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

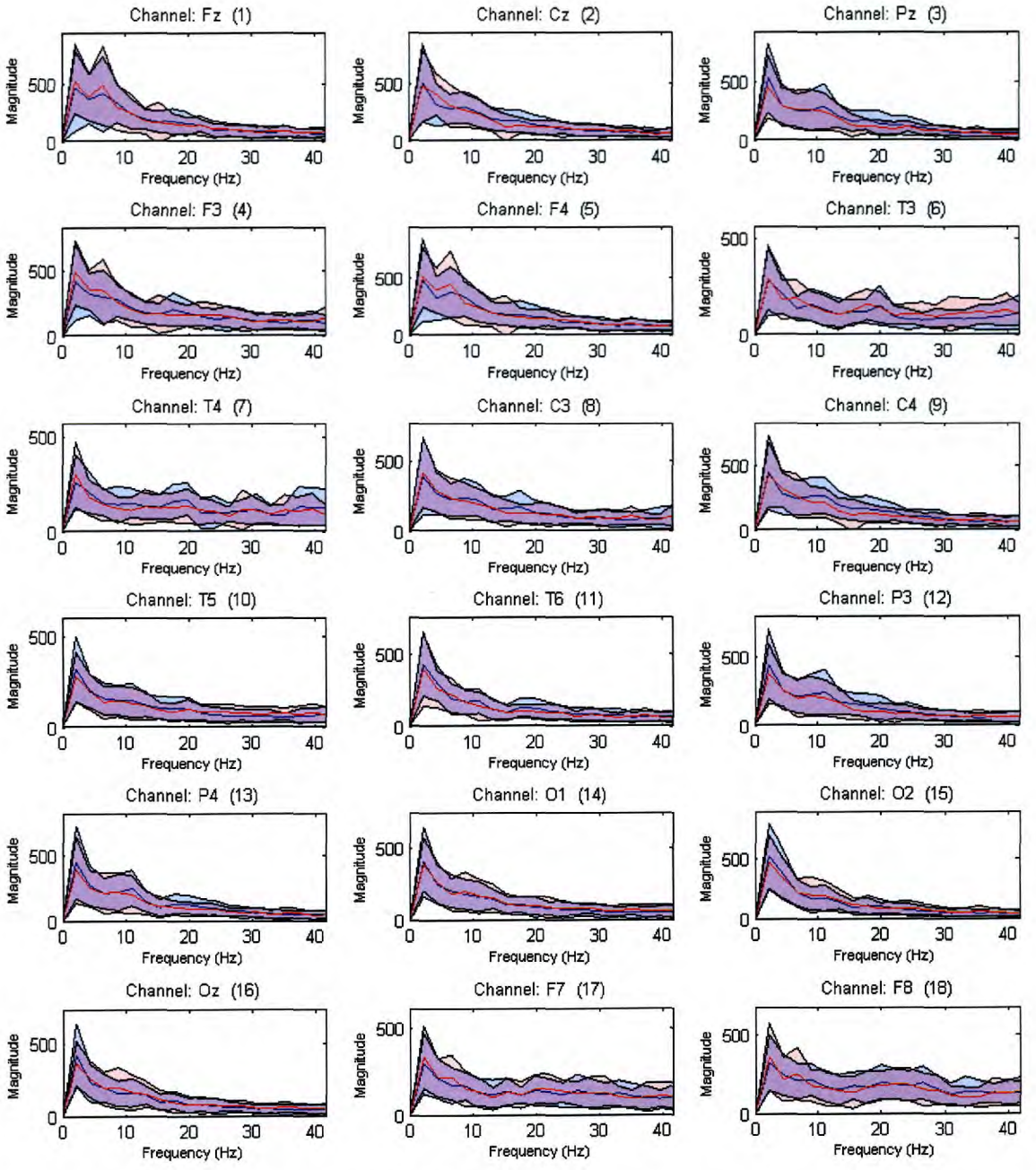


Figure A.19: Subject 13. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

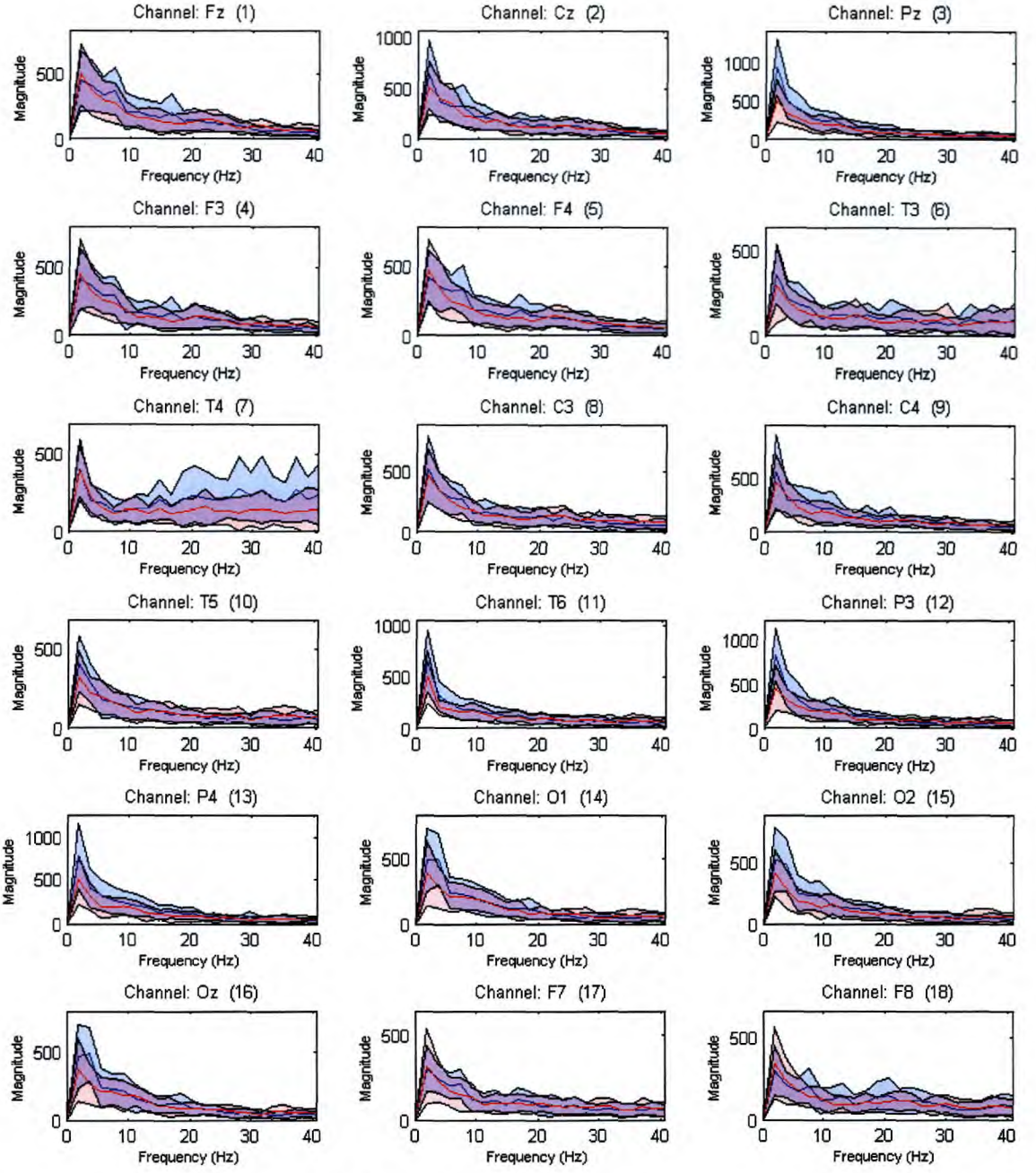


Figure A.20: Subject 14. The blue and red lines are the averages of the magnitude of the spectrum over the EEG signals of valid trials for classes "success" (blue) and "failure" (red). The area between the average signal plus and minus one standard deviation is marked with blue ("success") and red ("failure"). The spectrum plotted is for EEG signals truncated between the stimulus onset and the time point of the subject's quickest reaction.

Appendix B

Computation of the confidence interval

In the experiments of this thesis we gave to each classifier the feature vectors of the trials in the testing set and calculated an estimate for its correct classification rate as the percentage of the correctly classified trials. This procedure can be described as follows. Let us assume a random variable x which takes the value 1 when a trial is correctly classified and 0 in the opposite case. The correct classification rate is the mean value m of x. We make an estimate of the mean of x by sampling it L times, where L is the size of the testing set. Of course the larger L is, the more confident we are that the mean of x is close to the estimate we compute.

We are interested in finding a confidence interval for the correct classification rate m using the estimate we made. This is an interval where m lies with a high probability p. We choose here $p = 0.95$. The procedure of finding this interval is based on the procedure described in [21], with the difference that the variance of x is computed analytically as a function of m and incorporated in the results.

The procedure can be described as follows. Let us assume that we draw all possible samples of size L from the population of trials and for each sample we compute an estimate $\overline{x}$ of the mean of x. The probability distribution of $\overline{x}$ is named *sampling distribution.* If L is large enough (i.e. $L > 30$) then it can be proven using the Central Limit Theorem

that the sampling distribution is normal with mean m and variance s^2/L, where s^2 is the variance of x [21]. Then $z = \frac{\overline{x}-m}{s/\sqrt{L}}$ is $\mathcal{N}(0,1)$ and from the definition of the normal distribution we have that:

$$P\left(-1.96 < \frac{\overline{x}-m}{s/\sqrt{L}} < 1.96\right) = 0.95 \Leftrightarrow P\left(\overline{x} - 1.96\frac{s}{\sqrt{L}} < m < \overline{x} + 1.96\frac{s}{\sqrt{L}}\right) = 0.95 \tag{B.1}$$

Thus the 95% confidence interval of the correct classification rate m is given by the double inequality of Eq. (B.1). However, this inequality has to be simplified as the standard deviation s is a function of m. Indeed, if the probability of x taking the value 1, i.e. the classifier makes a correct classification, is p_1, then $m = p_1$ and $s^2 = E[x^2] - m^2 = p_1 1^2 + (1-p_1)0^2 - m^2 = m - m^2$. The results of the following analysis have also been reported in [92].

Let us first consider the right part of the inequality in Eq. (B.1), which gives:

$$m < \overline{x} + 1.96\frac{s}{\sqrt{L}} \Leftrightarrow s > \frac{\sqrt{L}(m-\overline{x})}{1.96} \tag{B.2}$$

The inequality of Eq. (B.2) is true for $m \leq \overline{x}$ as s is positive and for $m > \overline{x}$ we have:

$$\begin{aligned} s^2 > \frac{Lm^2 + \overline{x}^2 L - 2L\overline{x}m}{3.8416} \Leftrightarrow m - m^2 > \frac{Lm^2 + \overline{x}^2 L - 2L\overline{x}m}{3.8416} \Leftrightarrow \\ \Leftrightarrow (L+3.8416)m^2 - (3.8416 + 2L\overline{x})m + L\overline{x}^2 < 0 \end{aligned} \tag{B.3}$$

The two roots of the second order polynomial of Eq. (B.3) are:

$$\begin{aligned} D &= (3.8416 + 2L\overline{x})^2 - 4(L+3.8416)L\overline{x}^2 \\ m_1 &= \frac{3.8416 + 2L\overline{x} - \sqrt{D}}{2(L+3.8416)} \\ m_2 &= \frac{3.8416 + 2L\overline{x} + \sqrt{D}}{2(L+3.8416)} \end{aligned} \tag{B.4}$$

and thus the solution of Eq. (B.2) is the union of the intersection of $m_1 < m < m_2$ and $m > \overline{x}$ with $m \leq \overline{x}$. Equivalently, the left part of the inequality in Eq. (B.1), i.e.

$\overline{x} - 1.96\frac{s}{\sqrt{L}} < m$, has as solution the union of the intersection of $m_1 < m < m_2$ and $m < \overline{x}$ with $m \geq \overline{x}$. The intersection of the two unions is $m_1 < m < m_2$ so Eq. (B.1) can be written us:

$$P(m_1 < m < m_2) = 0.95 \tag{B.5}$$

where m_1 and m_2 are given from Eq. (B.4). Thus the interval between m_1 and m_2 is the 95% confidence interval for the correct classification rate. This means that if we get an estimate $\overline{x}$ for the correct classification rate m of one classifier then the latter lies with probability 95% in the interval between m_1 and m_2.

Appendix C

Significance test

Let us assume that we get an estimation $\widehat{m}_1$ for the correct classification rate of a classifier A and an estimation $\widehat{m}_2 > \widehat{m}_1$ for another classifier B. Let us denote by x_1 and x_2 the random variables of the two classifiers respectively, in the way defined in Appendix B. We also define the random variable $z \equiv x_2 - x_1$. We want to check whether the difference between $\widehat{m}_1$ and $\widehat{m}_2$ is adequate to be confident that the correct classification rate of B is larger than the one of A.

In order to check this we make the following significance test. We test the null hypothesis H_0 against the alternative hypothesis H_1, where:
H_0: Classifier A has equal or smaller classification rate than classifier B, i.e. $E[z] \leq 0$.
H_1: Classifier B has a larger classification rate than classifier A, i.e. $E[z] > 0$.

Let us assume that H_0 is true and particularly that the two classifiers have equal classification rates, i.e. $E[z] = 0$. Then the random variable $\overline{z}$, the values of which are the estimations of the mean of z using L samples, is $\mathcal{N}(0, s^2/L)$, where s^2 is the variance of z [21]. From the definition of the normal distribution we have:

$$P(\overline{z} \geq \widehat{m}_2 - \widehat{m}_1) = 1 - \frac{\sqrt{L}}{\sqrt{2\pi}s} \int_{-\infty}^{\widehat{m}_2 - \widehat{m1}} e^{-\frac{L\overline{z}^2}{2s^2}} d\overline{z} = p \qquad \text{(C.1)}$$

The variance of z s^2 can be computed from $s^2 = s_1^2 + s_2^2$, where s_1^2 and s_2^2 are

the variances of x_1 and x_2, respectively. This is because the variance of the difference of two independent random variables is equal with the sum of the variances of the two variables [21]. The variances s_1^2 and s_2^2 can be estimated from the available samples using formulae:

$$\begin{aligned} s_1^2 &= \sum_{i=1}^{L} \frac{(x_1^{(i)} - \widehat{m}_1)^2}{L-1} \\ s_2^2 &= \sum_{i=1}^{L} \frac{(x_2^{(i)} - \widehat{m}_2)^2}{L-1} \end{aligned} \tag{C.2}$$

Thus, the probability of observing a difference $\widehat{m}_2 - \widehat{m}_1$ or a larger one if the two classifiers have equal classification rates is p given by Eq. (C.1). The probability p is the observed level of significance of the observation $\widehat{m}_2 - \widehat{m}_1$ when $E[z] = 0$. In the general case of H_0 being true, i.e. $E[z] \leq 0 \Rightarrow E[\overline{z}] \leq 0$, the probability of observing a difference $\widehat{m}_2 - \widehat{m}_1$ or a larger one is smaller or equal with p.

In practice we calculate p using Eq. (C.1) and if it is sufficiently small (smaller than 5% or 1%) we reject the null hypothesis and we are confident that classifier B has a larger correct classification rate than classifier A. In the opposite case, we have to accept the null hypothesis which means that the difference between the estimated classification rates $\widehat{m}_1$ and $\widehat{m}_2$ is not large enough, for the given number of samples L, to be confident that B is better than A.