DIRECT BRAIN-COMPUTER COMMUNICATION THROUGH SCALP RECORDED EEG SIGNALS

THÈSE N^o 3019 (2004)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

INSTITUT DE TRAITEMENT DES SIGNAUX

SECTION D'ÉLECTRICITÉ

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Gary Nelson Garcia Molina

Ingénieur électricien diplômé EPF
de nationalité bolivienne

acceptée sur proposition du jury:

Prof. Juan Mosig, président du jury
Prof. Touradj Ebrahimi, directeur de thèse
Dr. Thomas Koenig, rapporteur
Prof. Yves Biollay, rapporteur
Prof. Ferran Marqués, rapporteur
Dr. Jean-Marc Vesin, rapporteur

Lausanne, EPFL
2004

Acknowledgments

This thesis is the result of three years work and it would never have been possible for me to accomplish this without the help, support and encouragement from many people. First, I wish to express my gratitude to Prof. Touradj Ebrahimi for giving me the possibility to do this thesis in his group. His engagement in my research and the time he spent with me discussing different problems ranging from philosophical issues to technical details have been essential for the results presented here.

I am specially grateful to Dr. Jean-Marc Vesin with whom I have been working very closely in most of the research presented here. His never ending stream of ideas and his passion for research were a source of inspiration for me.

My gratitude to Prof. Juan Mosig, Prof. Ferran Marqués, Prof. Yves Biollay, and Dr. Thomas Koenig for accepting to be part of the committee, and for their valuable comments on my work. I would like to express special thanks to Prof. Yves Biollay for his interest in my work and the fruitful discussions we had.

For his valuable help in shaping many ideas for this work, I would like to acknowledge Ulrich Hoffmann with whom I had the privilege to work in close collaboration.

I also wish to thank Lam Dang who provided me with valuable inputs to improve the comprehensibility of this document and foremost because during these years we shared our interest for signal processing and machine learning.

Several friends, including Jonathan Nieto, Emir Vela, and Abel Villca participated in the experiments, and made useful suggestions to improve the usability of the system.

For the help and support in the technical and administrative matters, I would like to thank Gilles Auric, Marianne Marion, and Fabienne Vionnet.

Finally, my sincere gratitude goes to my family for their continuous support and encouragement.

Contents

Version Abrégée

Les signaux obtenus par électroencéphalogramme (EEG) fournissent des indices sur l'activité synaptique combinée des groupes de neurones. En plus de leurs applications cliniques, les signaux EEG peuvent être utilisés en tant que support pour le développement d'interfaces de communication directe entre cerveau et ordinateur (interface cerveau-ordinateur ICO).

Lorsque des activités mentales sont exécutées, des caractéristiques spécifiques apparaissent dans l'EEG. Si des actions (produites par l'ICO) sont mises en correspondance avec de types de caractéristiques associées à des activités mentales qui n'impliquent aucun effort physique, alors la communication par la simple pensée devient possible. L'utilisateur opère l'ICO en exécutant des activités mentales qui sont reconnues par l'ICO grâce à des modèles de reconnaissance ayant été établis lors d'une phase d'entraînement.

Dans le cadre de cette thèse, nous considérons le positionnement d'un objet dans un environnement bidimensionnel généré par ordinateur (EGO). L'objet peut être déplacé suivant quatre directions correspondant à des activités mentales différentes. Le fonctionnement de l'ICO est asynchrone, à savoir que le systéme est actif en permanence et génère de mouvements de l'objet seulement lorsqu'il reconnaît l'une des activités mentales correspondantes. L'ICO analyse de segments d'EEG et génère de mouvements d'après un ensemble de règles (règles d'action) qui sont adaptées au niveau d'expérience de l'utilisateur lors du contrôle de l'application.

Les signaux EEG sont de faible amplitude et sont donc particulièrement sensibles à des perturbations extérieures. De plus, les changements abrupts apparaissant lors des activités musculaires, en particulier oculaires (artefacts), peuvent entraver le fonctionnement de l'ICO et même mener à des conclusions erronées sur la capacité de l'utilisateur à contrôler l'ICO. Ainsi, il est particulièrement important de filtrer les perturbations extérieures et détecter les artefacts. Les perturbations extérieures sont filtrées à l'aide de techniques de traitement de signaux classiques et les artefacts sont détectés en utilisant un algorithme de détection d'événements basé sur des méthodes dites du type noyau. Les paramètres de détection sont calibrés au début de chaque expérience de façon interactive. Lorsqu'un artefact est détecté dans un segment d'EEG, l'ICO en avertit l'utilisateur au moyen d'un événement particulier qui se produit dans l'EGO.

L'analyse des propriétés des signaux EEG en temps, fréquence et phase fourni des mesures statistiques (attributs) qui sont utiles pour la reconnaissance des activités mentales à partir de segments d'EEG. Cependant l'analyse extensive dans les domaines temps, fréquence et phase produirait un très grand nombre d'attributs. Moyennant des hypothèses sur la nature des signaux EEG il est possible de réduire le nombre d'attributs nécessaires. Les attributs sont groupés au sein d'un vecteur d'attributs à partir duquel les modèles de reconnaissance sont établis en utilisant de concepts d'apprentissage artificiel. Du point de vue de l'apprentissage artificiel, des vecteurs d'attributs à faible dimension sont préférables car ils réduisent le risque de sur-apprentissage.

Les modèles de reconnaissance sont construits sur la base de la théorie de l'apprentissage statistique et plus particulièrement des méthodes du type noyau. L'avantage d'une telle approche réside dans le fait que les modèles de reconnaissance ainsi construits atteignent de taux de reconnaissance supérieurs aux autres tout en étant très flexibles.

Un fonctionnement adéquat de l'ICO requiert l'adaptation continue des modèles de reconnaissance à de possibles changements pouvant apparaître dans les signaux EEG, et résultant des conditions externes différentes et de l'habituation de l'utilisateur à l'ICO. Cette adaptation est implémentée au moyen de l'apprentissage dynamique des paramètres des modèles de reconnaissance. Ainsi, ces paramètres peuvent être mis à jour continuellement et de façon considérablement efficace en termes de temps de calcul.

A la fin d'une première série de séances d'apprentissage, les méthodes pour l'extraction des vecteurs d'attributs sont choisies (d'après un critère d'optimalité lié à l'erreur de reconnaissance), les modèles de reconnaissance pour chaque activité mentale son construits et les règles d'action sont établies. Durant ces séances les activités mentales son présentées suivant un plan qui est défini par rapport à un protocole d'apprentissage.

Dans les séances d'apprentissage suivantes, l'ICO donne un retour à l'utilisateur pour lui indiquer le degré de reconnaissance de l'activité mentale qu'il lui a été demandé d'exécuter. Ainsi, l'utilisateur peut moduler son activité cérébrale afin d'obtenir un retour positif. A la fin de chaque séance les modèles de reconnaissance sont mis à jour. Ceci est accompli aisément due à la nature dynamique des paramètres des modèles. Puisque les modèles de reconnaissance changent dynamiquement, les règles d'action doivent changer en conséquence. Ceci se fait automatiquement car les règles d'action dépendent des paramètres des modèles.

L'ICO développé dans le cadre de cette thèse, a été validé par des expériences sur six sujets ayant participé à neuf séances d'apprentissage. Les trois premières séances ont servi à choisir les méthodes d'extraction de vecteurs d'attributs, construire les modèles de reconnaissance initiaux et établir les règles d'action. Dans les six dernières séances, en plus de l'expérience avec retour, des expériences de positionnement de l'objet ont étés réalisées afin d'évaluer l'expérience acquise lors de chaque séance. L'évaluation a été effectuée suivant deux critères, à savoir le calcul théorique du taux de transfert d'information en considérant l'erreur de reconnaissance moyen sur les activités mentales et la mesure expérimentale du taux de transfert d'information associée au test de positionnement. Cette dernière présente l'avantage de refléter plus étroitement les capacités réelles du sujet. Les deux mesures de taux de transfert d'information ont augmenté au cours des six dernières séances et ont atteint un taux moyen (sur les sujets) de 126 et 25 bits par minute respectivement.

Abstract

Scalp recorded electroencephalogram signals (EEG) reflect the combined synaptic and axonal activity of groups of neurons. In addition to their clinical applications, EEG signals can be used as support for direct brain-computer communication devices (Brain-Computer Interfaces BCIs). Indeed, during the performance of mental activities, EEG patterns that characterize them emerge. If actions executed by the BCI, are associated with classes of patterns resulting from mental activities that do not involve any physical effort, communication by means of thoughts is achieved. The subject operates the BCI by performing mental activities which are recognized by the BCI through comparison with recognition models that are set up during a training phase.

In this thesis we consider a 2D object positioning application in a computer-rendered environment (CRE) that is operated with four mental activities (controlling MAs). BCI operation is asynchronous, namely the system is always active and reacts only when it recognizes any of the controlling MAs. The BCI analyzes segments of EEG (EEG-trials) and executes actions on the CRE in accordance with a set of rules (action rules) adapted to the subject controlling skills.

EEG signals have small amplitudes and are therefore sensitive to external electromagnetic perturbations. In addition, subject-generated artifacts (ocular and muscular) can hinder BCI operation and even lead to misleading conclusions regarding the real controlling skills of a subject. Thus, it is especially important to remove external perturbations and detect subject-generated artifacts. External perturbations are removed using established signal processing techniques and artifacts are detected through a singular event detection algorithm based on kernel methods. The detection parameters are calibrated at the beginning of each experimental session through an interactive procedure. Whenever an artifact is detected in an EEG-trial the BCI notifies the subject by executing a special action.

Features that are relevant for the recognition of the controlling MAs are extracted from EEG-trials (free of artifacts) through the statistical analysis of their time, frequency, and phase properties. Since a complete analysis covering all these aspects, would result in a very large number of features, various hypotheses on the nature of EEG are considered in order to reduce the number of needed features.

Features are grouped into feature vectors that are used to build the recognition models using machine learning concepts. From a machine learning point of view, low dimensional feature vectors are preferred as they reduce the risk of over-fitting.

Recognition models are built based on statistical learning theory and kernel methods. The advantage of these methods resides in their high recognition accuracy and flexibility. A particular requirement of BCI systems is to continuously adapt to possible EEG changes resulting from external factors or subject adaptation to the BCI. This requirement is fulfilled by means of an online learning framework that makes the parameters of the recognition models easily updatable in a computationally efficient way.

After the completion of a series of training sessions, the feature extraction methods are chosen (according to an optimality criterion based on the recognition error), the initial recognition models are built for each controlling MA, and the action rules are set. In these sessions, the subject is asked to perform the controlling MAs in accordance to a training protocol which determines the training schedule.

In posterior training sessions, the BCI provides feedback indicating the subject how well the asked MA was recognized by the BCI. Thus, the subject can modulate his brain activity so as to obtain positive feedback. Furthermore, at the end of each session the BCI updates its recognition models. Such updating is straightforward as the recognition models can be dynamically updated, i.e. their parameters can be updated as new training data becomes available while progressively forgetting the contribution of old data. Because of the adaptation of the recognition models, the action rules must be adapted as well. This is achieved by considering, in the definition of the action rules, variables that change along with the recognition model parameters. The training schedule is decided based on the recognition error associated with each controlling MA, thus those MAs with large recognition errors are trained more often.

The BCI developed in this thesis was validated by experiments on six subjects who participated in nine training sessions. The first three training sessions served to select the feature extraction methods, build the initial recognition models, and set the action rules. In the last six sessions, in addition to the training with feedback, positioning tests were carried out to measure the controlling skills acquired by them during each session. The evaluation was done following two criteria, namely the computation of the theoretical information transfer rate using estimates of the average recognition errors over the controlling MAs, and an experimental measure of the information transfer rate corresponding to the positioning tests. The latter has the advantage of corresponding to a real controlling situation and consequently reflects more closely the actual controlling skills of a subject. Both information transfer rates increased during the last six sessions and reached an average, over subjects of 126 and 25 bits per minute respectively.

Notation and Terminology

Variables and Constants

$\aleph$	Space of ADBs power spectral densities
$\aleph_{\text{cal}}$	Calibration set to train the artifact detection algorithm
α_l	Expansion coefficient associated with training vector x_l
a_i	i-th autoregressive coefficient
$\mathbf{A}(i)$	i-th Autoregressive matrix
$\mathcal{A}_s$	Analytic signal associated with the signal s
β_n	Notch band
b_k	Offset of the membership function f_k
B	Frequency band
C_c	Center of the artifact detection sphere
$\daleth_k$	Action strength associated with MAk
f	Frequency in Herz
f_n	Notch frequency
f_s	Sampling frequency
ϕ_s	Instantaneous phase associated with the signal s
f_k	membership function associated with MAk
FTE	Fraction of training errors
FSV	Fraction of support vectors
H	Functional space
H_s	Hilbert transform of a signal s
κ	Penalization constant for the artifact detection procedure
k	Index used for mental activities ($k = 1, \ldots, \mathrm{N}_{\mathrm{MA}}$)
L	Number of elements in the training set
μ	Notch bandwidth
m	Electrode index
n	Time index

N_{B}	Number of frequency bands
$\mathrm{N}_{\mathrm{cal}}$	Number of elements in the calibration set
N_{e}	Number of electrodes
N_{MA}	Number of mental activities used to operate the BCI
$\mathrm{N}_{\mathrm{spt}}$	Number of samples per EEG-trial
ν_k	Training error bound associated with MAk
P_s	Power of signal s
$\mathrm{P}_m(\mathrm{B}_i)$	Power of s_m in the frequency band B_i
Q_m	Autoregressive order associated with s_m for the mappings: ψ_{AR} and ψ_{NAR}
Q	Order of the ψ_{MVAR} mapping
ρ_k	Threshold associated with the membership function f_k
R_c	Radius of the artifact detection sphere
σ	Parameter of the Gaussian Kernel
$\hat{s}$	Fourier transform of a signal s
$\bar{s}$	Analytic signal associated with a signal s
s_m	univariate signal recorded at the m^{th} electrode
S	Artifact-free EEG-trial
$\tilde{S}$	Non-preprocessed EEG-trial
τ	Time lag
$\mathrm{T}_{\mathrm{act}}$	Action period
U_m	Spectral order associated with s_m in the ψ_{NAR} mapping
υ	Frequency lag
ϑ	Discrete frequency index
x	Feature vector
X_k	Target set associated with MAk
$\mathcal{X}_k$	Feature vector space associated with MAk
y	Label $\in \{-1, +1\}$ of a pattern vector
$\mathrm{Y}(m1, m2, \mathrm{B})$	Synchronization between s_{m1} and s_{m2}, in the frequency band B
ζ_k	Normalized membership associated with MAk
Ω	Set of EEG-trials
Ω_k	Set of EEG-trials produced during the performance of MAk

Functions

$A_s(\cdot, \cdot)$	Ambiguity function of s
$A_{m1,m2}(\cdot, \cdot)$	Inter ambiguity function of signals s_{m1} and s_{m2}
$\langle \cdot, \cdot \rangle_{\mho}$	Inner product in vector space $\mho$
$\|\cdot\|_{\mho}$	Norm in vector space $\mho$
$c_k(\cdot, \cdot, \cdot)$	k-th loss function associated with MAk
$\mathrm{C}_{m1,m2}(\cdot)$	Coherence function of signals s_{m1} and s_{m2}

$\delta(\cdot)$	Discrete Dirac's delta function
$\mathcal{E}(\cdot)$	Prediction error power associated with the order of an AR model
$E[\cdot]$	Mathematical expectation
$\phi_k(\cdot)$	k-th application from the set of feature vectors $\mathcal{X}$ into the functional space H
$\vec{f}(\cdot)$	Vector of memberships
$f_k(\cdot)$	Membership function associated with MAk
$K_d(\cdot,\cdot)$	Polynomial Kernel function of degree d
$K_\sigma(\cdot,\cdot)$	Gaussian Kernel function with parameter σ
$\psi_{\mathrm{AR}}(\cdot)$	Autoregressive mapping
$\psi_{\mathrm{C}}(\cdot)$	Coherence mapping
$\psi_{\mathrm{NAR}}(\cdot)$	Non-stationary autoregressive mapping
$\psi_{\mathrm{MVAR}}(\cdot)$	Multivariate autoregressive mapping
$\psi_{\mathrm{P}}(\cdot)$	PSD mapping
$\psi_{\mathrm{Y}}(\cdot)$	Synchronization mapping
$R[\cdot]$	Risk functional
$R_{\mathrm{emp}}[\cdot]$	Empirical risk
$R_{\mathrm{reg}}[\cdot]$	Regularized risk
$R_\xi[\cdot]$	Relaxed risk
$R_{\mathrm{stoch}}[\cdot]$	Stochastic risk
$\mathrm{R}_s(\cdot,\cdot)$	Time autocorrelation function of signal s
$\mathcal{R}_s(\cdot,\cdot)$	Frequency autocorrelation function of signal s
$\mathscr{R}_s(\cdot,\cdot,\cdot,\cdot)$	Time-frequency autocorrelation function of signal s
$\mathrm{R}_{m1,m2}(\cdot,\cdot)$	Time inter-correlation function of signals s_{m1} and s_{m2}
$\mathcal{R}_{m1,m2}(\cdot,\cdot)$	Time inter-correlation function of signals s_{m1} and s_{m2}
$V[\cdot]$	Variance
$W_s(\cdot,\cdot)$	Wigner-Ville transform of s
$W_{m1,m2}(\cdot,\cdot)$	Inter Wigner-Ville transform of signals s_{m1} and s_{m2}

Abbreviations

ADB	Artifact detection block
AIC	Akaike information criterion
BCI	EEG based brain-computer interface
CRE	Computer-rendered environment
EEG	Scalp recorded electroencephalogram
ERD	Event related desynchronization
ERS	Event related synchronization
ERP	Event related potentials
FPE	Final prediction error
fMRI	Functional magnetic resonance imaging

HCI	Human-computer interface
HSMA	Hemispheric specialized mental activity
MA	Mental activity
MDL	Minimum description length
MEG	Meagnetoencephalogram
MRA	Movement related artifact
PET	Positron emission tomography
PSD	Power spectral density
RKHS	Reproducing kernel Hilbert space
SCPS	Slow cortical potential shift
SSVER	Steady state visual evoked response

List of Figures

List of Tables

Introduction 1

1.1 Context

The interaction between humans and computers has been an expanding field of research and development in recent years. The last two decades have witnessed the emergence of innovative human-computer interfaces that use voice, vision, haptics, and a combination of these (multimodality) as communication support.

Over the past decade effective attempts to achieve communication based on the analysis of electrical brain signals have begun. They were mainly fostered by the will to help people suffering from severe neuromuscular disorders by providing them with new communication channels. Recent advances in neuroscience, psychology, signal processing, machine learning, and hardware equipment made it possible to develop direct brain-computer communication systems (brain-computer interfaces BCI). A BCI is a communication system in which the messages or commands that the subject sends to the external world do not pass through the brain normal output pathways of peripheral nerves and muscles.

Like any communication system a BCI has inputs (electrophysiological signals resulting from brain activity monitoring) outputs (actions executed on an active device), elements that transform inputs into outputs, and a protocol that determines its operation.

Subjects control the active device by performing mental activities (MAs) which are associated with actions depending on the BCI application. Typical applications include cursor positioning, spelling programs, and command of external devices. The association between MAs and actions requires the selection of a set of MAs to operate the BCI and the identification of signatures in the brain activity that univocally characterize each MA.

Brain activity produces a wide variety of phenomena that can be measured with adequate sensors and have potential use in a BCI. Among the current monitoring methods scalp recorded electroencephalogram (EEG) constitutes an attractive choice for BCI systems because of its noninvasiveness, relative simplicity and low cost. We therefore focus our attention on the design and development of a scalp recorded electroencephalogram based BCI.

The types of MAs used in current BCIs were chosen in accordance with brain hemispheric specialization studies which suggest that the two brain hemispheres are specialized for different cognitive functions. The left hemisphere appears to be predominantly involved during verbal and other analytical functions and the right hemisphere in spatial and holistic processing. Thus typical MAs include: evoked response to external stimuli, imagined limb movement, and spatial, geometrical, arithmetical, and verbal operations.

Following the type of MAs they use, BCIs are categorized into evoked response and operant conditioning based ones. Evoked response based BCIs rely on subject attention-focusing to particular stimuli that are associated with actions. Operant conditioning based BCIs react to particular MAs (controlling MAs) executed by the subject. These MAs are recognized by the BCI through recognition models that are built during a training phase and continuously updated.

BCI operation can be of two types: synchronous, and asynchronous. In synchronous BCIs, the system is active only during some periods defined by the operator and the subject. Conversely, asynchronous BCIs are always active and react only when the subject performs the controlling MAs.

1.2 Objectives and Approaches

This thesis focuses on the development of an asynchronous operant conditioning based BCI which operates with four MAs in the framework of a 2D cursor positioning application.

The achievement of successful BCI operation depends on the mutual adaptation of both the subject and the system. Three levels of adaptation are identified:

First, when a subject accesses the BCI for the first time the latter adapts to that subject's signal characteristics. This level of adaptation is achieved during an initial training phase in which several signal processing methods to analyze EEG signals are considered and the optimal ones are selected.

Second, continuous adjustments are necessary to maintain subject controlling skills and reduce the impact of possible EEG variations. This level of adaptation is handled by means of recognition models based on kernel methods, which in addition to be highly suitable for the recognition task, have parameters that are automatically updated as more training data become available.

Third, the adaptive capacity of the brain needs to be engaged in the sense that, through feedback the brain activity will modify itself so as to produce those EEG patterns that best control the BCI. This level of adaptation is implemented through the establishment of operation rules that take into account subject controlling skills. Such rules are updated along with the recognition models.

In addition to these requirements, the following criteria must be fulfilled to achieve reliable communication: proper external noise removal, detection of subject generated artifacts (e.g. ocular and muscular artifacts), and adequate design of the training protocols and the evaluation scheme. Detection of artifacts is especially important as they can lead to misleading conclusions about the subject's ability to control the BCI. Indeed, the subject might be (voluntarily or not) controlling the BCI by generating artifacts.

1.3 Main contributions

The main contributions of this thesis can be summarized as follows.

- Definition of a general conceptual framework for BCI operation.
- Development of an efficient artifact detection algorithm whose parameters can be easily and rapidly adjusted during a short calibration phase that takes place at the beginning of each experimental session.
- Establishment of a general EEG analysis framework that can be extended to the analysis of any multivariate process. By assuming different hypotheses on the nature of the EEG process different feature extraction methods result.
- Establishment of a theory of recognition models based on kernel methods whose parameters can be efficiently updated as new training data become available.
- Derivation of a method to automatically select the parameters of the recognition models from statistical learning considerations.
- Choice of the optimal feature extraction method for each mental activity.
- Definition of action rules that determine BCI operation in function of the mental activities and subject controlling skills.
- Evaluation of the subject controlling skills through a theoretical and experimental measure of the information transfer rate.

1.4 Outline

The dissertation is organized in seven chapters. In chapter 2, the BCI architecture, operation mode and main concepts are defined. Furthermore, state-of-the-art implementations are presented and compared. The general architecture presented in this chapter serves as a thread for subsequent chapters in which we detail our solution.

Chapter 3 presents the basic elements of electroencephalography, the EEG acquisition procedure, and the preprocessing algorithms aiming at removing external noise and detecting artifacts in EEG signals. In chapter 4, a general framework for the analysis of EEG is developed and through the establishment of hypotheses on the nature of EEG signals, we derive several feature extraction methods. Chapter 5 describes the algorithms used to establish recognition models for the mental activities used to control the BCI as well as the dynamic updating of these models parameters. Chapter 6 presents the application of our BCI implementation in the framework of an asynchronous 2D positioning application.

Conclusions and an outline of some interesting future research directions are presented in Chapter 7. Complementary details on the nature of the recognition models discussed in Chapter 5 are given in Appendix A. Finally, complementary numerical results corresponding to the experiments that we carried out are provided in Appendix B.

Conceptual framework and state of the art 2

"An expert is a man who has made all the mistakes,
which can be made, in a very narrow field ." *Niels Bohr*

2.1 Introduction

A BCI is a communication system which allows a subject to act on his environment only by means of his thoughts, without using the brain's normal output pathways of peripheral nerves and muscles. Like any communication system a BCI has inputs (electrophysiological signals that result from brain activity monitoring) outputs (device actions), elements that transform inputs into outputs, and a protocol that determines its operation [167, 175].

The subject controls the active device by performing mental activities (MAs) which are associated with actions that are dependent on the BCI application (see Fig. 2.1). Typical BCI applications include control of the elements in a computer-rendered environment (e.g. cursor positioning [63, 176], visit of a virtual apartment [11, 12]), spelling programs (e.g. virtual keyboard [120]), and command of an external device (e.g. robot [107], prosthesis [128]).

The association between MAs and actions requires the selection of a set (controlling set) of MAs to which the BCI responds, and the identification of signatures in the brain activity that characterize each MA in the controlling set. These signatures are identified through the analysis of the electrophysiological signals recorded during the performance of the MAs in the controlling set.

The basic BCI design is depicted in Fig. 2.1. The monitoring of the subject's brain activity results in electrophysiological signals that are analyzed by the signal processing block. The latter computes measurements on these signals (features) that are grouped into a feature vector which is sent to the translation-into-commands block. This block recognizes

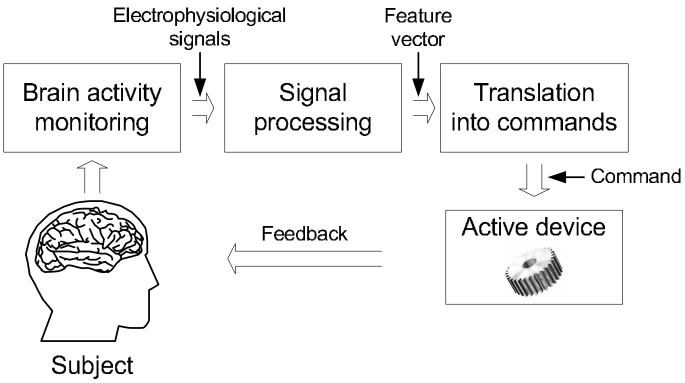
Electrophysiological signals
Feature vector
Brain activity monitoring
Signal processing
Translation into commands
Command
Active device
Feedback
Subject

Figure 2.1. Basic BCI design. Like any communication system a BCI has inputs (electrophysiological signals that result from brain activity monitoring) outputs (device actions), elements that transform inputs into outputs, and a protocol that determines its operation

the signatures characterizing the MAs in the controlling set and triggers the corresponding action on the active device. As this action can be noticed by the subject, it constitutes a feedback that he can use to modulate his mental activities so as to obtain the desired result.

In this chapter, we review the approaches in terms of brain activity monitoring and types of MAs used in current BCIs. In particular, we focus our attention on scalp recorded electroencephalogram based BCIs, propose a detailed architecture for such BCIs, and discuss different implementations in the framework of existing BCIs.

2.2 Brain activity monitoring

Brain activity produces a wide variety of phenomena that can be measured with adequate sensors and have potential use in a BCI. Among the current methods to monitor brain activity we mention: electrical potentials measurement (i.e. electroencephalogram and invasive electrophysiological methods), functional magnetic resonance imaging (fMRI), magneto-encephalogram (MEG), and positron emission tomography (PET). A more complete discussion on different methods to monitor brain activity can be found in [174].

At present, fMRI and MEG systems are large, very expensive, and require a magnetically shielded environment. Furthermore, since fMRI and PET depend on blood flow, they have long time constants and consequently they are less amenable to rapid communication [167]. Thus, electrical potential measurements constitute an adequate choice to monitor brain activity for BCI applications. Indeed, they have good time resolution, there is clear evidence that observable changes in the corresponding signals result from the performance of given mental activities, and put aside the invasive modalities, are easily acquired [65].

Electrical potentials can be measured as action or field potentials. Action potentials reflect the activity of individual neurons and are measured by electrodes placed in the brain cortex while field potentials reflect the combined activity of groups of neurons [167]. The latter are recorded as electroencephalogram (EEG) measurements from the scalp (in which case, they reflect the activity in large areas of brain cortex), from small electrodes within the brain (in which case, they reflect the activity in small immediately adjacent areas of tissue), or from epidural or subdural locations in between these two extremes. In general, the more the electrodes are invasive, the better the topographical resolution and the signal-to-noise ratio [167].

The spatial scale of an intracortical electrode (10 μm to 1 mm) depends on the size of the electrode tip, whereas the scale of unprocessed scalp EEG (6 to 10 cm) is largely independent of electrode size. Scalp EEG scale may be reduced (to 2 to 3 cm) by using a combination of multiple electrode arrays (64 or 128 electrodes) and high-resolution EEG algorithms (spatial filtering). Intracortical (or invasive) electrodes achieve higher spatial resolution at the expense of spatial coverage and significant increase in cost and risk [117, 118].

Invasive methods need neurosurgical implantation and were first used to record action potentials in the cerebral cortices of awake animals during movements [45, 76]. With operant conditioning methods, several studies showed that monkeys could learn to control the discharge of single neurons in their motor cortex [47, 149, 179, 180]. From such work, came the expectation that humans could develop similar control and use it to communicate or to operate neuroprostheses. Evaluation of this possibility was delayed by the lack of intracortical electrodes suitable for human use and capable of stable long-term recording from single neurons. Presently, few research groups are active in invasive BCIs in humans. In [88, 90, 91], a special type of cone electrodes is used to record stable action potentials from neurons in the motor cortex; such potentials are used to move a cursor to select icons or letters on a computer screen. The signal to noise ratio can be increased substantially by invasive technologies. However, research in this field is rather limited as people may be reluctant to agree to brain implants for research purposes especially because, at present, successful control or communication with an invasive BCI cannot be guaranteed.

Because it combines high temporal resolution, relative simple acquisition, and low cost, scalp recorded EEG is predominantly used in current BCIs. In the following, we concentrate on describing the architecture and functioning of scalp recorded EEG based BCIs (henceforth simply called BCI), and the mental activities used to operate such BCIs.

2.3 Type of mental activities in the controlling set

The MAs used in current BCIs were chosen in accordance with brain hemispheric specialization studies which suggest that the two hemispheres of the human brain are specialized for different cognitive functions. In particular, the left hemisphere appears to be predominantly involved in verbal and other analytical functions and the right one in spatial and holistic processing [5, 55]. Thus, typical MAs include: evoked responses to external stimuli, imagined limb movement, and spatial, geometrical, arithmetical and verbal operations.

BCIs can be categorized, by the type of MAs they use into evoked response and operant conditioning based BCIs. Both types are presented hereafter.

2.3.1 Evoked response based BCIs

Evoked responses are related to cognitive methods in psychology [163, 167] which consider the mind as an information processing device whose output depends on the relationship between stimuli and the activation of cognitive processes.

External visual or auditory events (e.g. blinking objects on a computer screen, flashing elements on a grid or brief sounds) elicit transient signals in the EEG that are characterized by voltage deviations known as event related potentials (ERPs). When the subject pays attention to a particular stimulus, an ERP that is time locked with that stimulus appears in his EEG. The changes in the EEG signals induced by the ERP can be detected by using averaging or blind source separation methods [75, 99]. If actions are associated with stimuli, the subject can gain control of the BCI by focusing his attention on the stimulus corresponding to the desired action.

Examples of BCIs functioning under evoked conditions are those using the P300 and the steady state visual evoked responses. We briefly present them hereafter.

P300 based BCIs

Infrequent or particularly significant auditory, visual, or somatosensory stimuli, when mixed with frequent or routine stimuli, typically evoke in the EEG over the parietal cortex a positive peak at about 300 milliseconds after the stimulus presentation [39, 40, 46, 155, 171].

The BCI presents different stimuli (previously associated with specific actions) to the subject. The P300 is prominent only in the responses elicited by the desired choice, and the BCI uses this effect to determine the subject's intent. In online experiments and offline simulations, a variety of different algorithms (e.g. averaging [75], independent component analysis [99]) for recognizing the desired choice have been evaluated, and the relationship between the number of trials per selection and BCI accuracy has been described [39, 40]. In people with visual impairments, auditory or tactile stimuli might be used [66, 143].

In [40], the user faces a 6×6 matrix of letters, numbers, and/or other symbols or commands. Every 125 milliseconds, a single row or column flashes; and, in a complete series of 12 flashes, each row or column flashes twice. The user makes a selection by counting how many times the row or column containing the desired choice flashes. EEG over the parietal cortex is digitized, the average response to each row and column is computed, and P300 amplitude for each possible choice is computed. In related works, single trial P300 evoked potentials were used to control some elements in a virtual environment [11, 12].

A P300-based BCI has an apparent advantage in that it requires no initial user training: P300 is a typical, or naive, response to a desired choice. At the same time, P300 and related potentials change in response to conditioning protocols [66, 109, 143, 154]. A P300 used in a BCI is also likely to change over time. Studies up to the present have been short-term. In the long term, P300 might habituate [138] so that BCI performance can deteriorate.

Steady state visual evoked response (SSVER) based BCIs

Flicker stimuli of variable frequency (2-90 Hz) elicit a SSVER in the EEG which is characterized by an oscillation at the same frequency as the stimulus. Thus, an SSVER can be detected by examining the spectral content of the signals recorded in the visual region, namely electrodes O1 and O2 of the 10-20 international system (see Chapter 3, Section 3.2.3).

When actions are associated with targets flickering with different frequencies the subject can control the BCI by gazing at the target corresponding to the desired action [25, 28, 56, 105]. BCIs based on this principle depend on the subject's ability to control gaze direction.

The advantage of BCIs based on evoked functional conditions resides in the fact that little or no training is necessary for a new subject to gain control of the BCI. However, the communication can be slow because of the averaging that is required to reliably detect an event related potential [175] and, in the case of an ERP-based BCI, the waiting time for the relevant stimulus presentation [11, 40]. Furthermore, the amplitude of the evoked response can diminish over time resulting from the user habituation to the stimulus [138].

2.3.2 Operant conditioning based BCIs

Operant conditioning is related to behavioral methods in psychology [153, 167]. According to them, the subject can acquire control skills through adequate feedback (operant conditioning feedback).

Effective attempts to provide control through operant conditioning feedback begun in the 50's [44, 84, 141] when some clinicians used the so-called neurofeedback to treat people suffering from attention deficit, hyperactivity, depression and even epilepsy. Based on the principle that functions of the autonomous and central nervous systems can be retrained for better adaptive functioning, neurofeedback practitioners trained their patients to self-regulate their brain activity through operant feedback. In some cases, they obtained astonishing results [141].

While a subject can learn to modulate his brain activity in order to make the BCI accomplish his intents, it is well known that the learning process (if successful) can take a considerable amount of time (up to several weeks [84, 141]). To handle this, the BCI simultaneously adapts to the user through machine learning algorithms. Thus, BCI operation relies upon the adaptation of two controllers, namely the subject and the computer [167].

In theory, the subject could choose any MA to control the BCI. Indeed, as the BCI learns how to recognize such an MA, one can expect that by means of machine learning algorithms and adequate feedback, the subject eventually will control the BCI using such an MA. However, MAs for which evidence that they are recognizable from EEG exists, are preferred in current BCIs; in particular, MAs studied in the framework of brain hemispheric specialization.

The subject and BCI adapt to each other according to the following process. In initial training sessions (training without feedback), the subject is asked to perform the MAs in the controlling set, while his EEG is recorded. The data are analyzed and using machine learning methods, a model to recognize each MA is set up. In subsequent training sessions (training with feedback), the subject is asked to perform a given MA (in the controlling set) and a feedback is provided, indicating the degree to which the BCI could identify the MA using the model computed in the previous training session. At the end of each training session, the recognition models are updated. This process is usually repeated many times in the course of BCI operation. Thus, the BCI is constantly being adapted to the subject and he can evaluate and improve his performance.

Examples of BCIs functioning under operant conditioning use slow cortical potential shifts, oscillatory sonsorimotor activity, and other hemispheric specialized MAs. We briefly present such BCIs hereafter.

Slow Cortical Potential Shifts (SCPSs) based BCIs

SCPSs last from a few hundred milliseconds up to several seconds and indicate the overall preparatory excitation level of a cortical network and they are universally present in the human brain. Negative SCPSs are typically associated with movement and other functions involving cortical activation, while positive SCPSs are usually associated with reduced cortical activation [18, 142].

In [73], several methods, ranging from low pass filtering to wavelet decomposition, for the extraction of SCPSs from EEG are described. For online applications low-pass filtering appears more suitable.

Subjects can learn through operant feedback to produce a SCPS in an electrically positive or negative direction for binary control [18, 124]. This skill can be acquired if the subjects are provided with a feedback on the course of their SCPS production and if they are positively reinforced for correct responses [17, 18].

In [18, 124], the binary control provided by the regulation of SCPSs and semantic considerations were used to implement a spelling program through which locked-in patients could communicate at a rate of one word per minute.

Oscillatory sensorimotor activity based BCIs

Populations of neurons can form complex networks which are at the origin of oscillatory activity. In general, the frequency of such oscillations decreases with an increase in the number of synchronized neuronal assemblies [152]. Two types of oscillations are especially important: the Rolandic mu rhythm in the range from 7 to 13 Hz and the central beta rhythm above 13 Hz, both originating in the sensorimotor cortex [80]. Sensory stimulation, motor behavior, and mental imagery can change the functional connectivity within the cortex and result in an amplitude suppression (event-related desynchronization ERD) or in an amplitude enhancement (event-related synchronization ERS) of mu and central beta rhythms [125].

Preparation and planning of self-paced hand movement results in a short-lasting desynchronization (ERD) of Rolandic mu and central beta rhythms. In [32, 160], electrocorticographic recordings exhibit ERD in the alpha band associated with hand and foot movement. The general finding is that similarly to the mu rhythm (around 10 Hz), beta oscillations desynchronize during the preparation and execution of a motor act [126].

Motor imagery may be seen as mental rehearsal of a motor act without any overt motor output. It is broadly accepted that mental imagination of movements involves brain regions/functions similar to that involved in programming and preparing such movements [81]. For example, during the imagination of a right-hand or left-hand movement, an ERD over the contralateral hand area was found [126]. This ERD is characteristic of the planning or preparation of a real movement [13].

Thus, the main difference between performance and imagery is that in the latter case execution would be blocked at some corticospinal level [35]. This observation opens the possibility of using motor imagery to provide a control option in BCI applications.

Oscillatory sensorimotor activity produced by the imagination of left/right hand movement and foot movement are used in a virtual keyboard application and to manipulate a hand orthosis in [128]. In a related work, the vertical movement of a 2D cursor is controlled by changes in the mu and beta rhythms in [176].

BCIs that use other hemispheric specialized mental activities (HSMAs)

In addition to imagined motor tasks, other mental activities for which evidence for hemispheric specialization was found, are: geometrical MA [116] (e.g. imagination of a geometric 3D object and the rotation of such an object), verbal MA [54, 94] (e.g. mental composition of a letter) and arithmetic MA [144](e.g. mental counting, multiplication, etc.)

Few research groups considered these mental activities for BCI applications. According to the results reported in [6, 60, 63, 106, 107], the communication bit rates and classification error percentages are comparable with those of other BCIs. Little attention was given to these mental activities because they did not seem ”natural” to control moving objects (e.g. prostheses, cursor on a computer screen, etc.). However, HSMAs open the possibility to implement more control capabilities and in certain cases they are easier to perform than imagined motor mental activities.

The categorization that we presented so far is merely conceptual. Indeed, different combinations of mental activities and functional conditions can be used in a BCI. For instance, in [89], an approach to decide on the type of recording (invasive or non-invasive), mental activities, and operating modes that are best suited for locked-in patients is presented. In this thesis, we use a combination of oscillatory activities, and other hemispheric specialized MAs in a 2D cursor positioning application (see Chapter 6).

2.4 BCI architecture and operation

The BCI architecture depicted in Fig. 2.2 is a detailed version of the general scheme in Fig. 2.1. The brain activity monitoring block is replaced by a scalp EEG acquisition system, the signal processing block is composed of the preprocessing and feature extraction modules, and the translation block is composed of the pattern recognition and action generation

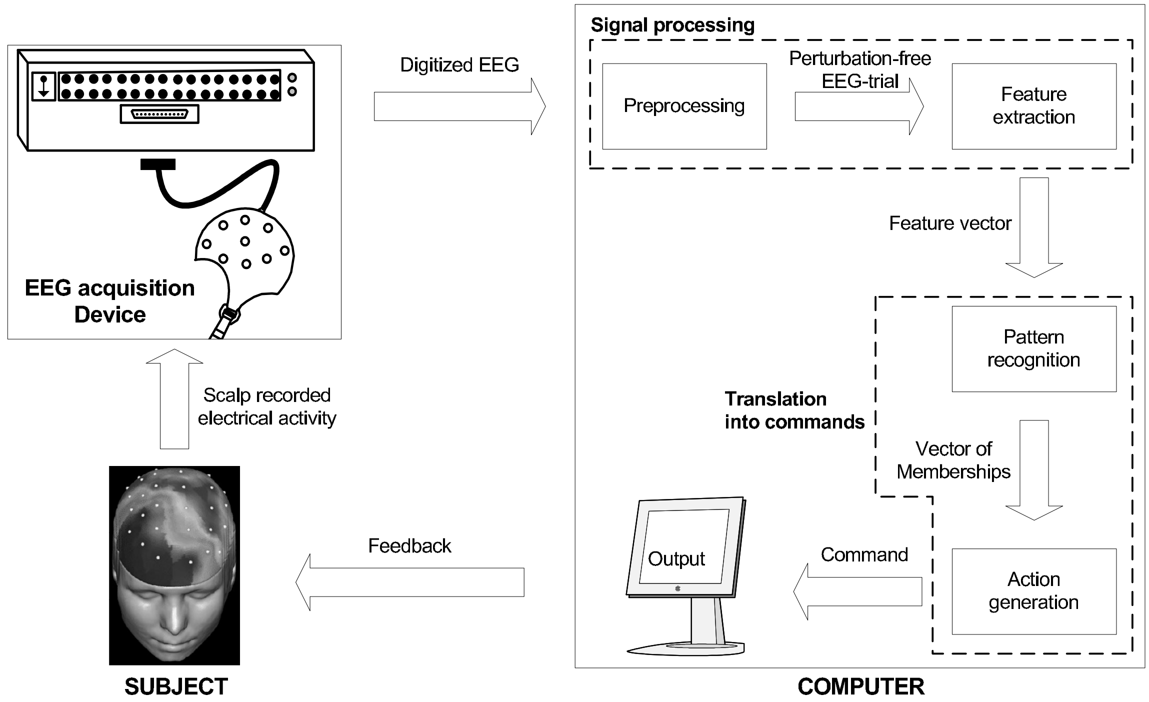

Figure 2.2. Architecture of the BCI system. With respect to the the basic design in Fig. 2.1, the brain activity monitoring block is substituted by a scalp EEG acquisition system, the signal processing block is subdivided into the preprocessing and feature extraction modules, and the translation block into the pattern recognition and action generation modules. The actions are displayed on a computer screen and constitute a feedback to the subject who can modulate his mental activity to make the BCI accomplish his intents.

modules. The actions being displayed on a computer screen constitute a feedback to the subject who can modulate his mental activity in order to make the BCI accomplish his intents.

EEG is recorded by using an array of electrodes which are affixed on the subject's scalp; the acquired signals are amplified, digitized and sent to the computer. EEG signals are analyzed in segments (EEG-trials) of a given duration that depends on the operation mode (i.e. whether the BCI operates in a synchronous or asynchronous manner) and type of mental activities. For instance, in a SCPS (see Section 2.3.2) based BCI the EEG-trial duration is in the order of eight seconds. Each EEG-trial is preprocessed so as to remove external (e.g. power line noise) and subject generated perturbations (e.g. ocular and muscular movement artifacts). EEG-trials (free of perturbations) are sent to the feature extraction module which extracts statistical measurements that are relevant for the recognition of the MAs in the controlling set, and groups them into a vector (feature vector) which is in turn sent to the pattern recognition module. The latter computes scores that indicate the likelihoods that the feature vector was produced during the performance of each of the MAs in the controlling set.

Since each MA defines, in the feature vector space, a subset of feature vectors produced

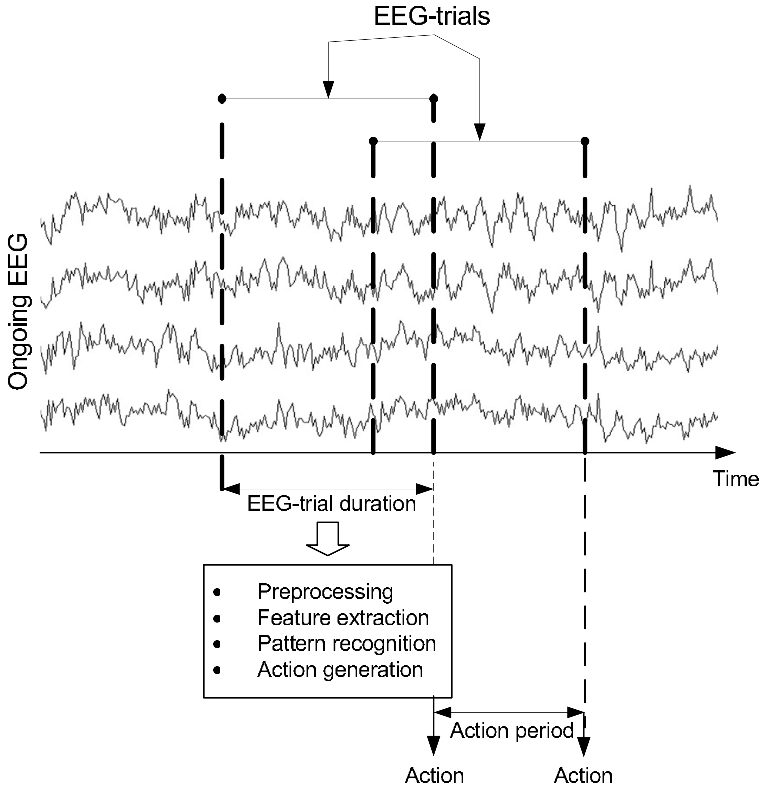

Figure 2.3. BCI operation. EEG signals are analyzed in segments (EEG-trials) of a given duration (EEG-trial duration) that depends on the operation mode and type of mental activities. Each EEG-trial is preprocessed in order to remove the external noise and detect artifacts. Then, relevant features are computed and grouped into a feature vector which is used to determine the likelihoods that the EEG-trial was generated during the performance of each MA in the controlling set. Finally, an action is executed that depends on the BCI action rules. Usually [128, 176], the unique action rule consists in executing the most likely MA. Other approaches consider the past actions as well [63].

during the performance of such MA, the scores determined by the pattern recognition module characterize the membership, with respect to each MA subset, of the current feature vector. Such scores are grouped into a vector of memberships that is sent to the action generation module which decides on the action that the BCI executes. Such an action depends on the BCI application and on the action rules. For instance in [128, 175], the action taken corresponds to the MA associated with the highest membership score while in [63], the action depends on the vector of memberships and on past actions. The time length between two consecutive actions is the action period (see Fig. 2.3).

The action rules depend on the BCI operation mode. In synchronous or cue-based BCIs [18, 128, 176], the system is active, i.e. generates actions, during some ”windows of opportunity” defined by the operator and the subject. Conversely, in asynchronous BCIs [19, 63, 107], the system is always active and a neutral action is executed when the current feature vector is considered as not belonging to any of the MAs in the controlling set. Clearly, the latter approach is more suitable for real applications. Indeed, an action should be executed only in response to any of the MAs in the controlling set. At present, by adequately adjusting the recognition parameters, it is possible to make the BCI generate an action only when a certain level of confidence on the recognition exists. On the opposite, it remains difficult to ensure adequate functioning when the subject is simultaneously engaged in other activities (e.g. speaking).

Current BCIs can be described by the general architecture in Fig. 2.2. In the following, we discuss the implementations of each module, the BCI applications, and evaluation criteria, in the framework of existing BCIs.

2.5 Preprocessing

The preprocessing module removes the external noise from EEG-trials and detects the presence of artifacts. In this thesis, the term noise refers to external perturbations, e.g. power line noise, and artifact to subject generated perturbations, e.g. muscular and ocular artifacts (see Chapter 3, Section 3.3). In general, the EEG-trials containing artifacts are discarded [70, 87, 101, 176] because the relevant information contained in the trial is masked by the artifact. Indeed, at frontal, temporal, and occipital locations particularly, ocular artifacts can exceed EEG [31, 68, 71] in amplitude.

Furthermore, the presence of artifacts can lead to misleading conclusions about the subject's ability to control the BCI. Indeed, the subject might be (voluntarily or not) controlling the BCI by generating artifacts [167].

In [107] it is suggested that artifacts do not need to be identified because the BCI is trained to recognize the MAs in the controlling set and consequently it automatically rejects artifacts. In this thesis, the artifacts are detected and special actions are generated to indicate to the subject whether an ocular or muscular artifact was detected (see Chapter 3). Thus, the subject can auto-regulate the artifacts he produces.

Electromagnetic and EEG equipment noise are narrow band pass signals. Thus, removing them through hardware or software filtering is straightforward. Typically [102, 106, 128], EEG signals are filtered in the 0.5-40 Hz frequency band, i.e. the effective EEG frequency support. As power line and other electromagnetic noise sources have frequency supports beyond 40 Hz such filtering removes most of this noise.

It is worth mentioning that while in the BCI framework they are treated as artifacts, muscular and eye movements are used as information support in other human-machine interaction systems [8, 158].

In this thesis, the power line noise is removed through notch filtering and the artifacts are detected by adapting the outlier detection framework presented in [157].

2.6 Feature extraction

The feature extraction module is in charge of computing statistical measurements or features on the EEG-trial (free of perturbations), delivered by the preprocessing module, that are relevant for the recognition of the MAs in the controlling set. Such measurements are grouped into a feature vector.

Thus, the feature extraction module maps the EEG-trial set into a feature space. The mapping properties are determined by the type of features (see Chapter 4). For instance, a mapping can be defined by the coefficients of an autoregressive model [6, 123, 146] fitted to the EEG-trial, the synchronization coefficients [16], or the powers in different frequency bands [128, 176]. It appears that the mappings achieving the best recognition performance in each MA in the controlling set are different(see Chapter 6). Thus, for its optimal operation a BCI should support several mappings. In [41], low recognition errors are obtained through the combination of mappings for the off-line classification of EEG-trials recorded during the imagination of left and right hand finger movements.

Most of the current BCIs use features based on the parametric and nonparametric spectral representations of EEG signals. In fact, such methods were extensively used to analyze EEG signals recorded during sleep, cognitive functions, epilepsy, and other clinical applications [78].

Nonparametric spectral representations are obtained through the discrete Fourier transform [114]. In [126], the power in the alpha and beta bands at electrodes located near the motor cortex are used to discriminate between EEG-trials produced between the imagination of left and right index finger movements. In [132, 175], the powers in the alpha band at frontal, central and occipital electrodes are used in a 1D cursor positioning application. In [107, 106], the powers in 2 Hz wide frequency bands from 8 to 30 Hz at every electrode are used as features to discriminate between five MAs, namely relaxing, imagination of left and right hand movements, rotation of a cube, and arithmetic.

Parametric spectral representations include: autoregressive (AR) [123] and autoregressive-moving-average (ARMA) [87] models for each EEG channel, and multivariate autoregressive models that characterize all the channels simultaneously [6].

The above mentioned approaches require EEG to be stationary. Since stationarity is not necessarily satisfied for EEG signals, alternative approaches which do not require stationarity were considered. Thus, in [129] lower recognition errors, with respect to the analysis of the powers in the alpha and beta band are reported by using adaptive autoregressive models. In [57, 58, 59, 60, 61, 62], we used time-frequency analysis to obtain features for the recognition of imagined left and right index finger movements, and mental multiplication in a cursor positioning application.

In [110], a set of space filters, optimized to discriminate between EEG-trials generated during the imagination of left and right index finger movements, are designed using the eigenvalue feature extraction [53] method. This method basically consists in simultaneously diagonalizing the mean autocorrelation matrices of EEG-trials recorded during the performance of each MA. By using this method in different frequency bands, we derive in [63] a set of space frequency filters to characterize each MA in the controlling set. Furthermore, in [60] we combine the diagonalization of the mean autocorrelation matrices with the analysis of time-frequency correlations to derive a set of features to discriminate between three mental activities, namely imagined left and right index finger movements and mental counting. To handle the continuous adaptation to possible changes in the user's EEG dynamics, the joint diagonalization of the mean autocorrelation matrices needs to be done periodically.

Nonlinear feature extraction methods were sporadically used in the BCI context. In [162], phase space reconstruction of the chaotic attractor associated with signals recorded at electrodes O1 and O2 are used to discriminate between three cognitive mental states, namely eyes open subject alert, eyes closed subject alert, and eyes closed subject performing visualization tasks. Phase synchronization [16] measures appear to be adequate for the recognition of certain mental states from EEG. Furthermore in [161] measures of the EEG complexity are used to provide 1D control.

In this thesis, we use several mappings from the EEG-trial set into different feature spaces. Such mappings are defined, for example, by the powers in certain frequency bands, autoregressive coefficients, coherence values, and synchronization. For each MA we select the mapping that produces the lowest recognition error (see Chapters 4 and 6).

2.7 Pattern recognition

As mentioned earlier, the feature vectors produced during the performance of a given MA define a set associated with such MA. The pattern recognition module measures the membership of a feature vector with respect to each MA set by means of machine learning approaches.

In [86], the distribution of the feature vectors belonging to a given MA set is assumed to be (multi-dimensional) Gaussian, thereby an optimal Bayesian classifier [53] is used to determine the memberships. If in addition to the Gaussian distribution assumption, one considers that the covariance matrices are all equal, linear discriminant analysis (LDA) can be used [64]. LDA classifiers are simple and can be easily updated. They are used in [129] to discriminate between EEG-trials produced during left and right hand movement imagination, and in [21, 87] to recognize the readiness potentials associated with the finger movements. Moreover, when the feature vectors are considerably high dimensional [58, 60] LDA classifiers appear to be the most suitable. However, if the separating boundary between different MA sets, in the space of feature vectors, is significantly different from linear, LDA classifiers can lead to high recognition error rates.

Clustering methods such as vector quantization and distance sensitive vector quantization (a variant of vector quantization in which the distance takes into account the discriminative power of each feature) are used in [83] and [133] respectively.

Multi-layer neural networks are used in [6, 106] to deal with general separation boundaries between the MA sets. Auto-associative neural networks which are able to directly operate in the time domain (the feature extraction procedure is therefore no longer necessary) are used in [38]. Linear and nonlinear classifiers are compared in [112] for BCI applications.

Other approaches include logistic regression classifiers [123], hidden Markov models [120] and microstates decomposition [52, 122] to classify sequences of features, and Bayesian time series classification [137].

Machine learning state-of-the-art kernel methods such as support vector machines and kernel novelty detection algorithms are used in [62, 59, 63, 61].

Few research groups [127] mention their strategy to continuously adapt their recognition algorithms. The usual approach consists in the periodical retraining of such algorithms. When kernel methods are used, a simple updating strategy can be applied [63] by considering the fact that in such methods, a reduced set of support vectors characterizes the data (training data) that was used to set up the recognition algorithm. Thus, when new training data become available, they are used in conjunction with the old support vectors to update the recognition algorithm.

In this thesis, we efficiently solve the updating problem by using an online learning of kernel-based classifiers (see Chapter 5).

2.8 Training protocols

The recognition algorithms mentioned in the previous section have parameters that are estimated during a training phase. In general, a first estimation of the recognition parameters take place after some training-without-feedback sessions in which the subject is asked to perform those MAs in the controlling set. Then, to improve control, in subsequent training sessions a feedback is provided, telling the subject how successfully the BCI recognized the MA he was asked to perform. Thus, in parallel to the subject, who can improve his controlling skills through feedback, the BCI updates the recognition parameters after each training session. Such training-with-feedback sessions need to be periodically programmed in order to maintain the BCI updated.

The schedule according to which the MAs are trained is random [128, 176], or decided by the subject [107]. In this thesis, the training schedule is determined according to the subject's performance; so that he can achieve similar controlling skills with each MA (see Chapter 6, Section 6.5). Thus, the MAs having the largest recognition errors are trained more frequently.

2.9 Action generation

The action executed by the BCI's active device is decided upon action rules which depend on the BCI application, its operation mode, and the vector of memberships delivered by the pattern recognition module. A dependency on the previous actions can be introduced in order to predict the next action (this can be especially important in a spelling application [124]) or to avoid abrupt changes [63].

In synchronous BCIs, the simplest action rule consists in executing the action corresponding to the highest score in the vector of memberships [128, 176]. On the opposite, when the BCI is asynchronous, a neutral action is executed whenever the BCI considers that the subject did not perform any of the MAs in the controlling set [63, 107].

In [63], we use an action rule in which the transition from one action to another depends on the probability of confusion between these actions. For large confusion probabilities the transition is done only if the subject confirms it a sufficient number of times. By improving his performance (i.e. decreasing the confusion probability), the subject can make the transition faster.

In this thesis, an action has an associated strength coefficient (see Chapter 6, Section 6.4) which depend on the level of confidence with which the corresponding MA is recognized. When this level is large the strength with which the action is executed is large.

2.10 Evaluation

The BCI performance can be evaluated as: 1) speed and accuracy in specific applications and 2) theoretical performance measured as an information transfer rate. The information transfer rate, as defined in [151], is the amount of information communicated per unit of time. This parameter encompasses speed and accuracy in a single value. The bit rate can be used for comparing different BCI approaches and for the measurement of system improvements [167].

The bit rate (in bits per minute) [46, 177] for a BCI with N mental activities in its controlling set, mean accuracy p_a (i.e. $1 - p_a$ is the mean recognition error), and action period $\mathrm{T}_{\mathrm{act}}$ (in seconds), is:

$$\text{Bit rate} = \frac{60}{\mathrm{T}_{\mathrm{act}}} \left(\log_2 N + p_a \log_2 p_a + (1 - p_a) \log_2 \frac{1 - p_a}{N - 1} \right)$$

In Fig. 2.4 we depict the bit rate (in bits per action period) for some typical values of N. Obviously, these curves make sense for values of p_a that are larger than $\frac{1}{N}$, i.e. the chance threshold.

2.11 BCI applications

BCI applications are in a continuum, from a binary on/off switch [18] at one end to hand orthosis at the other end [128]. Along this continuum, more and more degrees of control appear, and these may show finer gradations of control going from binary on-off to analog positioning.

In most of present-day BCIs, the output device is a computer screen. Cursor movement [63, 128, 176], spelling programs [124, 128], and environmental control panels [40] are among the applications rendered on a computer screen. Recently, virtual reality environments have been introduced to simulate the use of a BCI in real situations. For instance, in [11], the BCI is used to control a virtual apartment. Moreover, external hardware is used in [128], to control a hand orthosis and in [107], to control a robot.

The application can influence the choice of signals that are used to control the BCI. In a relatively precise control application, the slow changes of some EEG signals may be inadequate whereas changes in the oscillatory activity can be more suitable.

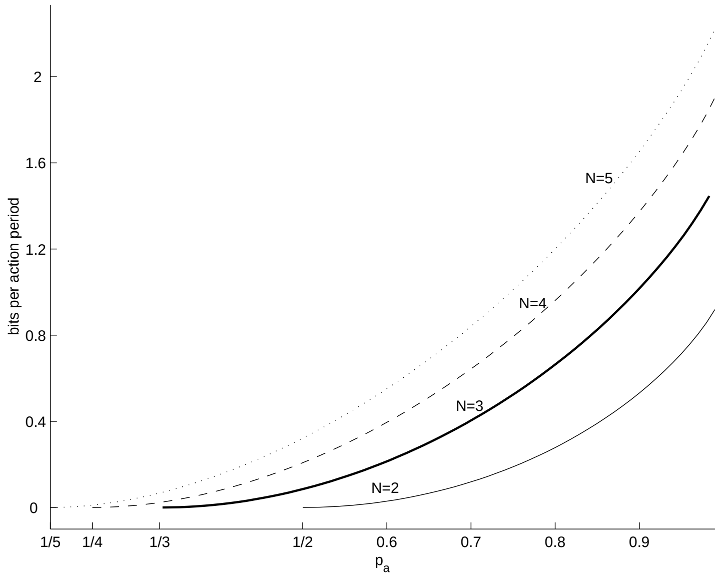

Figure 2.4. Bit rate in bits per action period. For a BCI with N mental activities in its controlling set and mean accuracy p_a, the information transfer, in bits per action period is: $\log_2 N + p_a \log_2 p_a + (1 - p_a) \log_2 \frac{1-p_a}{N-1}$

In this thesis we consider an asynchronous two dimensional cursor positioning application (see Chapter 6).

2.12 Summary

In this chapter we have presented the general architecture of a brain-computer interface and considered the possible choices in terms of brain activity monitoring and types of mental activities.

Because of its relative simplicity, low cost, and high time resolution, scalp recorded EEG constitutes the most used brain monitoring method in current BCIs. The choice of mental activities and conditions (evoked and operant) under which they are executed, were inspired by the results from brain hemispheric specialization studies and behavioral and cognitive psychological methods.

We focused our attention to scalp recorded EEG based BCIs that are controlled by mental activities performed under operant functional conditions. The detailed architecture, operating protocol, and implementation of current BCIs were then discussed. In particular, we considered: the implementation of the preprocessing, extraction of relevant features from EEG-trials, recognition algorithms, training protocols, and the rules that govern the execution of actions by the BCI. In table 2.1, we report the main features of current BCI implementations [1].

[1] Some systems were not included because their published descriptions did not contained enough information about their implementations or parameters choice

In the next chapters we present our implementation for each module of the general architecture presented here.

Group	**MAs** **EEG-trial duration/** **Action period** (milliseconds)	**Electrodes** **Features** **Recognition algorithm**	**Application** **Number of subjects** **Bit rate**	**Training time**
ABI Project European Union JRC [106, 107]	• Relax, imagination of left and right hand movement, cube rotation, and subtraction • 1000/500	• F3, F4, C3, Cz, C4, P3, Pz, P4 • Power in 2 Hz wide bands from 8 to 30 Hz • Neural network	• Asynchronous control of a mobile robot • Five subjects • 33 bit/min (max)	Days
EPFL Switzerland [63]	• Imagined left and right finger movements, mental counting, and object rotation. • 2000/500	• Fp1, Fp2, F7, F3, F4, F8, T3, C3, C4, T4, T5, P3, P4, T6, O1, O2 • Several types of feature vectors (see Chapter 4) • Online kernel based algorithm (Chapter 5)	• Asynchronous 2D Object positioning • Six subjects • 25 bits/min (avg) 35 bits/min (max)	Days
Neil Squire Foundation Canada [19, 101]	• Recognition of movement imagination against other MAs • 1000/62.5	• Bipolar recordings: F1-FC1, Fz-FCz, F2-FC2, FC1-C1, FCz-Cz, FC2-C2 • Bi-scale wavelength analysis • One-Nearest neighbor classifier	• Asynchronous switch • Seven subjects • 51 bits/min (max)	Weeks

Table 2.1. Comparison of current BCI systems (continued on next page)

Comparison of current BCI systems (continued from previous page)

Group	**MAs** **EEG-trial duration/** **Action period**	**Electrodes** **Attributes** **Recognition algorithm**	**Application** **Number of subjects** **Bit rate**	**Training time**
Technical University of Graz Austria [127]	• Imagination of left and right hand, and foot movements • 1000 to 3000/500	• 2 electrodes 2.5 cm anterior and posterior to electrode positions C3 and C4 respectively • Power in the alpha and beta band. Autoregressive coefficients • Linear discriminant analysis, hidden Markov models	• Synchronous virtual keyboard, hand orthosis control, and cursor movement • Four subjects, 17 bits/min (avg), 99 subjects (with the cursor positioning application [69] avg 11.3 bits/min)	Days
Tsinghua University China [28, 181]	• Steady state visual evoked response • 3000/3000	• O1 and O2 • Identification of the peaks in the spectrum, corresponding to the desired choice frequency	• Synchronous selection of targets on a panel for environmental control • Thirteen subjects • 27 bits/min (avg)	Minutes
University of Illinois USA [40, 46]	• P300 component of the event related potentials • 1500/1500	• Fz, Cz, Pz, O1, O2 • Averaging • Thresholding	•Synchronous 6 × 6 virtual keyboard • Ten subjects • 9 bits/min	Minutes

Comparison of current BCI systems (continued from previous page)

Group	**MAs** **EEG-trial duration/ Action period**	**Electrodes** **Attributes** **Recognition algorithm**	**Application** **Number of subjects** **Bit rate**	**Training time**
University of Rochester USA [11, 12]	• P300 component of the event related potentials • 1600/1600	• Fz, Cz, Pz, P3, P4 • Averaging • Thresholding	• Synchronous control of five elements in a virtual apartment • Nine subjects • 12 bits/min (avg)	Minutes
University of Tübingen Germany [18, 73]	• Control of slow brain potentials • 8000/8000	• Fz, Pz, Cz • Low-pass filtering • Thresholding	• Synchronous on/off switch • Eleven locked-in patients • 6 bits/min (avg)	Months
Wadsworth center USA [104, 103, 176]	• Mu and beta rhythm modulation • 200/100	• 64 EEG electrodes • Power in the mu and beta band • Linear classifier	Synchronous 2D positioning of a cursor • Eight subjects • 22.5 bits/min (avg)	Weeks

3 Acquisition and Preprocessing

"Nature uses as little as possible of anything" *Johannes Kepler*

3.1 Introduction

In the previous chapter we presented the general architecture of a BCI based on scalp recorded electroencephalogram, and discussed different implementations and operation modes. In this chapter we present a review of the physiological principles of electroencephalography, the recording procedure, and the methods we use to remove external noise and detect artifacts. A more detailed description of electroencephalography and related fields can be found in [115].

The extraction of information from EEG data is hindered by external noise and subject generated artifacts. Most sources of external noise can be avoided by appropriately controlling the environment in which the measurement takes place. Thus, power line noise can be easily filtered since it occupies a narrow frequency band that is located beyond the EEG band.

Subject generated artifacts (eye movements, eye blinks and muscular activity) can produce voltage changes of much higher amplitude than the endogenous brain activity. Even when artifacts are not correlated with tasks, they make it difficult to extract useful information from the data. In this situation the data are discarded and the subject is notified by a special action executed by the BCI. If the data containing artifacts were not discarded they could lead to misleading conclusions about the controlling performance of a subject. For instance, a subject could (voluntarily or not) be controlling the BCI by producing artifacts.

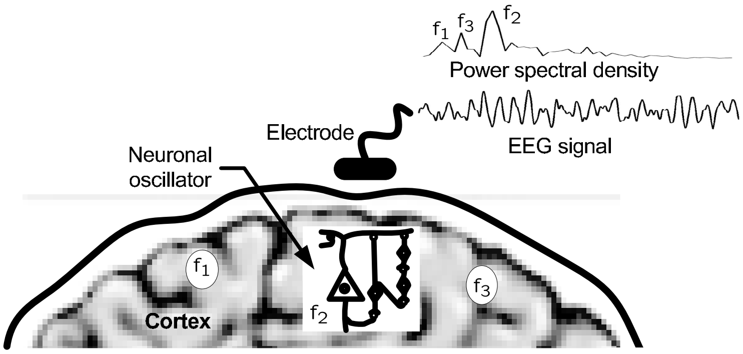

Figure 3.1. Origins of the rhythmic activity observed in EEG signals. The signal recorded at a particular electrode is composed of rhythms whose frequencies are visible in the signal power spectral density. These rhythms are produced by neuronal oscillators whose natural frequencies are determined by their internal cytoarchitecture.

3.2 An overview of electroencephalography

3.2.1 EEG origins

The generators of electric fields that can be registered with scalp electrodes are groups of neurons with uniformly oriented dendrites. Neurons communicate with each other by sending electrochemical signals from the synaptic terminal of one cell to the dendrites of other cells. These signals affect dendritic synapses, inducing excitatory and inhibitory postsynaptic potentials [44, 174]. The EEG is a result of the summation of potentials derived from the mixture of extracellular currents generated by populations of neurons. Hereby the EEG depends on the cytoarchitectures of the neuronal populations, their connectivity, including feedback loops, and the geometries of their extracellular fields. The main physical sources of scalp potentials are the pyramidal cells of cortical layers III and V[1].

The appearance of EEG rhythmic activity in scalp recordings results from the coordinated activation of groups of neurons, whose summed synaptic events become sufficiently large. The rhythmic activity may be generated both by pacemaker neurons having the inherent capability of rhythmic oscillations, and by neurons which cannot generate a rhythm on their own but can coordinate their activity through excitatory and inhibitory connections in such a manner that they constitute a network with pacemaker properties. The latter may be designated as neuronal oscillators [174]. The oscillators have their own discharge frequency (Fig. 3.1) which depends on their internal connectivity. The neuronal oscillators start to act in synchrony after application of external sensory stimulation or hidden signals from internal sources, e.g. resulting from cognitive loading.

[1] The brain cortex is composed of six layers, namely molecular layer (I), external granular layer (II), external pyramidal layer (III), internal granular layer (IV), internal pyramidal layer (V) and polymorphic or multiform layer (VI)

3.2.2 EEG Rhythms

The usual classification of the main EEG rhythms based on their frequency ranges is as follows: delta (2 to 4 Hz), theta (4 to 8 Hz), alpha (8 to 13 Hz), beta (13 to 30 Hz), and gamma (higher than 30Hz). However, this classification only partially reflects the functional variation of rhythmic activities. For example, EEG rhythms within the alpha range may be distinguished by their dynamics, place of generation and relation to certain behavioral acts. Since the pioneering work of Hans Berger in 1929 [15], the main EEG rhythm (the alpha one) has been known. This rhythm is typical of a resting condition and disappears when the subject perceives a sensory signal or when he makes mental efforts. It was shown that the alpha rhythm is generated by reverberating propagation of nerve impulses between cortical neuronal groups and some thalamic nuclei, interconnected by a system of excitatory and inhibitory connections and resulting in rhythmic discharges of large populations of cortical neurons [33].

The theta rhythm originates from interactions between cortical and hippocampal neuronal groups [108]. It appears in periods of emotional stress and during rapid-eye-movement sleep.

The delta rhythm appears during deep sleep, anesthesia, and is also present during various meditative states involving willful and conscious focus of attention in the absence of other sensory stimuli [48].

The neuronal oscillators, which generate the beta rhythm are located presumably inside the cortex [33]. The beta rhythm is typical of periods of intense activity of the nervous system and occurs principally in the parietal and frontal regions.

The basis for gamma oscillations is interneuronal feedback with quarter-cycle phase lags between neurons situated close to each other in local areas of the cortex [51]. It is thought that gamma oscillations are associated with attention, perception and cognition.

Most of the rhythms are rather widespread in brain structures. Induced gamma, theta and alpha rhythms were found in cortex, hippocampus, thalamus, and brain stem. In [50], the expression “common modes” was used for the existence of similar rhythms in various networks of the brain. This may play a role in the integration of activities of neuronal oscillators distributed over various brain structures. The candidate mechanism for such integration is coordination of the distant neuronal oscillators activity. The coordination concept (see Chapter 4 for a mathematical treatment) encompasses the interaction in time (as measured by the correlation function), frequency (as measured by the coherence function), time-frequency (as measured by the ambiguity function), and phase (as measured by the synchronization function).

The analysis of EEG rhythms and their interactions provide indices that are correlated with mental states such as attention [65], memory encoding [156], motor imagery [7, 128, 176] and perception/recognition [159].

3.2.3 EEG recording procedure

EEG recording is a routine procedure that usually includes the following steps: 1. The subject is seated in a comfortable chair in a dimly illuminated room; 2. Electrodes are placed on his head according to a certain scheme; 3. The reference electrode(s) are chosen; 4. Calibration of the acquisition system is executed; 5. EEG is recorded.

The silver/silver chloride (Ag-AgCl) electrodes are the most appropriate ones to record scalp EEG because they avoid potential shifts due to electrode polarization. To improve the conduction between the skin and electrode surfaces, electrode gel or salt solutions are used.

The scheme generally used for electrode placement is the 10-20 scheme [79], which is shown in Fig. 3.2. Even numbers indicate electrodes located on the right side of the head and odd numbers indicate electrodes on the left side. The letter before the number designates the general area of the cortex on which the electrode is located. A stands for auricular, C for central, Fp for frontal pole, F for frontal, P for parietal, O for Occipital, and T for temporal.

The most common way to place the electrode array on the scalp is the use of a cap (or helmet) with the electrodes fixed on it. Such devices can be placed and removed rapidly and cause a minimal unpleasant feeling. This is especially important in the BCI framework in which the subjects wear the cap for relatively long time (in the order of one hour). These caps automatically provide the electrode placement with appropriate interelectrode distance.

One of the important questions in EEG recording is that of the reference electrodes which should be placed on a presumed “inactive” zone. Frequently, this is the left or right earlobe or both of them (labelled as A1 and A2 in Fig. 3.2). If one earlobe electrode is used as a reference, there is the systematic decrease of EEG amplitudes at the electrodes which are closer to the reference side. If linked earlobes are used, this kind of asymmetry is avoided, but this linking distorts the EEG picture since the electric current flows inside the linking wire. Alternatively, the EEG may be recorded with any scalp electrode as a reference, and then the average reference is computed as a mean of all electrodes. The latter avoids all kinds of asymmetry and makes the EEG recorded in various laboratories comparable. In this thesis, we use Cz as physical reference and re-reference with respect to the signals average.

Since the frequency content of EEG signals is mainly confined to the 0-40 Hz band a minimum sampling rate of 100 samples/second is recommended [115]. This rate permits to analyze frequencies up to 50 Hz because the maximal allowed frequency of the input signal (the Nyquist frequency) should be half the sampling rate.

3.3 EEG perturbations

In the context of EEG driven BCIs, the signal is the endogenous brain activity measured as voltage changes at the scalp while a perturbation is any voltage change generated by

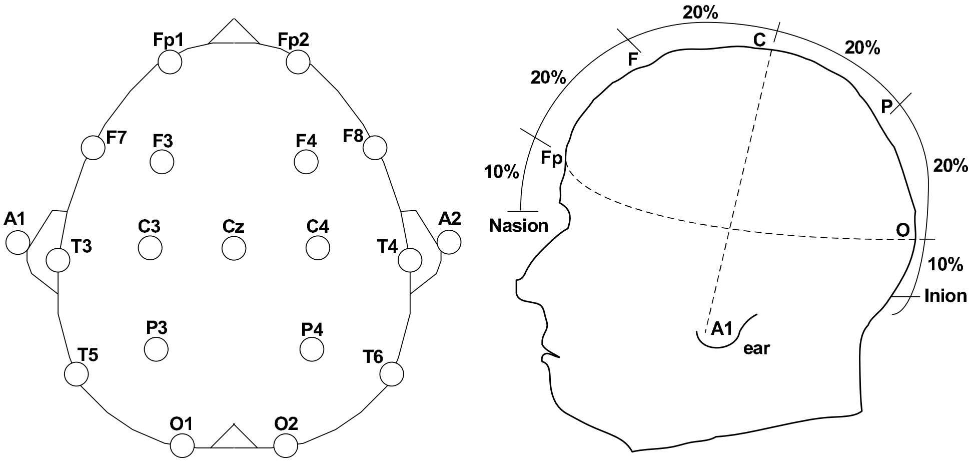

Figure 3.2. Electrodes placement according to the 10-20 international system. Even numbers indicate electrodes located on the right side of the head and odd numbers indicate electrodes on the left side. Capital letters are used to reference each cortical zone, namely frontal (F), central (C), parietal (P), temporal (T), and occipital (O). Fp and A stand for frontal pole and auricular respectively. The designation 10-20 comes from the percentage ratio of the inter-electrode distances with respect to the nasion-inion distance.

other sources. The perturbation sources include: electromagnetic interferences, eye blinks, eye movements and muscular activity (particularly head muscles). While the terms “noise” and “artifact” are often used interchangeably, in this thesis the term noise is used for external perturbations (e.g. power line noise) and artifact for subject related perturbations (e.g. muscular and eye movement artifacts).

- *Electromagnetic interferences.* Most of these interferences can be avoided or at least attenuated by controlling the environment in which the measurements are carried out. Nonetheless, since the BCI setup requires at least an amplifier connected to a computer, the EEG data can be corrupted by the noise from the A/C power supplies. These perturbations are usually well localized in frequency and located beyond the EEG band (see Fig. 3.3b).

- *Eye blink and eye movement artifacts.* Eye blink artifacts are very common in EEG data; they produce low-frequency high-amplitude signals that can be quite greater than EEG signals of interest (see Fig. 3.3c). Indeed, while regular EEG amplitudes are in the range of -50 to 50 microvolts eye blink artifacts have amplitudes up to 100 microvolts.

  Eye movement artifacts are caused by the reorientation of the retinocorneal dipole [121]. They are recognized by their quasi square shape and their amplitude in the range of that of regular EEG [121].

  Eye blink and eye movement artifacts (henceforth called ocular artifacts) often occur at close intervals as shown in Figure 3.3c. They are mainly reflected at frontal sites (e.g. electrodes Fp1, Fp2) although they can corrupt data on all electrodes, even those at the back of the head.

- *Muscular movement artifacts.* These artifacts can be caused by activity in different muscle groups. However, the activity in neck and facial muscles has more influence in EEG recordings. Muscular artifacts are characterized by their wide frequency content (see Fig. 3.3d). Depending on the location of the source muscles they can be distributed across different sets of electrodes. They mainly appear in temporal and parietal electrodes.

3.3.1 Perturbations handling

External interferences can often be attenuated by carefully controlling the environment in which the recordings are made. However, we have to deal with power noise as most of the equipment we use (computers, monitors, etc.) is connected to the power grid. Since the power line noise is well localized in frequency it can be easily filtered using a notch filter (as presented in Section 3.4.1).

Even if muscular and ocular artifacts are not correlated with the mental activities that the subject is executing, they make it difficult to extract useful information from the data. Furthermore, artifacts can lead to erroneous conclusions about the BCI controlling performance of a subject. Indeed, the BCI could be responding to muscular or ocular activity instead of genuine EEG. To prevent these errors we detect these artifacts, discard the corresponding data and notify the subject. The detection method is described in Section 3.4.2.

3.4 EEG preprocessing

As mentioned in Chapter 2, the EEG signals are processed in segments (EEG-trials) in which the BCI attempts to recognize the MAs in the controlling set. In this section we present the methods to remove the power line noise and detect eye and muscular artifacts in an EEG-trial. When an artifact is detected in an EEG-trial, the latter is not sent to the feature extraction module (see Fig. 3.14). Instead, the BCI notifies the subject by generating special actions that indicate if the detected artifact was muscular or ocular.

A digitized EEG-trial is represented by a real $\mathrm{N}_e \times \mathrm{N}_{\mathrm{spt}}$ matrix where N_e is the number of electrodes and $\mathrm{N}_{\mathrm{spt}}$ the number of samples per EEG channel. We denote as $\tilde{S}$ a non-preprocessed (raw) EEG-trial and as S an EEG-trial in which the power line noise was removed and no artifact was detected, so that it can be sent to the feature extraction module (see Fig. 3.14).

The rows of $\tilde{S}$ and S, which correspond to the EEG channels are denoted as $\tilde{s}_1, \ldots, \tilde{s}_{\mathrm{Ne}}$, and $s_1, \ldots, s_{\mathrm{Ne}}$ respectively. The indexes m and n are used to reference the electrode and time index respectively. Thus, $\tilde{s}_m(n)$ corresponds to the n-th sample of the m-th EEG channel or the (m, n)-th element of the matrix $\tilde{S}$.

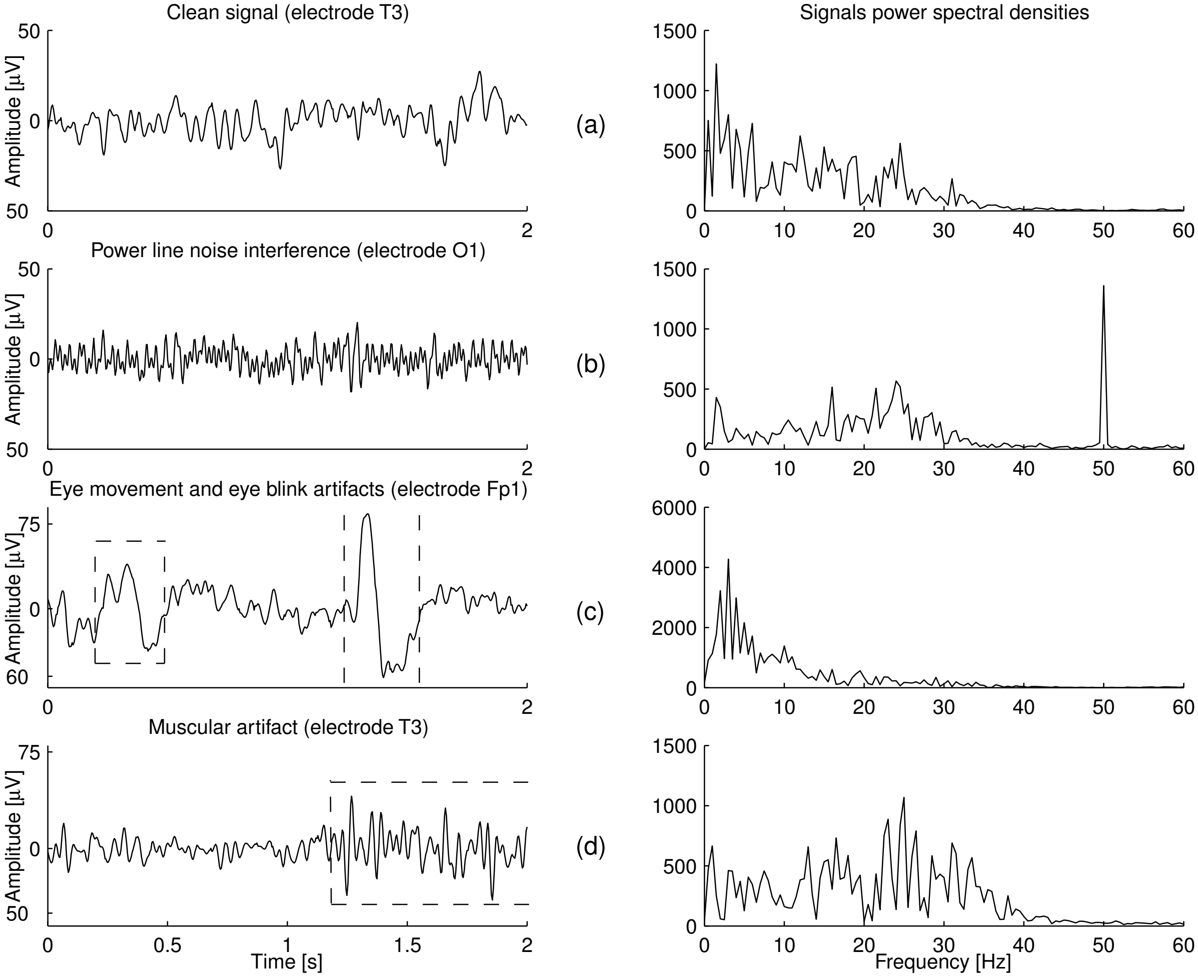

Figure 3.3. EEG signals perturbed by noise and artifacts and their corresponding power spectral densities (PSD) (*a*): clean EEG signal recorded at electrode T3. (*b*): EEG signal, recorded at electrode O1, perturbed by power line noise. The corresponding PSD shows clearly the perturbation at 50 Hz. (*c*): Signal recorded at electrode Fp1 containing an eye movement (left) and eye blink artifacts (right). The corresponding PSD reveals a concentration of the power in the theta band (4-8 Hz). (*d*): Signal recorded at electrode T3, containing a muscular movement artifact. The corresponding PSD shows that the power is concentrated in the beta band (13 to 30 Hz).

As mentioned in Section 3.2.3, the raw EEG-trial is first re-referenced with respect to the average of the EEG channels. In addition, the time average of every EEG channel is subtracted from the corresponding EEG channel. Therefore, the following relations hold.

$$\begin{aligned} \sum_{n=0}^{\mathrm{N_{spt}}-1} s_m(n) = 0 \qquad & m = 1, \ldots, \mathrm{N}_e \\ \sum_{m=1}^{\mathrm{N}_e} s_m(n) = 0 \qquad & n = 0, \ldots, \mathrm{N_{spt}} - 1 \end{aligned}$$

3.4.1 Power line noise filtering

The power line noise is concentrated around a single frequency (50 Hz in Europe) that falls beyond the EEG band. Therefore, it can be filtered using a notch filter [74] which highly attenuates a single frequency while leaving nearby frequencies relatively unchanged. The digital notch filter z-transform is given by [67, 74]:

$$H_{\mathrm{n}}(z) = \frac{1 + a_2 - 2a_1 z^{-1} + (1 + a_2)\, z^{-2}}{1 - a_1 z^{-1} + a_2 z^{-2}} \tag{3.1}$$

where

$$a_1 = \frac{2\cos\left(\frac{2\pi \mathrm{f_n}}{\mathrm{f_s}}\right)}{1 + \tan\left(\frac{\pi\beta_{\mathrm{n}}}{\mathrm{f_s}}\right)} \qquad a_2 = \frac{1 - \tan\left(\frac{\pi\beta_{\mathrm{n}}}{\mathrm{f_s}}\right)}{1 + \tan\left(\frac{\pi\beta_{\mathrm{n}}}{\mathrm{f_s}}\right)}$$

$\mathrm{f_n}$ is the notch frequency at which there is no transmission through the filter, and $\mathrm{f_s}$ is the sampling frequency. Within the frequency band centered at $\mathrm{f_n}$ and of width β_{n} (3-dB band) all signal components are attenuated by more than 3 dB. The smaller β_{n} the lower the attenuation of the notch frequency (see Fig . 3.4).

To determine the tradeoff between the width of the 3-dB band and the attenuation of the notch frequency, we estimate the power line noise level by measuring the signals coming from the electrodes before the conducting gel was applied. Depending on this level we select the adequate value of β_{n} using the graph depicted in Figure 3.4b.

If no artifact is detected in the raw EEG-trial $\tilde{S}$, the rows of the preprocessed EEG-trial S (that is sent to the feature extraction module) are obtained through the difference equation (which is obtained directly from Eq. 3.1):

$$s_m(n) - a_1 s_m(n-1) + a_2 s_m(n-2) = (1 + a_2)\, \tilde{s}_m(n) - 2a_1 \tilde{s}_m(n-1) + (1 + a_2)\, \tilde{s}_m(n-2) \tag{3.2}$$

for $m = 1, \ldots, \mathrm{N}_e$.

3.4.2 Artifact detection

The presence of eye movements, eye blinks and muscular artifacts in EEG signals can be easily detected from simple observation (Fig. 3.3). As a matter of fact, each type of artifact has characteristics in time and frequency that make it distinguishable from regular EEG.

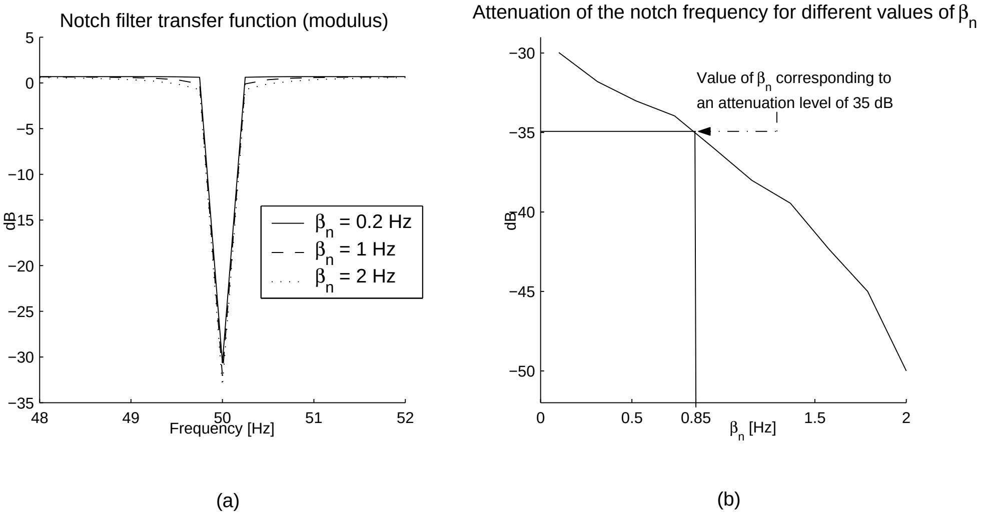

Figure 3.4. Notch filter characteristics. (*a*): Modulus of the notch filter (centered at the power line frequency, i.e. 50 Hz) transfer function. (*b*): The attenuation of the notch frequency increases with the width of the 3-dB band, β_{n}. The power line noise should be estimated in order to select the adequate value of β_{n}.

Ocular artifacts have large amplitudes, their spectral content is mainly concentrated in the theta band and are more prominent at frontal pole electrodes, i.e. Fp1 and Fp2. As it can be seen in Fig. 3.5, the time-frequency representation of a signal containing a series of ocular artifacts exhibits an abnormal concentration of the power in the theta band when ocular artifacts appear.

Muscular artifacts have amplitudes in the order of that of regular EEG but their spectral content is concentrated in the beta band. These artifacts are more noticeable in central temporal and parietal electrodes, i.e. electrodes T3, T4, T5, P3, P4 and T6 [164]. As depicted in Fig. 3.6, the time-frequency representation of a signal containing a muscular artifact reveals the presence of the artifact by exhibiting an abnormal concentration of the power in the beta band.

Artifacts can be considered as singular events in the time-frequency plane that appear randomly in EEG signals. To detect the presence of artifacts in an EEG-trial we divide it into one-second long segments (that overlap by 500 milliseconds) and check if an artifact is present in any of the segments. For instance, if the EEG-trial is 1500 milliseconds long, two segments are considered, namely from zero to 1000 milliseconds and from 500 to 1500 milliseconds.

The detection of an artifact in a one-second long segment (we call it artifact detection block ADB) is based on the following two facts. First, an ocular artifact implies that the power spectral densities of the signals at electrodes Fp1 and Fp2 are concentrated in the theta band and second, a muscular artifact at a given electrode makes its power spectral density concentrated in the beta band.

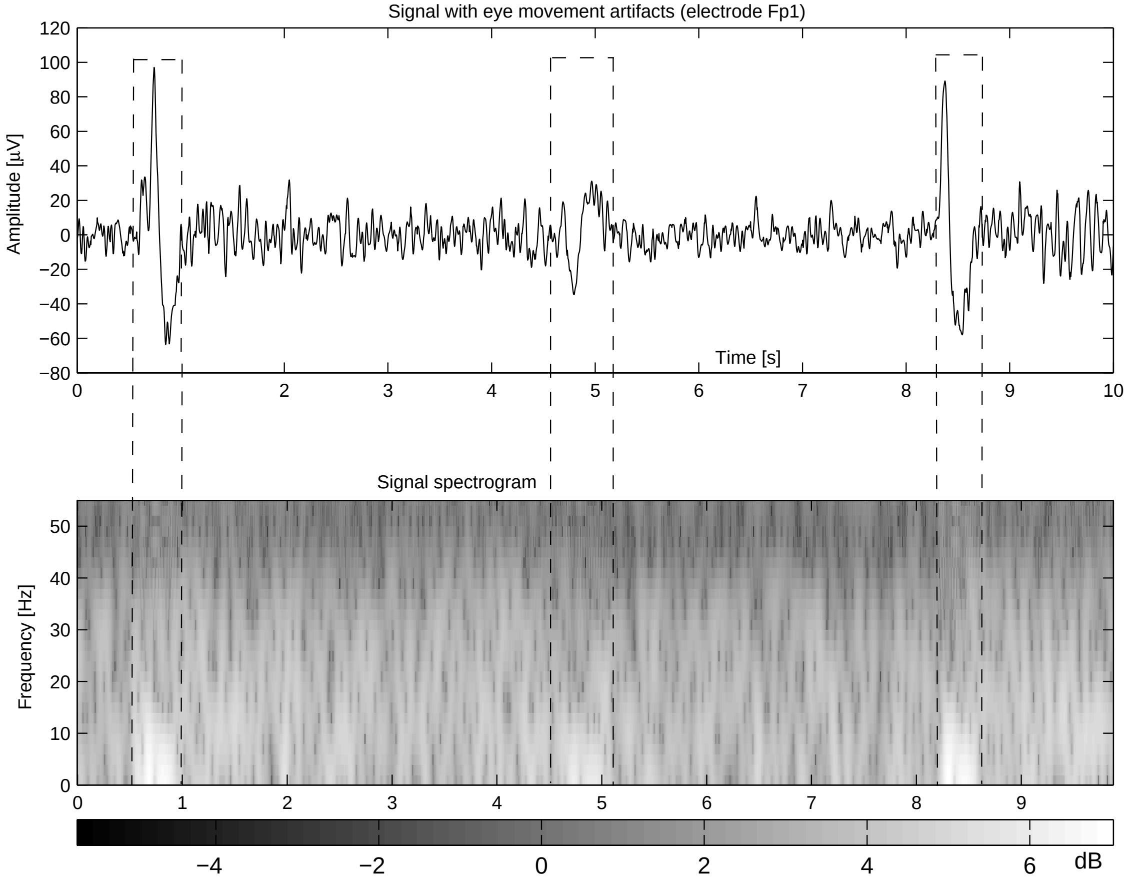

Figure 3.5. *Top*: Signal at electrode Fp1 containing three ocular artifacts delimited by the dashed lines. There is a considerable difference of amplitudes between the first and third artifact and the clean part of the signal. However, the amplitudes present in the second artifact are in the range of that of the clean part. A simple threshold on the signal amplitude is therefore insufficient to reliably detect the ocular artifacts. *Bottom*: Time-frequency representation of the signal. The times at which the ocular artifacts appear are characterized by a concentration of the signal power in the theta band. Thus, the frequency domain constitutes a good candidate to host the detection of ocular artifacts. Furthermore, it is important to note that an ocular artifact generally implies a strong correlation between the signals recorded at electrodes Fp1 and Fp2. Therefore, we take into account the frequency content of both electrodes in the detection procedure.

The time-frequency representation was obtained using the short term Fourier transform [30] which breaks the signal into chunks (which usually overlap each other) and computes the Fourier transform of each chunk.

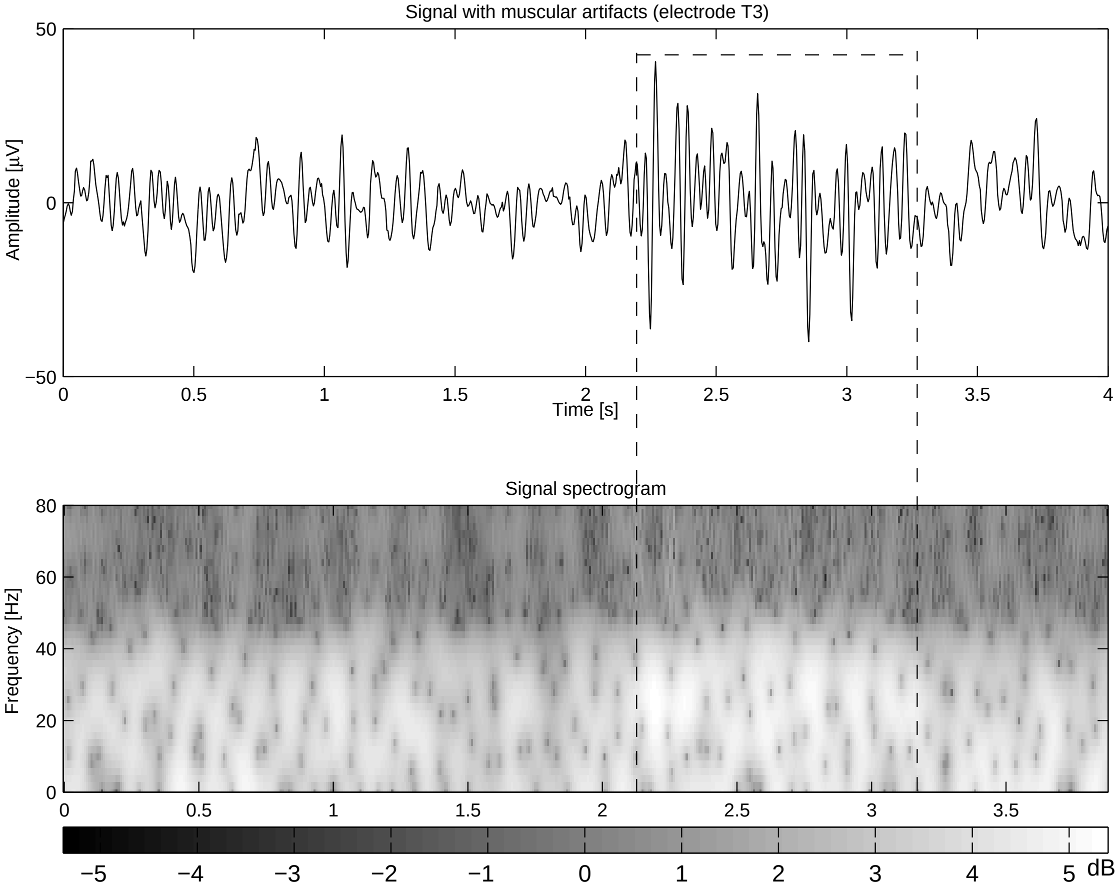

Figure 3.6. *Top*: Signal at electrode T3 containing a muscular artifact which is delimited by the dashed line. The difference between the signal amplitudes in the clean part and those in the muscular artifact is not as important as in the case of ocular artifacts. *Bottom*: Time-frequency representation of the signal. The signal power is concentrated in the beta band at the periods in which the artifact is present. As in the case of ocular artifacts, the frequency domain appears as more suitable than the time domain to host the detection of muscular artifacts. Muscular artifacts are more noticeable in temporal and parietal electrodes, i.e. electrodes T3, T4, T5, P3, P4 and T6. We thus, take into account the frequency content of the signal recorded at these electrodes in the detection procedure.

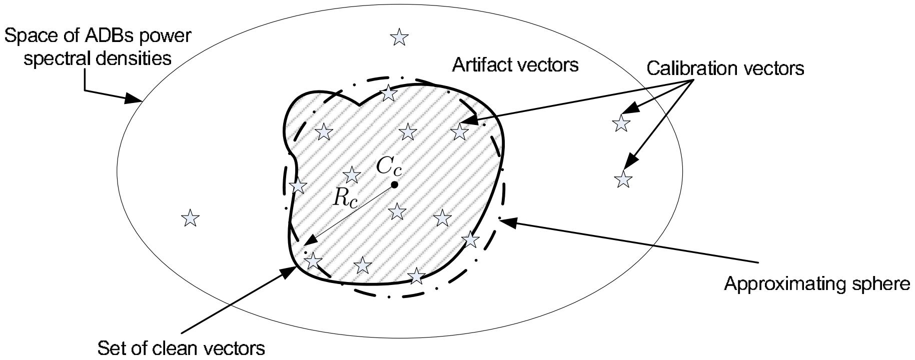

Figure 3.7. Set of clean ADBs in the space of their power spectral densities. The shape of this set depends on the subject and the environmental conditions at the time of recording, hence a calibration phase to adjust the artifact detection parameters is needed. The initial shape of the set of clean vectors is approximated by a sphere whose parameters are estimated using the calibration set.

From the above considerations it can be said that in the space of ADBs power spectral densities $\aleph$, the clean ADBs lie close to each other. This means that the set of clean ADBs lies in a small region of the space that is surrounded by ADBs containing artifacts (see Fig. 3.7). The shape of the set of clean ADBs depends on the subject and on the environmental conditions at the time of recording. Hence, the detection parameters need to be adapted at the beginning of each recording session (calibration phase).

For reasons of robustness and execution speed, the detection of ocular and muscular artifacts is performed separately. The space in which ocular artifacts are detected (ocular space) is composed of vectors containing the powers in 2 Hz wide bands from 2 to 40 Hz at electrodes Fp1 and Fp2. The space in which muscular artifacts are detected (muscular space) is composed of vectors containing the powers (in the same bands as in the ocular space) at electrodes T3, T4, T5, P3, P4 and T6. Therefore, the vectors are 38 and 114 dimensional in the ocular and muscular spaces respectively. The band powers are estimated using the Welch method, presented in Chapter 4 (Section 4.3).

The detection procedure is the same for both types of artifacts. Only its parameters need to be adapted to each artifact type during the calibration phase which lasts for a period varying from five to ten minutes. During the calibration, the subject is asked to blink his eyes and to execute slight head and hand movements, about 30 times each, at randomly chosen times. The resulting EEG is segmented into ADBs and the ocular and muscular vectors are computed for each ADB. At the end of the calibration phase two sets (one set per type of artifact) of vectors are available. In each of these sets we approximately know the percentage of vectors corresponding to ADBs containing artifacts (the exact percentage cannot be known since the subject could have generated additional artifacts).

In the following we present the general detection procedure which was adapted from the novelty detection framework presented in [147, 157].

Artifact detection procedure

Let $\aleph$ be the space of vectors computed from every possible ADB. We call artifact (clean) vector a vector resulting from an ADB that contains (does not contain) an artifact. The shape of the set of clean vectors is unknown. To effectively discriminate between clean and artifact vectors we seek for a criterion that evaluates whether or not a given vector belongs to the clean set.

The detection criterion is built using the calibration set $\aleph_{\text{cal}} = \{V_1, \ldots, V_{\text{N}_{\text{cal}}}\} \subset \aleph$ where N_{cal} is the number of ADBs recorded in the calibration phase. Since we ask the subject to produce a certain number of artifacts we approximately know the fraction of artifact vectors in the calibration set[1]. We denote as r_a the expected fraction of artifact vectors.

From the considerations in Section 3.4.2, we know that the clean vectors belonging to the calibration set must lie in a compact region of $\aleph$ (the assumption of compactness is reasonable since the clean vectors are close to each other with respect to their Euclidean distance). To start, we assume that this region can be approximated by a sphere of radius R_c centered at $C_c \in \aleph$ (see Fig. 3.7). The radius and the center are found by solving the optimization problem:

$$\underset{R_c, C_c, \xi_i}{\text{minimize}} \left(R_c^2 + \kappa \sum_{i=1}^{\text{N}_{\text{cal}}} \xi_i \right) \tag{3.3}$$

under constraints

$$\begin{aligned} \|V_i - C_c\|_{\aleph}^2 &\leqslant R_c^2 + \xi_i & (3.4) \\ \xi_i &\geqslant 0 & (3.5) \end{aligned}$$

$$\text{for } i = 1, \ldots, \text{N}_{\text{cal}}$$

where κ is a penalization constant whose value is linked to the fraction of artifact vectors (see Eq. 3.19) and $\|\cdot\|_{\aleph}$ is the Euclidean norm in the space $\aleph$. The positive slack variable ξ_i controls the position of V_i with respect to the approximating sphere. Indeed, if the value of ξ_i at the optimum is larger than zero then, V_i lies outside the approximating sphere and is therefore considered as an artifact.

To solve the optimization problem (3.3) under constraints (3.4) and (3.5), one introduces positive Lagrange multipliers $\mu_1, \ldots, \mu_{\text{N}_{\text{cal}}}, \gamma_1, \ldots, \gamma_{\text{N}_{\text{cal}}}$ to obtain the primal Lagrangian [130]:

$$\text{L}_g = R_c^2 + \kappa \sum_{i=1}^{\text{N}_{\text{cal}}} \xi_i - \sum_{i=1}^{\text{N}_{\text{cal}}} \gamma_i \left(R_c^2 + \xi_i - \|V_i - C_c\|_{\aleph}^2 \right) - \sum_{n=1}^{\text{N}_{\text{cal}}} \mu_i \xi_i \tag{3.6}$$

The primal Lagrangian should be minimized with respect to the primal variables, R_c, ξ_i, C_c and maximized with respect to the dual ones, γ_i, μ_i. Taking derivatives of L_g with respect

[1] Such fraction is only approximately known since the subject could have produced more artifacts

to the primal variables R_c, ξ_i, C_c and setting them to zero leads to the following results.

$$\partial_{R_c}\mathrm{L}_g = 0 \quad \Rightarrow \quad \sum_{i=1}^{\mathrm{N_{cal}}} \gamma_i = 1 \tag{3.7}$$

$$\partial_{C_c}\mathrm{L}_g = 0 \quad \Rightarrow \quad \sum_{i=1}^{\mathrm{N_{cal}}} \gamma_i V_i = C_c \tag{3.8}$$

$$\partial_{\xi_i}\mathrm{L}_g = 0 \quad \Rightarrow \quad \gamma_i + \mu_i = \kappa \tag{3.9}$$

By replacing (3.7), (3.8), and (3.9) in (3.6), one obtains the dual optimization problem:

$$\underset{\gamma_1,\dots,\gamma_{\mathrm{N_{cal}}}}{\text{maximize}} \left(\sum_{i=1}^{\mathrm{N_{cal}}} \gamma_i \left\langle V_i, V_i \right\rangle_{\aleph} - \sum_{i1,i2=1}^{\mathrm{N_{cal}}} \gamma_{i1}\gamma_{i2} \left\langle V_{i1}, V_{i2} \right\rangle_{\aleph} \right) \tag{3.10}$$

under the following constraints

$$0 \leqslant \gamma_i \leqslant \kappa \quad ; \quad i = 1, \dots, \mathrm{N_{cal}} \tag{3.11}$$

where $\left\langle V_{i1}, V_{i2} \right\rangle_{\aleph}$ is the inner (scalar) product of V_{i1}, V_{i2}. The dual optimization problem can be easily solved using standard quadratic optimization techniques [168]. By abuse of notation, we continue to write γ_i, μ_i, ξ_i for the values, at the optimum, of these parameters. Thus, the center and the radius of the approximating sphere are given by:

$$C_c = \sum_{i=1}^{\mathrm{N_{cal}}} \gamma_i V_i \tag{3.12}$$

$$\begin{aligned} R_c^2 &= \left\| V_{\hat{i}} - C_c \right\|_{\aleph}^2 \\ &= \left\langle V_i, V_i \right\rangle_{\aleph} - 2 \sum_{i=1}^{\mathrm{N_{cal}}} \gamma_i \left\langle V_{\hat{i}}, V_i \right\rangle_{\aleph} + \sum_{i1,i2=1}^{\mathrm{N_{cal}}} \gamma_{i1}\gamma_{i2} \left\langle V_{i1}, V_{i2} \right\rangle_{\aleph} \end{aligned} \tag{3.13}$$

where $V_{\hat{i}}$ is a vector that is on the approximating sphere, i.e. $0 < \gamma_{\hat{i}} < \kappa$ (see Fig. 3.8).

At the optimum, the Karush-Kuhn-Tucker conditions [95] imply that the following relations hold.

$$\gamma_i \left(R_c^2 + \xi_i - \| V_i - C_c \|_{\aleph}^2 \right) = 0 \tag{3.14}$$

$$\mu_i \xi_i = 0 \tag{3.15}$$

The position of V_i with respect to the approximating sphere depends on the value of γ_i. Three possibilities exist:

- If $\gamma_i = 0$, V_i is inside or on the approximating sphere. Therefore, V_i is considered as a clean vector (see Fig. 3.8a).
- If $0 < \gamma_i < \kappa$, V_i is on the approximating sphere. V_i is still considered as a clean vector (see Fig. 3.8b).

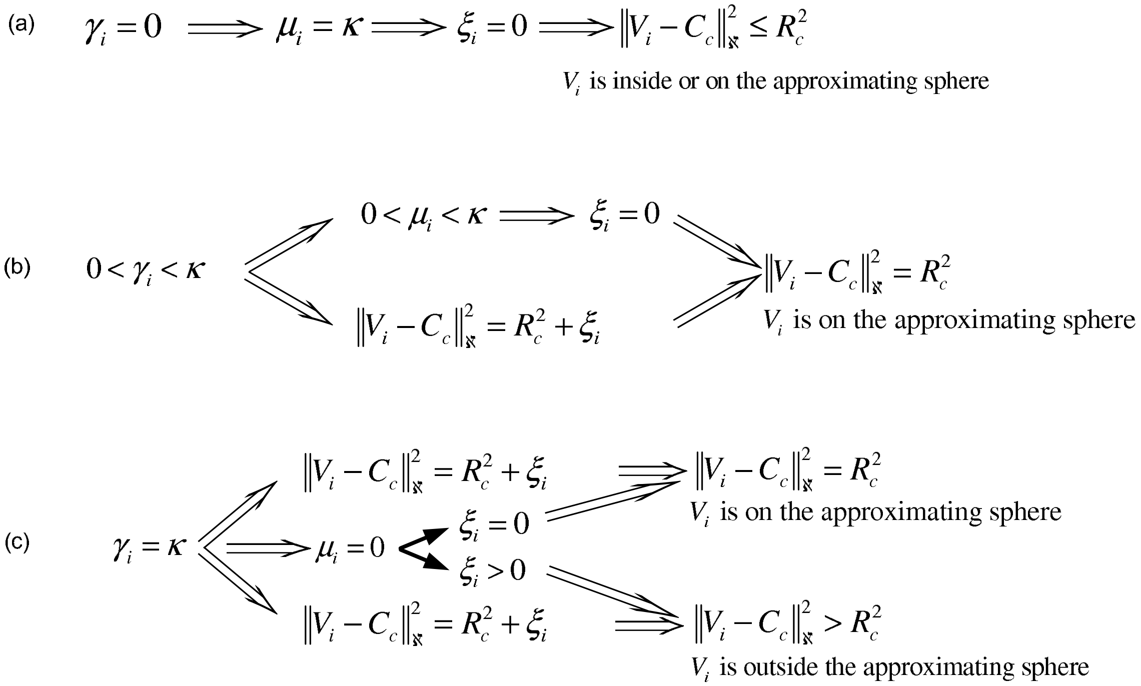
(a)
$\gamma_i = 0 \Longrightarrow \mu_i = \kappa \Longrightarrow \xi_i = 0 \Longrightarrow \|V_i - C_c\|_{\aleph}^2 \leq R_c^2$
V_i is inside or on the approximating sphere
(b)
$0 < \gamma_i < \kappa$
$0 < \mu_i < \kappa \Longrightarrow \xi_i = 0$
$\|V_i - C_c\|_{\aleph}^2 = R_c^2 + \xi_i$
$\|V_i - C_c\|_{\aleph}^2 = R_c^2$
V_i is on the approximating sphere
(c)
$\gamma_i = \kappa$
$\|V_i - C_c\|_{\aleph}^2 = R_c^2 + \xi_i$
$\mu_i = 0$
$\xi_i = 0$
$\xi_i > 0$
$\|V_i - C_c\|_{\aleph}^2 = R_c^2 + \xi_i$
$\|V_i - C_c\|_{\aleph}^2 = R_c^2$
V_i is on the approximating sphere
$\|V_i - C_c\|_{\aleph}^2 > R_c^2$
V_i is outside the approximating sphere

Figure 3.8. Position of V_i with respect to the approximating sphere for different values of γ_i. For values of γ_i in $[0, \kappa[$, the corresponding V_i is considered as a clean vector. Conversely, if $\gamma_i = \kappa$ the corresponding V_i is considered as a clean or artifact vector depending on the value of the slack variable ξ_i.

- If $\gamma_i = \kappa$, the position of V_i depends on the value of ξ_i. If $\xi_i > 0$, V_i is outside the approximating sphere and then considered as an artifact vector. Conversely, $\xi_i = 0$ implies that V_i is on the approximating sphere and hence, it is considered as a clean vector. (see Fig. 3.8c).

The constant κ controls the fraction of vectors in the calibration set $\aleph_{\text{cal}}$ that are considered as artifacts. To see this, we decompose (3.7) into the sum of the γ's corresponding to vectors that are inside (I), on (B) and outside (O) the approximating sphere:

$$\sum_{i \in \text{ I}} \gamma_i + \sum_{i \in \text{ B}} \gamma_i + \sum_{i \in \text{ O}} \gamma_i = 1 \tag{3.16}$$

The first term on the left vanishes, so:

$$\sum_{i \in \text{ B}} \gamma_i + \sum_{i \in \text{ O}} \gamma_i \quad = \quad 1 \tag{3.17}$$

$$\Rightarrow \sum_{i \in \text{ O}} \gamma_i \quad \leqslant \quad 1 \tag{3.18}$$

From Fig. 3.8 we know that if a vector V_i is outside the approximating sphere, the corresponding γ_i is equal to κ. Thus, we obtain the following inequality that links κ to the expected fraction of artifact vectors r_a.

$$\kappa \leqslant \frac{1}{r_a \text{N}_{\text{cal}}} \tag{3.19}$$

To decide whether a vector $\tilde{V}$, not belonging to the calibration set, is an artifact vector or not, we compute its square distance to the center of the approximating sphere:

$$\left\|\tilde{V} - C_c\right\|_{\aleph}^{2} = \left\langle \tilde{V}, \tilde{V} \right\rangle_{\aleph} - 2\sum_{i=1}^{\mathrm{N}_{\mathrm{cal}}} \gamma_i \left\langle \tilde{V}, V_i \right\rangle_{\aleph} + \sum_{i1,i2=1}^{\mathrm{N}_{\mathrm{cal}}} \gamma_{i1}\gamma_{i2} \left\langle V_{i1}, V_{i2} \right\rangle_{\aleph} \tag{3.20}$$

This distance is compared to R_c^2 to obtain the detection ratio:

$$\frac{\left\|\tilde{V} - C_c\right\|_{\aleph}^{2}}{R_c^2} \tag{3.21}$$

$\tilde{V}$ is considered as an artifact if the detection ratio is larger than 1 and as a clean vector otherwise.

It is worth mentioning that (3.20) depends only on those calibration vectors that are at the boundary or outside the approximating sphere (indeed, the vectors inside the approximating sphere have their corresponding γ equal to zero). Such vectors are usually called support vectors [157].

So far, we assumed that the shape of the set of clean vectors could be approximated by a sphere. This approximation permitted to obtain a simple detection criterion through the solution of a standard quadratic optimization problem.

However, there is no a priori reason that makes the sphere the preferred approximation shape for the set of clean vectors. In certain cases, especially when the clean set is non-convex the sphere approximation is clearly flawed. Thus, we need to consider more flexible shapes to approximate the clean set. This can be easily done by means of the ”Kernel trick” [1] which consists in replacing the inner products $\langle \cdot, \cdot \rangle_{\aleph}$ in the the detection procedure by a kernel function $K(\cdot, \cdot)$ that satisfies the Mercer conditions [1] (see Chapter 5, Section 5.2). One can show [169] that the latter amounts to project the space $\aleph$ into a high (possibly infinite) dimensional space $\mathscr{J}_{\aleph}$, through a map $\mathscr{J}$, such that $K(\cdot, \cdot)$ is the inner product in $\mathscr{J}_{\aleph}$. This means that the following relation holds.

$$K(V_1, V_2) = \langle \mathscr{J}(V_1), \mathscr{J}(V_2) \rangle_{\mathscr{J}_{\aleph}} \tag{3.22}$$

where $\mathscr{J}(V_i)$ is the image of V_i through the map $\mathscr{J}$.

The advantage of using kernels resides in the fact that a sphere in $\mathscr{J}_{\aleph}$ can represent a complex shape in the space $\aleph$. A kernel function that satisfy the Mercer conditions and permits to flexibly approximate the shape of clean vectors is the Gaussian kernel (see Chapter 5, Section 5.4):

$$K(V_1, V_2) = \exp\left(-\frac{\|V_1 - V_2\|_{\aleph}^{2}}{\sigma^2} \right) \tag{3.23}$$

where σ is the Gaussian kernel parameter. One can show that for fixed κ, the smaller σ the smaller the number of artifact vectors in the calibration set [157]. Because of the definition of the Gaussian kernel in terms of the ratio between the distance of its arguments and σ, we discuss the influence of this parameter by considering its normalized version, $\sigma_r = \frac{\sigma}{\Delta_m}$ where Δ_m is the minimum distance between two different calibration vectors.

From relation (3.19) linking κ to the expected fraction of artifact vectors in the calibration set, we can deduce that for fixed σ, the larger κ the smaller the fraction of artifact vectors in the calibration set. In geometrical terms we can think of κ as a factor limiting the generalized volume of the approximating region.

For the sake of visualization we illustrate the role of σ_r and κ in a 2D toy problem. In Fig. 3.9 we illustrate the influence of the Gaussian kernel parameter (for fixed κ) on the shape of the approximating region. In particular, a small σ_r makes the approximating region over-fit the data while a large σ makes the approximating region become a sphere. In Fig. 3.10 we illustrate the influence of κ (for fixed σ, which amounts to fix the shape of the approximating region) on the extent of the approximating region.

Thus, we can control the shape and volume of the approximating region through σ and κ respectively. The adequate choice of these parameters depends on the data. In Fig. 3.11 we report the fraction of artifact vectors as a function of σ_r for different values of κ (detection curves). As predicted by the relation (3.19), κ establishes an upper bound on the fraction of artifact vectors. This means that it is possible to fix κ by using the expected fraction of artifact vectors and then adjust σ to match the requirements in terms of detection sensibility.

3.4.3 Practical parameter setting

As mentioned before, we know the approximative number of artifacts that the subject produced during the calibration phase. The values of κ and σ are decided by the operator through the observation of the detection curves (as shown in Fig. 3.11) for four values of κ, namely: $\frac{1}{4r_a \mathrm{N}_{\mathrm{cal}}}$, $\frac{1}{2r_a \mathrm{N}_{\mathrm{cal}}}$, $\frac{3}{4r_a \mathrm{N}_{\mathrm{cal}}}$ and $\frac{1}{r_a \mathrm{N}_{\mathrm{cal}}}$ and σ_r ranging from 1 to 50 (the range for σ_r can be adjusted by the operator). For a given choice, the operator can visually check those ADBs identified as containing artifacts.

Thus, the operator can effectively control the sensibility of the artifact detection based on his own experience. In Figures 3.12 and 3.13, the detection procedure for ocular and muscular artifacts respectively is illustrated. As it can be seen, too small a value of σ_r diminishes the detection sensibility. On the other hand, large values of σ_r prompt the rejection of clean data.

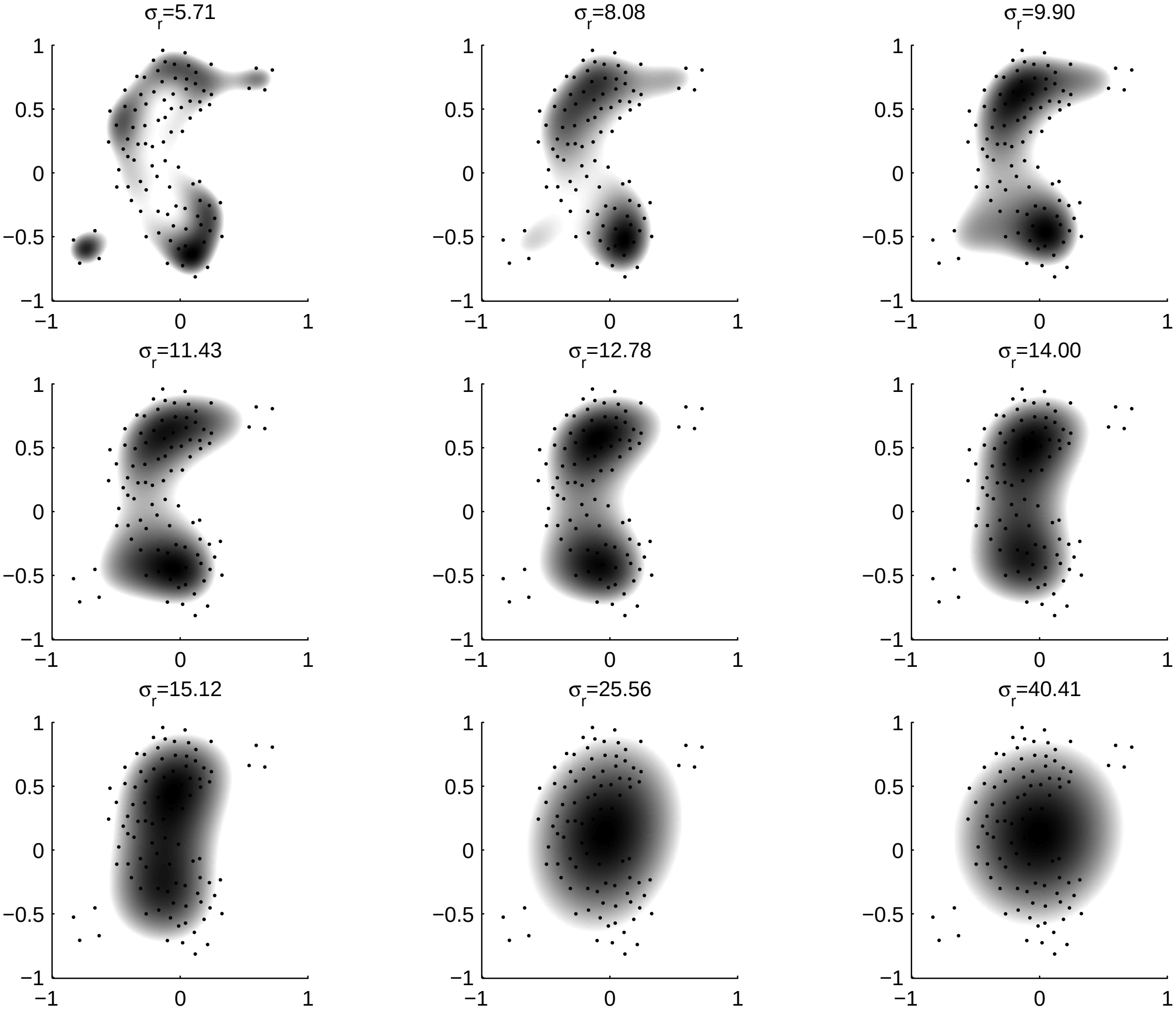

Figure 3.9. Influence of the Gaussian kernel parameter (for fixed $\kappa = 0.05$) on the shape of the approximating region. The data are represented by the black dots. The darker the region the smaller the detection ratio computed using (3.21). The white zone surrounding the approximating region corresponds to the region in which the rejected data lie.
As it can be seen, the shape of the approximating region is effectively controlled by σ_r. In particular, the smaller σ_r the smaller the fraction of rejected data (rejected data corresponds to artifact vectors in the framework of artifact detection). As σ_r increases the shape of the approximating region becomes more spherical.

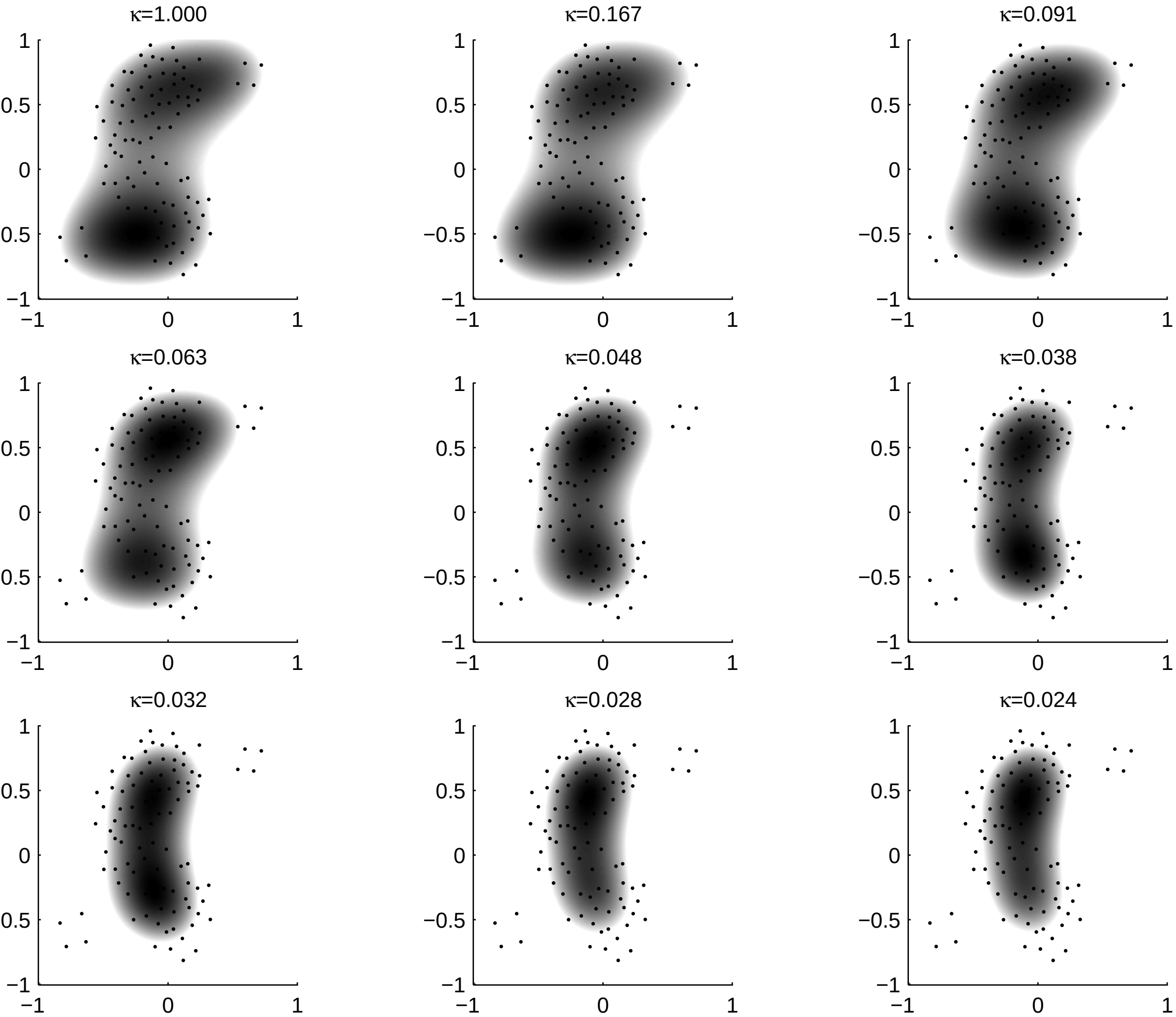

Figure 3.10. Influence of the parameter κ (for fixed $\sigma_r = 15$) on the volume of the approximating region in the context of the toy problem considered in Fig. 3.9. Once the shape of the approximating region is fixed by σ_r, its volume is limited by κ. Thus, the larger κ the smaller the fraction of rejected data (or the larger the allowed volume of the approximating region).

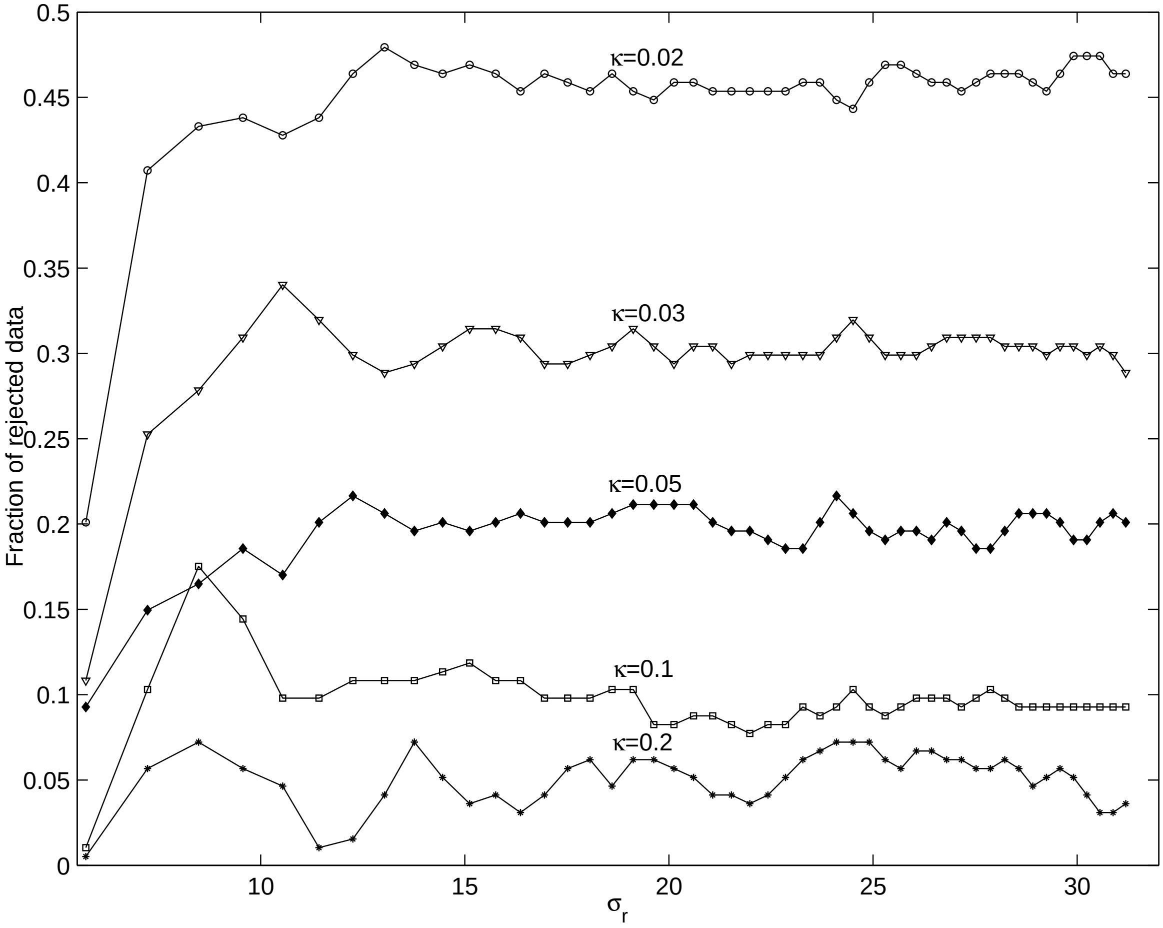

Figure 3.11. The parameters σ_r and κ allow us to control the shape and the volume of the approximating region respectively. The adequate selection of these parameters is data dependent. The detection curves (for the toy problem of Figs. 3.9 and 3.10) depicted here show the joint influence of the detection parameters on the fraction of rejected data. The limiting role of κ becomes evident on the detection curves.

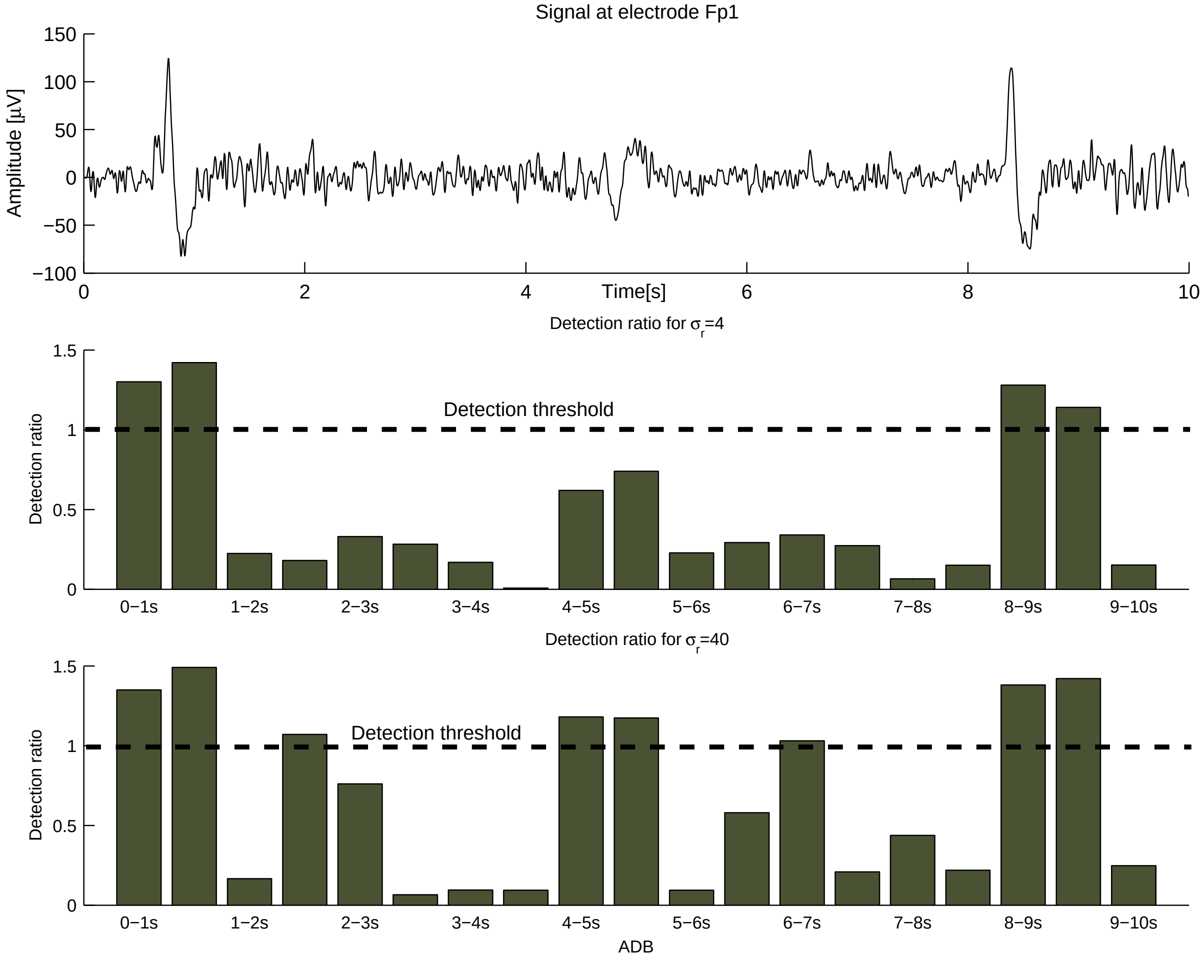

Figure 3.12. Detection of ocular artifacts. *Top*: Signal recorded at electrode Fp1 containing ocular artifacts. *Middle*: Detection ratio ($\sigma_r = 4$ and $\kappa = \frac{3}{4r_a \mathrm{N_{cal}}}$) for the ADBs (one-second long segments overlapped by half a second) of the signal on the top. *Bottom*: Detection ratio for $\sigma_r = 40$ and $\kappa = \frac{3}{4r_a \mathrm{N_{cal}}}$. In this example, the algorithm for $\sigma_r = 4$ fails to detect the artifact in the middle. On the opposite, $\sigma_r = 40$ leads to false artifact detections.

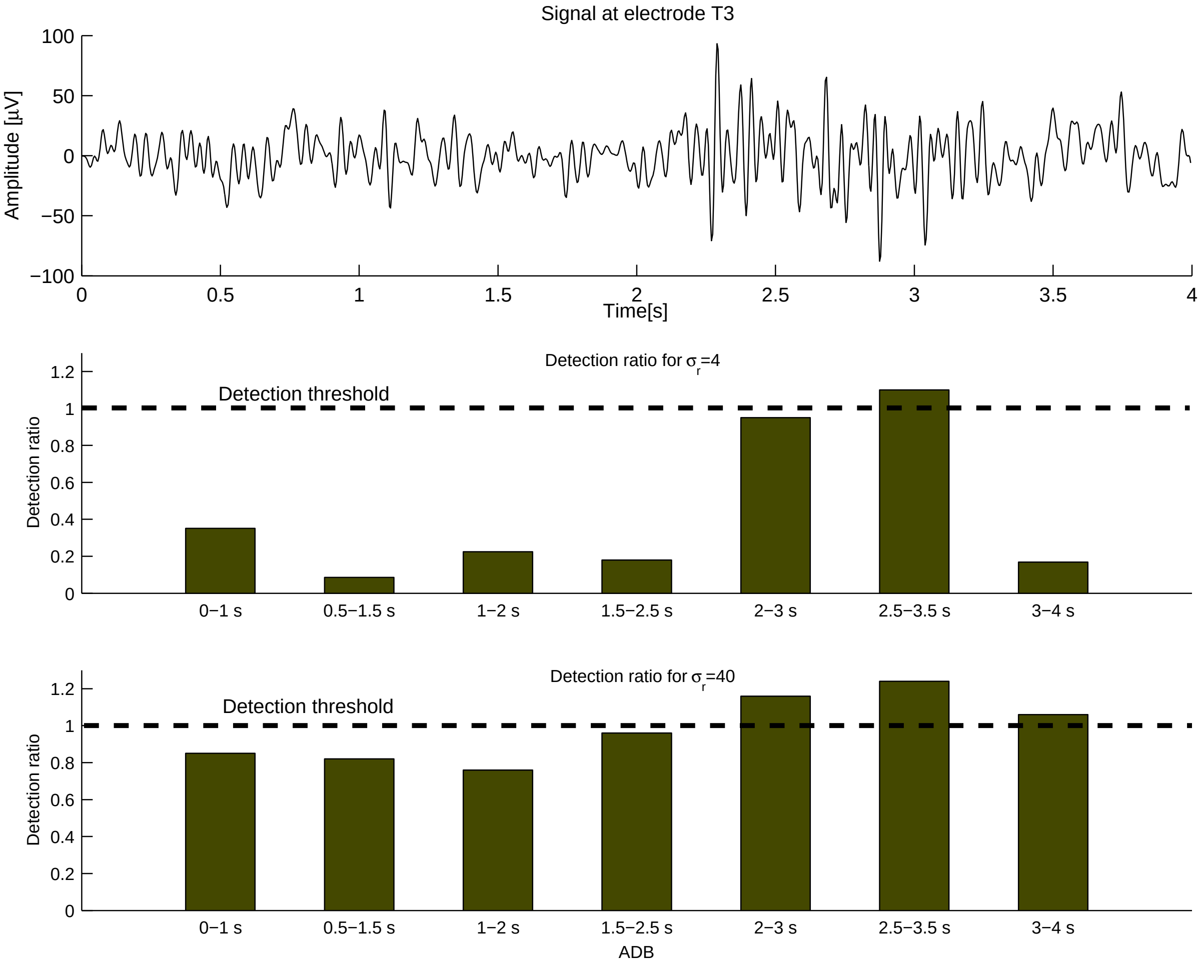

Figure 3.13. Detection of muscular artifacts. *Top*: Signal recorded at electrode T3 containing muscular artifacts. *Middle*: Detection ratio ($\sigma_r = 4$ and $\kappa = \frac{3}{4r_a \mathrm{N_{cal}}}$) for the ADBs (one-second long segments overlapped by half a second) of the signal on the top. *Bottom*: Detection ratio for $\sigma_r = 40$ and $\kappa = \frac{3}{4r_a \mathrm{N_{cal}}}$.

3.5 Summary

In this chapter we have presented an overview of electroencephalography concepts and the details of the EEG acquisition method used in this thesis. The electrode Cz was selected as physical reference and in a posterior step the signals were re-referenced with respect to their average.

The influence of external noise is attenuated by controlling the recording environment. As power line noise is almost unavoidable since most of the equipment used in the experiences need to be connected to the power grid, we filter it by using a notch filter centered at the power line frequency. The tradeoff between the degree of attenuation and the width of the filtered band is resolved by estimating the level of power noise.

The presence of ocular and muscular artifacts makes it difficult to extract useful information that can be exploited by the BCI. Furthermore, they can lead to erroneous conclusions about the control performance of a subject. To prevent these issues we discard the data containing artifacts. To implement the rejection criterion we considered the frequency domain characteristics of artifacts which make them easily identifiable from regular EEG.

By using an adapted version of the novelty detection algorithm presented in [157] we can easily control the artifact detection sensibility through two parameters that can be set by the operator in an interactive way.

In Fig. 3.14 we summarize the function of the acquisition and preprocessing modules within the BCI system. The raw EEG-trials delivered by the acquisition module are re-referenced and their power line noise is filtered. If the EEG-trial contains muscular or ocular artifacts the BCI does not attempt to generate an action command from such a trial. Instead, it notifies the subject by executing predefined actions depending on whether ocular or muscular artifacts were detected.

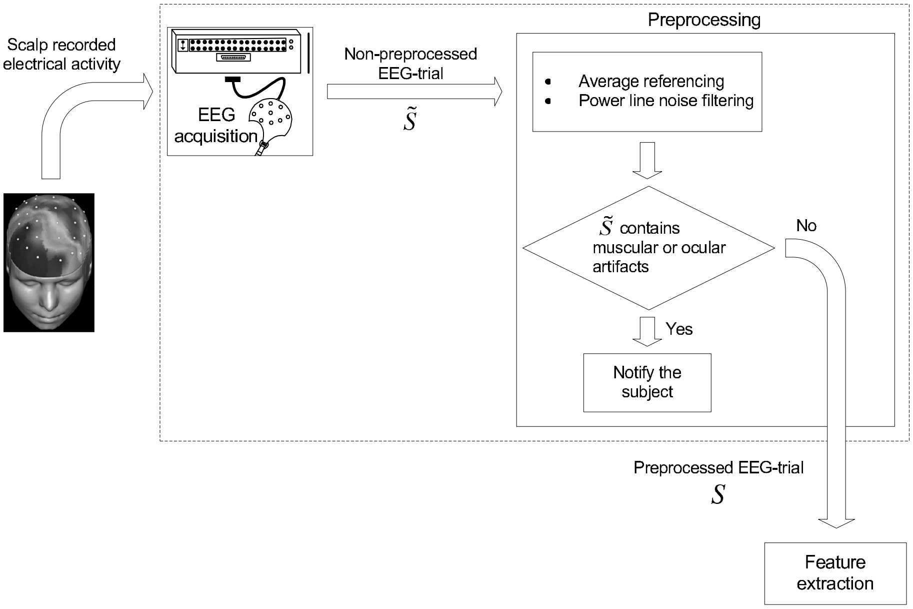
Scalp recorded electrical activity
EEG acquisition
Non-preprocessed EEG-trial
$\tilde{S}$
Preprocessing
Average referencing
Power line noise filtering
$\tilde{S}$ contains muscular or ocular artifacts
No
Yes
Notify the subject
Preprocessed EEG-trial
S
Feature extraction

Figure 3.14. Role of the EEG acquisition and preprocessing modules. The non-preprocessed EEG-trials delivered by the acquisition module are re-referenced and their power line noise is filtered. If the EEG-trial contains muscular or ocular artifacts the BCI does not attempt to generate an action command from such a trial. Instead, it notifies the subject by executing predefined actions depending on whether ocular or muscular artifacts were detected.

4 Feature extraction

"We are always paid for our suspicion by finding what we suspect." *Henry David Thoreau*

4.1 Introduction

In the previous chapter we presented the preprocessing procedure through which the external noise is removed and the EEG-trials containing artifacts are detected and discarded. In this chapter we focus on the estimation of statistical measurements (or features) from the perturbation free EEG-trials delivered by the preprocessing module. The features computed on a given EEG-trial are grouped into a vector called feature vector that is sent to the pattern recognition module which evaluates the likelihoods that the EEG-trial (represented by its feature vector) was produced during the execution of the MAs in the controlling set (see Fig. 4.1).

Features need to reflect properties of EEG that are relevant for the recognition of MAs. The choice of adequate features to characterize EEG has been the object of active research during the last decades [115, 174]. As a matter of fact, the techniques used to analyze EEG evolved in parallel with the development of novel signal processing concepts. In particular, the analysis of the generalized interaction (in time, frequency, and phase) between EEG channels has emerged as a tool to study EEG data [4, 43].

A complete analysis that takes into account time, frequency and phase would result in a very large number of features (Section 4.2.2) and consequently a high dimensional feature vector. Because of the particular requirements of BCI applications, according to which a continuous adaptation of the recognition models and a reasonable training time are required (Chapter 5), high dimensional feature vectors are clearly non-suitable.

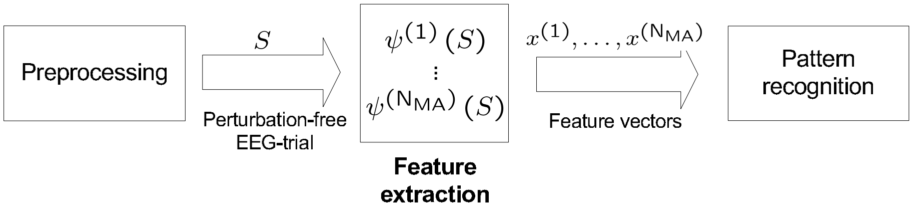
Preprocessing
S
Perturbation-free EEG-trial
$\psi^{(1)}(S)$
$\vdots$
$\psi^{(\mathrm{N_{MA}})}(S)$
Feature extraction
$x^{(1)},\dots,x^{(\mathrm{N_{MA}})}$
Feature vectors
Pattern recognition

Figure 4.1. The feature extraction module is in charge of computing statistical properties (features) on an EEG-trial (free of artifacts) S delivered by the preprocessing module. The mappings associated with each MA in the controlling set, $\psi^{(1)}(S),\dots,\psi^{(\mathrm{N_{MA}})}(S)$ are computed ($x^{(k)} = \psi^{(k)}(S)$) and sent to the pattern recognition module which evaluates the likelihoods that S was generated during the performance of each MA.

By assuming certain hypotheses on the properties of EEG, less features are required to characterize an EEG-trial. In this chapter, we present such hypotheses and derive different mappings from the EEG-trial set into feature spaces. A mapping is associated to a certain number of hypotheses that are used to define it. As presented in Chapter 6, depending on the subject a single mapping is not sufficient for the recognition of all the MAs in the controlling set. Therefore, the best mapping to recognize each MA has to be chosen. Such choice is carried out according to the optimality criterion presented in Chapter 6. The mapping associated to MA$_k$ is denoted as $\psi^{(k)}$ (see Fig. 4.1).

This chapter is organized as follows. First, the general time-frequency analysis of stochastic signals is considered. Second, the hypotheses that permit to obtain the mappings are discussed and finally the resulting mappings are presented.

4.2 An overview of time-frequency analysis for stochastic signals

Being composed of the univariate signals (or univariate components) recorded at each electrode, EEG signals can be modelled as realizations of a multivariate stochastic process. Time-frequency (TF) analysis of multivariate signals aims at describing the time variations of their intra and inter component spectral properties by means of a time-frequency representation (TFR). TF analysis is particularly useful for non-stationary signals (e.g. EEG) for which an analysis restricted to time or frequency is not sufficient to describe their dynamics.

TFRs are broadly categorized by their inherent mathematical structure as linear or quadratic. As we consider second order statistical moments to describe the EEG signals, we concentrate on quadratic TFRs. The latter may be further subdivided as power or correlation based, depending on whether they seek to combine power or correlation analysis in the TF plane.

It is worth noting that we focus on those concepts of TF analysis that are useful for our purpose, namely to extract relevant features to recognize MAs from EEG. More complete descriptions of TF analysis can be found in [30] and [49].

For convenience of exposition we first consider the TF analysis of univariate stochastic signals and then, generalize the obtained results to the multivariate case. As the signals that we consider are discrete, we develop our results in the discrete framework.

4.2.1 Time-frequency analysis of univariate stochastic signals

Let $\mathbf{s}$ be an univariate stochastic signal of length N, composed of the random variables: $\{\mathbf{s}(n) | \, n = 0, \ldots, N-1\}$ where n is the time index.

Time domain analysis

The properties of $\mathbf{s}$ can be described in time using first and second order moments computed on the random variables $\mathbf{s}(n)$. These moments are:

- The expectation of $\mathbf{s}(n)$: $\mathrm{E}_{p(\mathbf{s}(n))}\left[\mathbf{s}(n)\right]$, where $p\left(\mathbf{s}(n)\right)$ is the probability density function associated with $\mathbf{s}(n)$.
- The expectation of the product $\mathbf{s}(n_1)\mathbf{s}(n_2)$: $\mathrm{E}_{p(\mathbf{s}(n_1),\mathbf{s}(n_2))}\left[\mathbf{s}(n_1)\mathbf{s}(n_2)\right]$, where $p\left(\mathbf{s}(n_1), \mathbf{s}(n_2)\right)$ is the joint probability density function of $\mathbf{s}(n_1)$ and $\mathbf{s}(n_2)$.

Expectations taken with respect to the probability density functions associated with the random variables $\mathbf{s}(n)$ are called ensemble averages. For convenience of notation, we denote as $\mathrm{E}_{\mathbf{s}}\left[\cdot\right]$ any ensemble average over $\mathbf{s}$.

The signal power $\mathrm{P}_{\mathbf{s}}$ and time autocorrelation function $\mathrm{R}_{\mathbf{s}}(n,\tau)$ are defined as:

$$\mathrm{P}_{\mathbf{s}} = \frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\sum_{n=0}^{N-1} |\mathbf{s}(n)|^2\right] \tag{4.1}$$

$$\mathrm{R}_{\mathbf{s}}(n,\tau) = \mathrm{E}_{\mathbf{s}}\left[\mathbf{s}^*(n-\tau)\mathbf{s}(n)\right] \tag{4.2}$$

where * stands for the complex conjugate operator[1], n is the time at which $\mathrm{R}_{\mathbf{s}}$ is computed, and $\tau \in \{-N+1, \ldots, N-1\}$ is the time lag. Since $\mathrm{P}_{\mathbf{s}}$ can be written as an average over time of: $\mathrm{E}_{\mathbf{s}}\left[|\mathbf{s}(n)|^2\right] = \mathrm{R}_{\mathbf{s}}(n,0)$, it follows that $\mathrm{E}_{\mathbf{s}}\left[|\mathbf{s}(n)|^2\right]$ can be considered as the signal power density in the time domain (or power time density PTD). Thus, we can use $\mathrm{E}_{\mathbf{s}}\left[|\mathbf{s}(n)|^2\right]$ to compute the average, over the PTD of any time function $\gamma(n)$ as follows.

$$\langle\gamma(n)\rangle_{\mathrm{PTD}} = \frac{1}{N}\sum_{n=0}^{N-1}\gamma(n)\mathrm{E}_{\mathbf{s}}\left[|\mathbf{s}(n)|^2\right] \tag{4.3}$$

[1] It is worth nothing that even though we consider real signals, the complex conjugate in the definition of $\mathrm{R}_{\mathbf{s}}(n,\tau)$ facilitates further developments

Frequency domain analysis

The frequency properties of $\mathbf{s}$ can be examined using its discrete Fourier transform defined as:

$$\hat{\mathbf{s}}(\vartheta) = \sum_{n=0}^{N-1} \mathbf{s}(n) \exp\left(-j\frac{2\pi n\vartheta}{N}\right) \tag{4.4}$$

where ϑ is the frequency index. The correspondence between the frequency index ϑ and the actual frequency f (in Hz) is given by [114]:

$$\mathrm{f} = \frac{\mathrm{f_s}\vartheta}{N} \qquad \vartheta = 0, \ldots, \frac{N}{2} \tag{4.5}$$

where $\mathrm{f_s}$ is the sampling frequency. The values: $\hat{\mathbf{s}}\left(\frac{N}{2}\right), \ldots, \hat{\mathbf{s}}(N-1)$ correspond to the negative part of the spectrum of $\mathbf{s}$ [135]. In fact, one can easily verify that:

$$\hat{\mathbf{s}}(\vartheta) = \hat{s}^*(N-\vartheta) \qquad \vartheta = 1, \ldots, \frac{N}{2} \tag{4.6}$$

Using the discrete inverse Fourier transform, $\mathbf{s}$ can be obtained from $\hat{\mathbf{s}}$ as follows.

$$\mathbf{s}\,(n) = \frac{1}{N}\sum_{\vartheta=0}^{N-1} \hat{\mathbf{s}}(\vartheta) \exp\left(j\frac{2\pi n\vartheta}{N}\right) \tag{4.7}$$

This relation is easily verified by replacing $\hat{\mathbf{s}}(\vartheta)$ by its definition (4.4) and using the identities:

$$\frac{1}{N}\sum_{n=0}^{N-1} \exp\left(j\frac{2\pi n\vartheta}{N}\right) = \delta_d(\vartheta) \tag{4.8}$$

$$\sum_{\vartheta=0}^{N-1} g(\vartheta)\delta_d(\vartheta-\vartheta') = g(\vartheta') \qquad \vartheta' = 0, \ldots, N-1 \tag{4.9}$$

where $\delta_d\,(\cdot)$ is the digital delta function which is equal to one at zero and equal to zero elsewhere.

Similarly to the time domain, first and second order moments can be defined on the random variables $\hat{\mathbf{s}}(\vartheta)$. In particular, the frequency autocorrelation function can be defined as:

$$\mathcal{R}_{\mathbf{s}}\,(\vartheta, \upsilon) = \frac{1}{N}\mathrm{E}_{\hat{\mathbf{s}}}\left[\hat{\mathbf{s}}^*\,(\vartheta-\upsilon)\,\hat{\mathbf{s}}\,(\vartheta)\right] \tag{4.10}$$

where υ is the frequency lag and the normalization factor $\frac{1}{N}$ takes into account the Parseval identity (4.12). Note that the ensemble average in (4.10) is taken with respect to the joint probability density function: $p(\hat{\mathbf{s}}(\vartheta-\upsilon), \hat{\mathbf{s}}(\vartheta))$.

It is well known that if a new signal $\mathbf{s}'$ is obtained from $\mathbf{s}$ through an invertible function $\mathcal{F}$ then:

$$\mathrm{E}_{\mathbf{s}}\left[\mathcal{G}(\mathbf{s})\right] = \mathrm{E}_{\mathbf{s}'}\left[\mathcal{G}(\mathcal{F}^{-1}(\mathbf{s}'))\right] \tag{4.11}$$

where $\mathcal{G}(\cdot)$ is any function of $\mathbf{s}$. In the following, for brevity of notation we use $\mathrm{E}_{\mathbf{s}}[\cdot]$ to denote any ensemble average over $\mathbf{s}$ or any other signal obtained from $\mathbf{s}$.

Replacing $\mathbf{s}(n)$ by (4.7), in the power definition (4.1) yields:

$$\mathrm{P}_{\mathbf{s}} \quad = \quad \mathrm{E}_{\mathbf{s}}\left[\frac{1}{N^3}\sum_{\vartheta=0}^{N-1}\hat{\mathbf{s}}(\vartheta)\sum_{\vartheta_1=0}^{N-1}\hat{\mathbf{s}}^*(\vartheta_1)\sum_{n=0}^{N-1}\exp\left(j\frac{2\pi n\left(\vartheta-\vartheta_1\right)}{N}\right)\right]$$

using (4.8), (4.9), and (4.1), we get the Parseval identity:

$$\mathrm{P}_{\mathbf{s}} = \mathrm{E}_{\mathbf{s}}\left[\frac{1}{N^2}\sum_{\vartheta=0}^{N-1}\hat{\mathbf{s}}(\vartheta)\sum_{\vartheta_1=0}^{N-1}\hat{\mathbf{s}}^*(\vartheta_1)\delta_d(\vartheta-\vartheta_1)\right]$$

$$\mathrm{P}_{\mathbf{s}} = \mathrm{E}_{\mathbf{s}}\left[\frac{1}{N^2}\sum_{\vartheta=0}^{N-1}\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right] = \frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\sum_{n=0}^{N-1}\left|\mathbf{s}(n)\right|^2\right] \tag{4.12}$$

This result can be thought of as a power conservation relation between the time and frequency domains.

Since the signal power, according to (4.12), can be written as an average over ϑ of: $\frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right] = \mathcal{R}_{\mathbf{s}}(\vartheta, 0)$, it follows that $\frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right]$ can be considered as the signal power density in the frequency domain (or power spectrum density PSD). Thus, we can use $\frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right]$ to compute the average, over the PSD of any frequency function $g(\vartheta)$ as:

$$\langle g(\vartheta)\rangle_{\mathrm{PSD}} = \frac{1}{N^2}\sum_{\vartheta=0}^{N-1} g(\vartheta)\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right] \tag{4.13}$$

In particular, when $g\left(\vartheta\right) = \exp\left(j\frac{2\pi\tau\vartheta}{N}\right)$, one obtains the characteristic function of the power spectrum density (i.e. its inverse Fourier transform):

$$\left\langle \exp\left(j\frac{2\pi\tau\vartheta}{N}\right)\right\rangle_{\mathrm{PSD}} = \frac{1}{N^2}\sum_{\vartheta=0}^{N-1}\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right]\exp\left(j\frac{2\pi\tau\vartheta}{N}\right) \tag{4.14}$$

By replacing $\hat{\mathbf{s}}(\vartheta)$ by (4.4) in the PSD characteristic function (4.14) and using the definition of the time autocorrelation function (4.2), we obtain:

$$\begin{aligned}\left\langle \exp\left(j\frac{2\pi\tau\vartheta}{N}\right)\right\rangle_{\mathrm{PSD}} &= \frac{1}{N^2}\mathrm{E}_{\mathbf{s}}\left[\sum_{n_1=0}^{N-1}\mathbf{s}^*(n_1)\sum_{n=0}^{N-1}\mathbf{s}\left(n\right)\sum_{\vartheta=0}^{N-1}\exp\left(j\frac{2\pi\left(n_1-n+\tau\right)\vartheta}{N}\right)\right]\\ &= \frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\sum_{n_1=0}^{N-1}\mathbf{s}^*(n_1)\sum_{n=0}^{N-1}\mathbf{s}\left(n\right)\delta_d\left(n_1-n+\tau\right)\right]\\ &= \frac{1}{N}\sum_{n=0}^{N-1}\mathrm{E}_{\mathbf{s}}\left[\mathbf{s}\left(n\right)\mathbf{s}^*\left(n-\tau\right)\right] = \frac{1}{N}\sum_{n=0}^{N-1}\mathrm{R}_{\mathbf{s}}\left(n,\tau\right)\end{aligned} \tag{4.15}$$

From (4.15) and (4.14) it comes out that $\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right]$ is the Fourier transform of $\sum\limits_{n=0}^{N-1}\mathrm{R}_{\mathbf{s}}(n,\tau)$. Hence, we can write:

$$\frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}(\vartheta)\right|^2\right] = \frac{1}{N}\sum_{\tau=0}^{N-1}\sum_{n=0}^{N-1}\mathrm{R}_{\mathbf{s}}(n,\tau)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right) \tag{4.16}$$

Thus, the PSD of $\mathbf{s}$ can be obtained by taking the Fourier transform, with respect to the time lag variable τ of the sum over n of the time autocorrelation functions $\mathrm{R}_{\mathbf{s}}(n,\tau)$. This result constitutes a generalization of the Wiener-Khinchin theorem [135] for stochastic signals.

Following the same line of reasoning, the characteristic function of the power time density (4.3) is:

$$\left\langle \exp\left(j\frac{2\pi n\upsilon}{N}\right)\right\rangle_{\text{PTD}} = \frac{1}{N}\sum_{n=0}^{N-1} \mathrm{E}_{\mathbf{s}}\left[|\mathbf{s}(n)|^2\right]\exp\left(j\frac{2\pi n\upsilon}{N}\right) \tag{4.17}$$

By replacing $\mathbf{s}(n)$ by (4.7) in the above relation we obtain the dual form of the Wiener-Khinchin theorem:

$$\mathrm{E}_{\mathbf{s}}\left[|\mathbf{s}(n)|^2\right] = \frac{1}{N}\sum_{\upsilon=0}^{N-1}\sum_{\vartheta=0}^{N-1}\mathcal{R}^*_{\mathbf{s}}(\vartheta,\upsilon)\exp\left(-j\frac{2\pi n\upsilon}{N}\right) \tag{4.18}$$

Notice that the Parseval identity (4.12), the Wiener-Khinchin relation (4.16), and its dual form (4.18) connect time and frequency ensemble averages.

The time and frequency power densities: $\mathrm{E}_{\mathbf{s}}\left[|\mathbf{s}(n)|^2\right]$ and $\frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[|\hat{\mathbf{s}}(\vartheta)|^2\right]$ along with the time and frequency autocorrelation functions: $\mathrm{R}_{\mathbf{s}}(n,\tau)$ and $\mathcal{R}_{\mathbf{s}}(\vartheta,\upsilon)$ allow us to independently analyze $\mathbf{s}$ in time and frequency. We now turn to obtaining TF representations of $\mathbf{s}$ that permit to characterize the power and the correlation in the TF plane

Wigner-Ville transform

The fundamental power based TFR of a signal is its Wigner-Ville transform (WVT) [30]. The WVT of $\mathbf{s}$ is defined as:

$$W_{\mathbf{s}}(n,\vartheta) = \frac{1}{N}\sum_{\tau=0}^{N-1}\mathbf{s}^*(n-\tau)\mathbf{s}(n)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right) \tag{4.19}$$

the normalizing factor $\frac{1}{N}$ is introduced to satisfy the marginal properties (4.23) to (4.25).

The frequency version of the WVT is obtained by replacing $\mathbf{s}(n)$ by (4.7), in the WVT definition. This yields:

$$W_{\mathbf{s}}(n,\vartheta) \quad = \quad \frac{1}{N^2}\sum_{\upsilon=0}^{N-1}\hat{\mathbf{s}}^*(\vartheta)\hat{\mathbf{s}}(\vartheta-\upsilon)\exp\left(-j\frac{2\pi n\upsilon}{N}\right) \tag{4.20}$$

By taking ensemble averages on both sides in (4.19) and (4.20), and using the definitions of time (4.2) and frequency (4.10) autocorrelation functions, we obtain the expected WVT of $\mathbf{s}$:

$$\begin{aligned}\mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}(n,\vartheta)\right] &= \frac{1}{N}\sum_{\tau=0}^{N-1}\mathrm{R}_{\mathbf{s}}(n,\tau)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right) &(4.21)\\ &= \frac{1}{N}\sum_{\upsilon=0}^{N-1}\mathcal{R}^*_{\mathbf{s}}(\vartheta,\upsilon)\exp\left(-j\frac{2\pi n\upsilon}{N}\right) &(4.22)\end{aligned}$$

The expected WVT can be considered as an indicator of the signal power density in time and frequency. Indeed, $\mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}\left(n,\vartheta\right)\right]$ is real everywhere (since: $W_{\mathbf{s}}(n,\vartheta) = W^*_{\mathbf{s}}(n,\vartheta)$) and it satisfies the marginal properties, i.e. its sum over frequency (4.23) and time (4.24) gives the signal power density in time and frequency respectively, and the sum over time and frequency (4.25), scaled by N, gives the signal power.

$$\sum_{\vartheta=0}^{N-1} \mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}\left(n,\vartheta\right)\right] = \frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\sum_{\vartheta=0}^{N-1}\sum_{\tau=0}^{N-1} \mathbf{s}^*\left(n-\tau\right)\mathbf{s}\left(n\right)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right)\right] = \mathrm{E}_{\mathbf{s}}\left[\left|\mathbf{s}(n)\right|^2\right] \tag{4.23}$$

$$\sum_{n=0}^{N-1} \mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}\left(n,\vartheta\right)\right] = \frac{1}{N^2}\mathrm{E}_{\mathbf{s}}\left[\sum_{n=0}^{N-1}\sum_{\upsilon=0}^{N-1} \hat{\mathbf{s}}^*\left(\vartheta\right)\hat{\mathbf{s}}\left(\vartheta-\upsilon\right)\exp\left(-j\frac{2\pi n\upsilon}{N}\right)\right] = \frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[\left|\hat{\mathbf{s}}\left(\vartheta\right)\right|^2\right] \tag{4.24}$$

$$\frac{1}{N}\sum_{n=0}^{N-1}\sum_{\vartheta=0}^{N-1} \mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}\left(n,\vartheta\right)\right] = \mathrm{P}_{\mathbf{s}} \tag{4.25}$$

It is important to note that $\mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}\left(n,\vartheta\right)\right]$ is but an indicator of the signal power density. In fact, it cannot be interpreted in a point-wise sense because of the uncertainty principle, according to which the time and frequency power densities cannot both be made arbitrarily narrow[1]. In addition, $\mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}\left(n,\vartheta\right)\right]$ can be negative in some regions of the TF plane [30].

Since the WVT represents the signal in the TF plane, we can generalize the time (4.2) and frequency (4.10) autocorrelation functions and define the signal TF autocorrelation as:

$$\mathscr{R}_{\mathbf{s}}(n,\tau,\vartheta,\upsilon) = \frac{1}{N}\mathrm{E}_{\mathbf{s}}\left[W^*_{\mathbf{s}}(n-\tau,\vartheta-\upsilon)W_{\mathbf{s}}(n,\vartheta)\right] \tag{4.26}$$

where n and ϑ are the time and frequency at which the TF correlation is computed, and τ and υ are the time and frequency lags respectively.

Ambiguity function

Whereas the WVT seeks to combine power analysis in time and frequency, the fundamental correlative based TFR, namely the ambiguity function (AF) seeks to combine time and frequency correlation as embodied by the definitions (4.2), (4.10), and (4.26). The AF, is defined as the Fourier transform of the product: $\mathbf{s}^*(n-\tau)\mathbf{s}(n)$ with respect to time:

$$A_{\mathbf{s}}\left(\tau,\upsilon\right) = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{s}^*(n-\tau)\mathbf{s}(n)\exp\left(-j\frac{2\pi n\upsilon}{N}\right) \tag{4.27}$$

The frequency version of the AF is obtained by replacing $\mathbf{s}(n)$ by (4.7) in the above definition.

$$A_{\mathbf{s}}\left(\tau,\upsilon\right) = \frac{1}{N^2}\sum_{\vartheta=0}^{N-1}\hat{\mathbf{s}}^*(\vartheta-\upsilon)\hat{\mathbf{s}}(\vartheta)\exp\left(j\frac{2\pi\tau\vartheta}{N}\right) \tag{4.28}$$

[1] The uncertainty principle and its implications are detailed in [30]

By taking ensemble averages on both sides in (4.27) and (4.28), and using definitions (4.2) and (4.10), we obtain the expected AF of $\mathbf{s}$:

$$\begin{aligned} \mathrm{E}_{\mathbf{s}}\left[A_{\mathbf{s}}(\tau,\upsilon)\right] &= \frac{1}{N}\sum_{n=0}^{N-1}\mathrm{R}_{\mathbf{s}}(n,\tau)\exp\left(-j\frac{2\pi n\upsilon}{N}\right) \qquad (4.29)\\ &= \frac{1}{N}\sum_{\vartheta=0}^{N-1}\mathcal{R}_{\mathbf{s}}(\vartheta,\upsilon)\exp\left(j\frac{2\pi\tau\vartheta}{N}\right) \end{aligned}$$

The expected AF satisfies the marginal properties (4.30) and (4.31), i.e. the sum over the frequency lag gives the time autocorrelation computed at time $n = 0$ and the sum over the time lag gives the frequency autocorrelation function computed at frequency $\vartheta = 0$.

$$\sum_{\upsilon=0}^{N-1}\mathrm{E}_{\mathbf{s}}\left[A_{\mathbf{s}}(\tau,\upsilon)\right] = \mathrm{R}_{\mathbf{s}}(0,\tau) \qquad (4.30)$$

$$\sum_{\tau=0}^{N-1}\mathrm{E}_{\mathbf{s}}\left[A_{\mathbf{s}}(\tau,\upsilon)\right] = \mathcal{R}_{\mathbf{s}}(0,\upsilon) \qquad (4.31)$$

The expected square modulus of the AF, $\mathrm{E}_{\mathbf{s}}\left[\left|A_{\mathbf{s}}(\tau,\upsilon)\right|^2\right]$ is an indicator of the global TF correlation for all the TF points separated, in time by τ and in frequency by υ. Indeed, by taking the sum over n and ϑ of the TF autocorrelation definition (4.26) and using the WVT (4.19) and AF (4.27) definitions, we have:

$$\sum_{n=0}^{N-1}\sum_{\vartheta=0}^{N-1}\mathscr{R}_{\mathbf{s}}(n,\tau,\vartheta,\upsilon) = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{\vartheta=0}^{N-1}\mathrm{E}_{\mathbf{s}}\left[W_{\mathbf{s}}(n,\vartheta)W_{\mathbf{s}}^{*}(n-\tau,\vartheta-\upsilon)\right]$$

$$= \frac{1}{N^3}\mathrm{E}_{\mathbf{s}}\left[\sum_{\tau_1,\tau_2,n,\vartheta=0}^{N-1}\mathbf{s}^{*}(n-\tau_1)\mathbf{s}(n)\mathbf{s}(n-\tau-\tau_2)\mathbf{s}^{*}(n-\tau)\exp\left(j\frac{2\pi(\tau_2\upsilon-\tau_1\upsilon-\tau_2\upsilon)}{N}\right)\right]$$

$$\sum_{n=0}^{N-1}\sum_{\vartheta=0}^{N-1}\mathscr{R}_{\mathbf{s}}(n,\tau,\vartheta,\upsilon) = \mathrm{E}_{\mathbf{s}}\left[\left|A_{\mathbf{s}}(\tau,\upsilon)\right|^2\right] \qquad (4.32)$$

This result constitutes a global indicator of the interaction in the TF plane. Its generalization to the analysis of a multivariate signal permits to characterize the interaction between its univariate components.

4.2.2 Time-frequency analysis of multivariate stochastic signals

Let $\mathbf{S}$ be an M dimensional multivariate stochastic signal of length N, composed of the random vectors: $\left\{\mathbf{S}(n) = (\mathbf{s}_1(n)\ldots\mathbf{s}_M(n))^{\mathrm{t}}\,\middle|\, n = 0,\ldots,N-1\right\}$ where $\mathbf{s}_1,\ldots,\mathbf{s}_M$ are the univariate components of $\mathbf{S}$ and $^{\mathrm{t}}$ is the transpose operator.

The TF analysis of multivariate signals that we consider in this thesis is based on second order statistics. Thus, the results obtained in the previous section can be easily generalized by considering second order moments between the univariate components of $\mathbf{S}$.

Time and frequency inter-correlations

We define the time and frequency inter-correlation functions of $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ as[1]:

$$\mathrm{R}_{m1,m2}\left(n,\tau\right)=\mathrm{E}_{\mathbf{S}}\left[\mathbf{s}_{m1}^{*}\left(n-\tau\right)\mathbf{s}_{m2}\left(n\right)\right] \tag{4.33}$$

$$\mathcal{R}_{m1,m2}\left(\vartheta,\upsilon\right)=\frac{1}{N}\mathrm{E}_{\mathbf{S}}\left[\hat{\mathbf{s}}_{m1}^{*}\left(\vartheta-\upsilon\right)\hat{\mathbf{s}}_{m2}\left(\vartheta\right)\right] \tag{4.34}$$

where $\hat{\mathbf{s}}_{m1}$ and $\hat{\mathbf{s}}_{m2}$ are the Fourier transforms of $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ respectively. As we did in the previous section, we do not explicitly write the probability density functions with respect to which the ensemble averages are computed. Instead, we simply denote as $\mathrm{E}_{\mathbf{S}}\left[\cdot\right]$ any ensemble average over $\mathbf{S}$.

In the previous section we have seen that the PSD could be obtained by taking the Fourier transform, with respect to the time lag, of the sum over time of the time autocorrelation function (4.16). This result can be generalized to the multivariate case by substituting the time autocorrelation function in (4.15) by the time inter-correlation function (4.33). Hence, we obtain:

$$\frac{1}{N}\left[\sum_{\tau=0}^{N-1}\sum_{n=0}^{N-1}\mathrm{R}_{m1,m2}(n,\tau)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right)\right]=\frac{1}{N}\mathrm{E}_{\mathbf{S}}\left[\hat{\mathbf{s}}_{m1}^{*}(\vartheta)\hat{\mathbf{s}}_{m2}(\vartheta)\right] \tag{4.35}$$

As in the univariate TF analysis, we call $\frac{1}{N}\mathrm{E}_{\mathbf{S}}\left[\hat{\mathbf{s}}_{m1}^{*}(\vartheta)\hat{\mathbf{s}}_{m2}(\vartheta)\right]$, the power inter-spectrum density of $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$. In fact, this result generalizes the signal cross-spectrum definition [135].

Inter Wigner-Ville transform and inter ambiguity function

Similarly to the time and frequency inter-correlation functions, the inter-WVT and inter-AF of $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ can be respectively defined as:

$$W_{m1,m2}\left(n,\vartheta\right)=\frac{1}{N}\sum_{\tau=0}^{N-1}\mathbf{s}_{m1}^{*}(n-\tau)\mathbf{s}_{m2}(n)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right) \tag{4.36}$$

$$A_{m1,m2}\left(\tau,\upsilon\right)=\frac{1}{N}\sum_{n=0}^{N-1}\mathbf{s}_{m1}^{*}(n-\tau)\mathbf{s}_{m2}(n)\exp\left(-j\frac{2\pi n\upsilon}{N}\right) \tag{4.37}$$

The TF inter-correlation function of $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ at time n and frequency ϑ for a time lag τ and frequency lag υ is:

$$\mathscr{R}_{m1,m2}(n,\tau,\vartheta,\upsilon)=\frac{1}{N}\mathrm{E}_{\mathbf{S}}\left[W_{m1,m2}^{*}(n-\tau,\vartheta-\upsilon)W_{m1,m2}(n,\vartheta)\right] \tag{4.38}$$

The global TF inter-correlation between $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ at time lag τ and frequency lag υ is given by the sum over time and frequency of the TF inter-correlation function (4.38). Using (4.36) and (4.37), we obtain:

$$\sum_{n=0}^{N-1}\sum_{\vartheta=0}^{N-1}\mathscr{R}_{m1,m2}(n,\tau,\vartheta,\upsilon)=\mathrm{E}_{\mathbf{S}}\left[\left|A_{m1,m2}(\tau,\upsilon)\right|^{2}\right] \tag{4.39}$$

[1] We use the prefix inter in a general sense. When $m_1=m_2$, this prefix is usually replaced by intra

In consequence, the expectation of the modulus of the inter-AF of $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ gives an indication of the global TF interaction between these two signals.

It should be noted that:

$$\mathrm{E}_{\mathbf{S}}\left[|A_{m1,m2}(\tau,\upsilon)|^2\right] = \mathrm{E}_{\mathbf{S}}\left[|A_{m2,m1}(\tau,\upsilon)|^2\right]$$

and, if $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ are real then[1]:

$$|A_{m1,m2}(\tau,N-\upsilon)| = |A_{m1,m2}(\tau,\upsilon)| \qquad \upsilon = 1,\ldots,\frac{N}{2} \tag{4.40}$$

The multivariate stochastic signal $\mathbf{S}$ can therefore be characterized by the set:

$$I_{\mathbf{S}} = \left\{ \mathrm{E}_{\mathbf{S}}\left[|A_{m1,m2}\left(\tau,\upsilon\right)|^2\right] \middle| 1 \leqslant m_1 \leqslant m_2 \leqslant M; \tau = 0,\ldots,N-1; \upsilon = 0,\ldots,\frac{N}{2} \right\}$$

which contains the values of the modulus of the inter-AFs for every possible pair of univariate components, time lags and frequency lags. The number of elements in $I_{\mathbf{S}}$ is:

$$|I_{\mathbf{S}}| = \frac{MN\left(M+1\right)\left(N+2\right)}{4}$$

Typical values of M and N, in the EEG framework are in the order of ten channels and a few hundreds of samples respectively. With these values, the number of elements in $I_{\mathbf{S}}$ is in the order of hundred of thousands. Clearly, a mapping from the EEG-trial set to a feature vector space directly generated by $I_{\mathbf{S}}$ cannot be used for BCI applications.

Under certain hypotheses on the nature of the $\mathbf{S}$ it is possible to reduce the number of statistical measurements that are necessary to characterize it. In the following we present these hypotheses and establish their implications in the framework of the analysis of $\mathbf{S}$.

4.2.3 Stationarity

Stationarity of $\mathbf{S}$ implies that its statistical properties do not change with time. However, this condition is hardly met in practice. As we employ statistical moments up to second order, we consider a weaker form of stationarity called wide sense stationarity. In the following we use stationarity to refer to wide sense stationarity. Thus, $\mathbf{S}$ is stationary if:

- The average $\mathrm{E}_{\mathbf{S}}\left[\mathbf{s}_m\left(n\right)\right]$ is independent of n, i.e. $\mathrm{E}_{\mathbf{S}}\left[\mathbf{s}_m\left(n\right)\right] = \mu_m$
- The inter-correlation function of any pair of univariate components $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ depends only upon the time lag τ for every time n, i.e. $\mathrm{R}_{m1,m2}\left(n,\tau\right) = \mathrm{R}_{m1,m2}\left(\tau\right)$. In particular, when $m1 = m2 = m$, one has: $\mathrm{R}_m(n,\tau) = \mathrm{R}_m(\tau) = \mathrm{R}(-\tau)$.

Because of the stationarity conditions, the power inter-spectrum density of $\mathbf{s}_{m1}$ and $\mathbf{s}_{m2}$ (4.35) becomes simply the Fourier transform of the time inter-correlation function:

$$\frac{1}{N}\mathrm{E}_{\mathbf{S}}\left[\hat{\mathbf{s}}^*_{m1}(\vartheta)\hat{\mathbf{s}}_{m2}(\vartheta)\right] = \sum_{\tau=0}^{N-1} \mathrm{R}_{m1,m2}(\tau)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right) \tag{4.41}$$

[1] For convenience we assume that N is even

and the expected inter-WVT depends only upon the frequency. Indeed, by taking ensemble averages on both sides in (4.36) we have:

$$\mathrm{E}_{\mathbf{S}}\left[W_{m1,m2}\left(n,\vartheta\right)\right]=\frac{1}{N}\sum_{\tau=0}^{N-1}\mathrm{R}_{m1,m2}(\tau)\exp\left(-j\frac{2\pi\tau\vartheta}{N}\right) \tag{4.42}$$

This result implies that the spectral properties of $\mathbf{S}$ do not change over time. Furthermore, by taking ensemble averages on both sides of (4.37), we obtain:

$$\mathrm{E}_{\mathbf{S}}\left[A_{m1,m2}\left(\tau,\upsilon\right)\right]=\mathrm{R}_{m1,m2}\left(\tau\right)\delta_{d}\left(\upsilon\right) \tag{4.43}$$

Therefore, a stationary signal does not present any correlation in frequency.

Since the spectral properties of a stationary signal do not change over time and there is no frequency correlation, we can describe $\mathbf{S}$ using (4.41).

4.2.4 Ergodicity

For a stationary signal $\mathbf{S}$, ergodicity implies that ensemble averages can be replaced by time averages. Thus, the stationary time inter-correlation function, under the hypothesis of ergodicity becomes:

$$\mathrm{R}_{m1,m2}(\tau)=\frac{1}{N}\sum_{n=0}^{N-1}s_{m1}(n-\tau)s_{m2}(n) \tag{4.44}$$

Replacing this result in the stationary power inter-spectrum density (4.41) we obtain:

$$\frac{1}{N}\mathrm{E}_{\mathbf{S}}\left[\hat{\mathbf{s}}_{m1}^{*}(\vartheta)\hat{\mathbf{s}}_{m2}(\vartheta)\right]=\frac{1}{N}\hat{s}_{m1}^{*}(\vartheta)\hat{s}_{m2}(\vartheta) \tag{4.45}$$

where $\hat{s}_{m1}$ and $\hat{s}_{m2}$ are the Fourier transforms of s_{m1} and s_{m2} respectively.

If $m1=m2=m$ then, $\frac{1}{N}\left|\hat{s}_{m}(\vartheta)\right|^{2}$ represents the PSD of $\mathbf{s}_{m}$. Therefore, according to (4.12) and (4.45), the power of $\mathbf{s}_{m}$, is:

$$\mathrm{P}_{\mathbf{s}_{m}}=\frac{1}{N^{2}}\sum_{\vartheta=0}^{N-1}\left|\hat{s}_{m}(\vartheta)\right|^{2} \tag{4.46}$$

Replacing (4.6) in the above relation yields:

$$\begin{aligned}\mathrm{P}_{\mathbf{s}_{m}} &= \frac{1}{N^{2}}\left(\sum_{\vartheta=0}^{\frac{N}{2}}\left|\hat{s}_{m}(\vartheta)\right|^{2}+\sum_{\vartheta=\frac{N}{2}+1}^{N-1}\left|\hat{s}_{m}(\vartheta)\right|^{2}\right)\\ &= \frac{1}{N^{2}}\left(2\sum_{\vartheta=0}^{\frac{N}{2}}\left|\hat{s}_{m}(\vartheta)\right|^{2}-\left|\hat{s}_{m}\left(\frac{N}{2}\right)\right|^{2}\right) \qquad (4.47)\end{aligned}$$

Since the sampling frequency is at least twice the maximum frequency present in the spectrum of s_{m}, i.e. the sampling frequency is chosen in accordance with the sampling theorem [119, 150], the second term on the right in (4.47) is close to zero. Therefore, the following approximation holds.

$$\mathrm{P}_{\mathbf{s}_m} \approx \frac{2}{N^2} \sum_{\vartheta=0}^{\frac{N}{2}} |\hat{s}_m(\vartheta)|^2 \tag{4.48}$$

Using this approximation and the correspondence between ϑ and the real frequency (4.5), we can approximate the power contained in a frequency band, $\mathrm{B_f} = [\mathrm{f}_1; \mathrm{f}_2] \subset \left[0; \frac{\mathrm{f_s}}{2}\right]$, as follows.

$$\tilde{\mathrm{P}}_{\mathbf{s}_m}(\mathrm{B_f}) = \frac{2}{N^2} \sum_{\vartheta=\vartheta_1}^{\vartheta_2} |\hat{s}_m(\vartheta)|^2 \tag{4.49}$$

where:

$$\vartheta_i = \mathrm{nint}\left(\frac{N\mathrm{f}_i}{\mathrm{f_s}}\right) \qquad i = 1, 2$$

The function nint$(\cdot)$ gives the nearest integer to its argument. In particular, it can be said that $\frac{2}{N^2}\left|\hat{s}_m\left(\mathrm{nint}\left(\frac{N\mathrm{f}}{\mathrm{f_s}}\right)\right)\right|^2$ represents the power contained in an $\frac{\mathrm{f}_s}{N}$ wide band centered at f.

Ensemble averages allowed us to theoretically develop the TF analysis framework. In practice, such averages are difficult to compute as in practice, one has no access to the signal's generative mechanism. Thus, under the hypothesis of ergodicity this problem has been overcome in the framework of stationary signals.

The stationarity and ergodicity hypothesis are used in the stationary PSD (Section 4.3), autoregressive (Section 4.5), multivariate autoregressive (Section 4.7), and coherence (Section 4.4) mappings.

4.2.5 Absence of coupling between the univariate components

The hypothesis of absence of coupling between the univariate components of the $\mathbf{S}$ implies that the correlations between two different univariate components can be ignored. Thus, in the general characterization set $I_{\mathbf{S}}$ (see Section 4.2.2) only the terms of the form: $\mathrm{E}_{\mathbf{S}}\left[|A_{m,m}(\tau, \upsilon)|^2\right]$ need to be considered. In particular, if $\mathbf{S}$ is stationary only the PSDs associated with each component are considered.

This hypothesis is used in the stationary PSD (Section 4.3), and autoregressive (Section 4.5) mappings.

4.2.6 Existence of a linear prediction model

According to this hypothesis, $\mathbf{S}$ can be generated by a a linear prediction model of the form:

$$\mathbf{S}(n) = -\sum_{i=1}^{Q} \mathbf{A}(n,i)\mathbf{S}(n-i) + \mathbf{e}(n) \tag{4.50}$$

where Q is the model order, the $\mathbf{A}(n,i)$ are $N \times N$ matrices and $\mathbf{e}(n)$ (the prediction error) is an N-dimensional zero mean random vector with covariance matrix $C_{\mathbf{e}}(n)$. Thus, $\mathbf{S}$ is completely determined by the parameters of the model.

It can be shown [22] that if $\mathbf{S}$ is stationary and ergodic, the matrices $\mathbf{A}(n,i)$ are time independent, i.e. $\forall n, \mathbf{A}(n,i) = \mathbf{A}(i)$. In this case the linear prediction model is called a stationary autoregressive model (AR model). On the other hand, if $\mathbf{S}$ is non-stationary, the matrices $\mathbf{A}(n,i)$ are time dependent. In this case, the linear prediction model is called a non-stationary autoregressive model (NAR model).

The hypothesis of existence of a linear prediction model is used in the autoregressive (Section 4.5), non-stationary autoregressive (Section 4.6), and the multivariate autoregressive mappings (Section 4.7).

4.2.7 Weak coupling

This hypothesis is based on the assumption that the univariate signals composing $\mathbf{S}$ are generated by self-sustained oscillators[1] which are weakly coupled.

The interaction between two self-sustained oscillators of natural frequencies f_1 and f_2 (without loss of generality we can assume that $\mathrm{f}_1 < \mathrm{f}_2$) whose mutual influence is approximately symmetrical, entrains frequency locking as their observed coupled frequencies $\tilde{\mathrm{f}}_1$ and $\tilde{\mathrm{f}}_2$ are such that $\tilde{\mathrm{f}}_1 = \tilde{\mathrm{f}}_2 = \tilde{\mathrm{f}}$ where typically, $\mathrm{f}_1 < \tilde{\mathrm{f}} < \mathrm{f}_2$. Frequency locking implies a certain relation between the oscillators phases that depends not only on the frequency detuning, $\mathrm{f}_1 - \mathrm{f}_2$, and coupling strength, but also on the way in which the oscillators are interacting. It is well known that weak coupling affects primarily the phases of the oscillators but not their amplitudes [131]. Thus, when the natural frequencies obey the relation $i_1\mathrm{f}_1 = i_2\mathrm{f}_2$ (where i_1, i_2 are positive integer numbers), phase locking (or synchronization [16]) of order $i_1 : i_2$ arises. The condition of synchronization can be formulated as:

$$|i_1\varphi_1 - i_2\varphi_2| < \epsilon \tag{4.51}$$

where φ_1 and φ_2 are the phases of the coupled oscillators, and ϵ is a small value [131]. As the oscillators considered come from the same physiological system, only synchronization of order $1:1$ is considered [16].

Thus under the weak coupling hypothesis, the analysis of the interaction between the univariate components of $\mathbf{S}$ focuses on the computation of the degree of synchronization between them. Since phase locking implies frequency locking, synchronization should be determined in narrow frequency bands. This hypothesis is used in the synchronization mapping in Section 4.8.

So far we have presented the theoretical elements to analyze a multivariate stochastic signal $\mathbf{S}$ in time and frequency. In addition, we established hypotheses on the nature of $\mathbf{S}$ that make it possible to simplify the analysis. As mentioned in Section 4.1, in the framework

[1] A self-sustained oscillator is an active system that contains an internal source of energy that is transformed into oscillatory activity which is entirely determined by the oscillator internal parameters. Neuronal oscillators are good examples of self-sustained oscillators

of BCI applications we have at our disposal a realization of **S**, namely an EEG-trial from which a feature vector should be extracted. The computation of this feature vector is done through a mapping from the EEG-trial set into a feature vector space. In the following we present different mappings that are based on the hypotheses afore mentioned.

For all the mappings we assume that the EEG-trial:

$$S = \begin{pmatrix} s_1(0) & \dots & s_1(\mathrm{N_{spt}}-1) \\ \vdots & \vdots & \vdots \\ s_{\mathrm{Ne}}(0) & \dots & s_{\mathrm{Ne}}(\mathrm{N_{spt}}-1) \end{pmatrix}$$

where N_e and $\mathrm{N_{spt}}$ are the number of electrodes and number of samples per trial respectively, is given (for simplicity, we assume that $\mathrm{N_{spt}}$ is even). In addition, as mentioned in Chapter 3 the averages over time and electrode are both equal to zero. This implies:

$$\begin{aligned} \sum_{n=0}^{\mathrm{N_{spt}}-1} s_m(n) = 0 \qquad & m = 1, \dots, \mathrm{N}_e \\ \sum_{m=1}^{\mathrm{N}_e} s_m(n) = 0 \qquad & n = 0, \dots, \mathrm{N_{spt}}-1 \end{aligned} \tag{4.52}$$

4.3 Stationary PSD mapping

The stationary PSD mapping, denoted as ψ_{P} is built on the hypotheses of stationarity, ergodicity, and absence of coupling. Thus, each EEG channel can be independently analyzed by means of its power spectrum density.

To compute the feature vector $\psi_{\mathrm{P}}(S)$, $\mathrm{N_B}$ frequency bands $\{\mathrm{B}_1, \dots, \mathrm{B}_{\mathrm{N_B}}\}$ are chosen in accordance with physiological considerations (for instance the typical delta, theta, alpha, beta, and gamma frequency bands). The frequency bands used in this thesis are presented in Chapter 6.

The powers for each frequency band and EEG channel are computed and grouped into an $\mathrm{N}_e\mathrm{N_B}$ dimensional vector:

$$\psi_{\mathrm{P}}(S) = \begin{pmatrix} \mathrm{P}_1(\mathrm{B}_1) & \dots & \mathrm{P}_m(\mathrm{B}_i) & \dots & \mathrm{P}_{\mathrm{N}_e}(\mathrm{B}_{\mathrm{N_B}}) \end{pmatrix}^{\mathrm{t}} \in \mathbb{R}^{\mathrm{N}_e\mathrm{N_B}}$$

where $\mathrm{P}_m(\mathrm{B}_i)$ is the power of s_m in the frequency band B_i.

To compute $\mathrm{P}_m(\mathrm{B}_i)$, the PSD of s_m is estimated using the Welch method [172] (see Fig. 4.2). In this method, the signal s_m is segmented into N_β (possibly) overlapping blocks containing N samples each. The blocks are then multiplied by an N-point Hamming window [135] which smoothly reduces the samples in each block to zero at the end points (see Fig. 4.2). Windowing aims at attenuating the spectral leakage effect[1] due to the discontinuities in time introduced by the segmentation.

[1] Spectral leakage means that signal energy which should be concentrated only at one frequency instead leaks into all the other frequencies

The Welch estimate of the PSD of s_m, denoted as $\mathcal{W}_{s_m}(\vartheta)$, is the average of the PSDs of the windowed blocks. Using (4.46), we have:

$$\mathcal{W}_{s_m}(\vartheta) = \frac{1}{\mathrm{N}_\beta N} \sum_{\beta=1}^{\mathrm{N}_\beta} |\hat{s}_\beta(\vartheta)|^2 \tag{4.53}$$

where $\hat{s}_\beta(\vartheta)$ is the Fourier transform of the β-th windowed block.

Finally, using (4.49) the power in the frequency band $\mathrm{B}_i = [\mathrm{f}_{i,1}; \mathrm{f}_{i,2}]$ is:

$$\mathrm{P}_m(\mathrm{B}_i) = \frac{2}{N} \sum_{\vartheta=\vartheta_{i,1}}^{\vartheta_{i,2}} \mathcal{W}_{s_m}(\vartheta) = \frac{2}{\mathrm{N}_\beta N^2} \sum_{\vartheta=\vartheta_{i,1}}^{\vartheta_{i,2}} \sum_{\beta=1}^{\mathrm{N}_\beta} |\hat{s}_\beta(\vartheta)|^2 \tag{4.54}$$

where:

$$\vartheta_{i,l} = \mathrm{nint}\left(\frac{N\mathrm{f}_{i,l}}{\mathrm{f_s}}\right) \qquad l = 1, 2$$

Variants of this mapping, e.g. taking different frequency bands and subsets of electrodes (following physiological considerations) are used in numerous current BCIs [106, 128, 132, 176].

4.4 Coherence mapping

The coherence mapping, denoted as ψ_C is built on the hypotheses of stationarity and ergodicity of the EEG signals. Under these hypotheses, an EEG-trial can be analyzed by means of the power inter-spectrum densities (4.45) between its channels.

The coherence function evaluates the inter-spectrum density of two EEG channels normalized by the channel PSDs. The coherence between s_{m1} and s_{m2} at frequency ϑ is defined as:

$$\mathrm{C}_{m1,m2}(\vartheta) = \frac{|\hat{s}^*_{m1}(\vartheta)\hat{s}_{m2}(\vartheta)|^2}{|\hat{s}_{m1}(\vartheta)|^2\,|\hat{s}_{m2}(\vartheta)|^2} \tag{4.55}$$

where $\hat{s}_{m1}(\vartheta)$ and $\hat{s}_{m2}(\vartheta)$ are the Fourier transforms of s_{m1} and s_{m2} respectively.

Because of the normalization by the channel PSDs, the coherence function takes values in the interval $[0, 1]$. In particular, if s_{m1} and s_{m2} are uncorrelated their coherence is zero at all frequencies.

In practice, the coherence is not directly computed using (4.55), instead it is estimated by segmenting the observed s_{m1} and s_{m2} into N_β (possibly) overlapping N-length blocks and computing [14]:

$$\mathcal{C}_{m1,m2}(\vartheta) = \frac{\left|\sum\limits_{\beta=0}^{\mathrm{N}_\beta} \hat{s}^*_{m1,\beta}(\vartheta)\hat{s}_{m2,\beta}(\vartheta)\right|^2}{\sum\limits_{\beta=0}^{\mathrm{N}_\beta} |\hat{s}_{m1,\beta}(\vartheta)|^2 \sum\limits_{\beta=0}^{\mathrm{N}_\beta} |\hat{s}_{m2,\beta}(\vartheta)|^2} \tag{4.56}$$

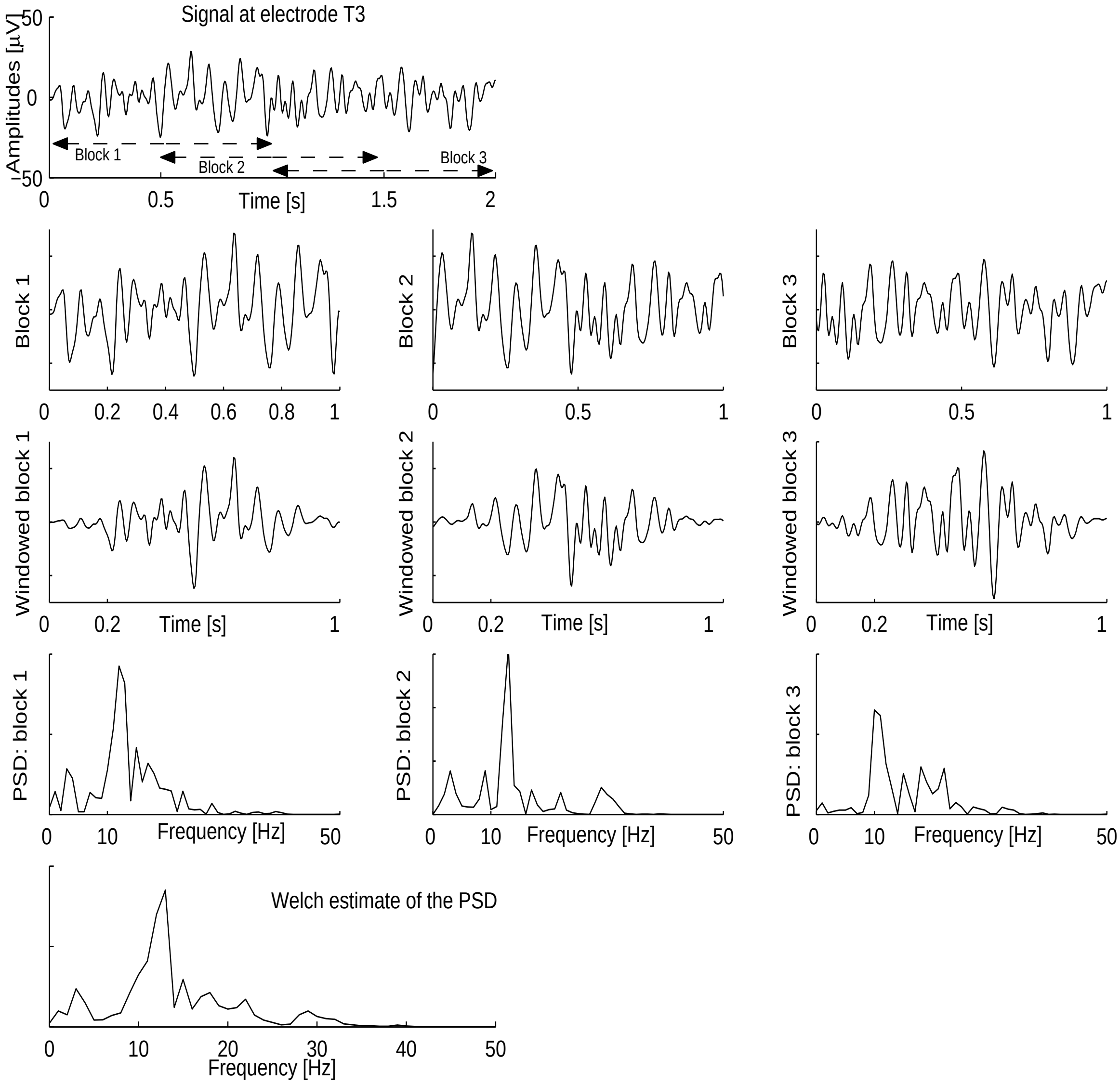

Figure 4.2. Welch method to estimate the power spectral density PSD. The signal is segmented into blocks that can overlap. These blocks are windowed by a Hamming window, their respective PSDs are computed and averaged. This average constitutes the estimated PSD. The signal under study was recorded at electrode T3 while the subject was reading a text on a computer screen.

where $\hat{s}_{m1,\beta}$ and $\hat{s}_{m2,\beta}$ are the Fourier transforms of the β-th blocks of signals s_{m1} and s_{m2} respectively (see Fig. 4.3).

To compute the feature vector $\psi_{\mathrm{C}}(S)$, $\mathrm{N_B}$ frequency bands $\{\mathrm{B}_1, \ldots, \mathrm{B}_{\mathrm{N_B}}\}$ are chosen and the average coherence for each frequency band and pair of EEG channels are computed and grouped into an $\frac{\mathrm{N}_e(\mathrm{N}_e-1)}{2}\mathrm{N_B}$ dimensional vector:

$$\psi_{\mathrm{C}}(S) = \Big(\; \langle \mathcal{C}_{1,2}(\vartheta) \rangle_{\mathrm{B}_1} \quad \ldots \quad \langle \mathcal{C}_{m1<m2,m2}(\vartheta) \rangle_{\mathrm{B}_i} \quad \ldots \quad \langle \mathcal{C}_{\mathrm{N}_e-1,\mathrm{N}_e}(\vartheta) \rangle_{\mathrm{B}_{\mathrm{N_B}}} \; \Big)^{\mathrm{t}} \in \mathbb{R}^{\frac{\mathrm{N}_e(\mathrm{N}_e-1)}{2}\mathrm{N_B}}$$

where $\langle \mathcal{C}_{m1,m2}(\vartheta) \rangle_{\mathrm{B}_i}$ is the average coherence in the frequency band $\mathrm{B}_i = [\mathrm{f}_{i,1}; \mathrm{f}_{i,2}]$. Such average is computed as follows.

$$\langle \mathcal{C}_{m1,m2}(\vartheta) \rangle_{\mathrm{B}_i} = \frac{1}{\vartheta_{i,2} - \vartheta_{i,1}} \sum_{\vartheta=\vartheta_{i,1}}^{\vartheta_{i,2}} \mathcal{C}_{m1,m2}(\vartheta) \tag{4.57}$$

where

$$\vartheta_{i,l} = \mathrm{nint}\left(\frac{N \mathrm{f}_{i,l}}{\mathrm{f_s}} \right) \qquad l = 1, 2$$

The coherence function is extensively used as a tool for quantifying the degree of interaction between two EEG channels in a frequency band. A large value of the average of the coherence function in a certain frequency band indicates that the corresponding oscillatory activities are of the same origin or interact with each other [14, 117, 118, 174].

4.5 Autoregressive mapping

The autoregressive (AR) mapping, denoted as ψ_{AR} is built on the hypotheses of stationarity, ergodicity, absence of coupling between the univariate components, and existence of a linear prediction model.

Since the coupling between the channels is ignored, the model in (4.50) can be split into linear prediction models corresponding to each univariate component. Thus, the m-th univariate component of $\mathbf{S}$ can be written in the form:

$$\mathbf{s}_m(n) = -\sum_{i=1}^{Q_m} a_m(n,i) \mathbf{s}_m(n-i) + \mathbf{e}_m(n) \tag{4.58}$$

where the $a_m(n,i)$ are the AR coefficients and Q_m is the AR order corresponding to $\mathbf{s}_m$, and $\mathbf{e}_m$ is the m-th prediction error process.

Furthermore, as stationarity and ergodicity are assumed, it can be shown [22] that the AR coefficients are time independent. Thus, the AR model for the m-th channel becomes:

$$s_m(n) = -\sum_{i=1}^{Q_m} a_m(i) s_m(n-i) + \mathrm{e}_m(n) \tag{4.59}$$

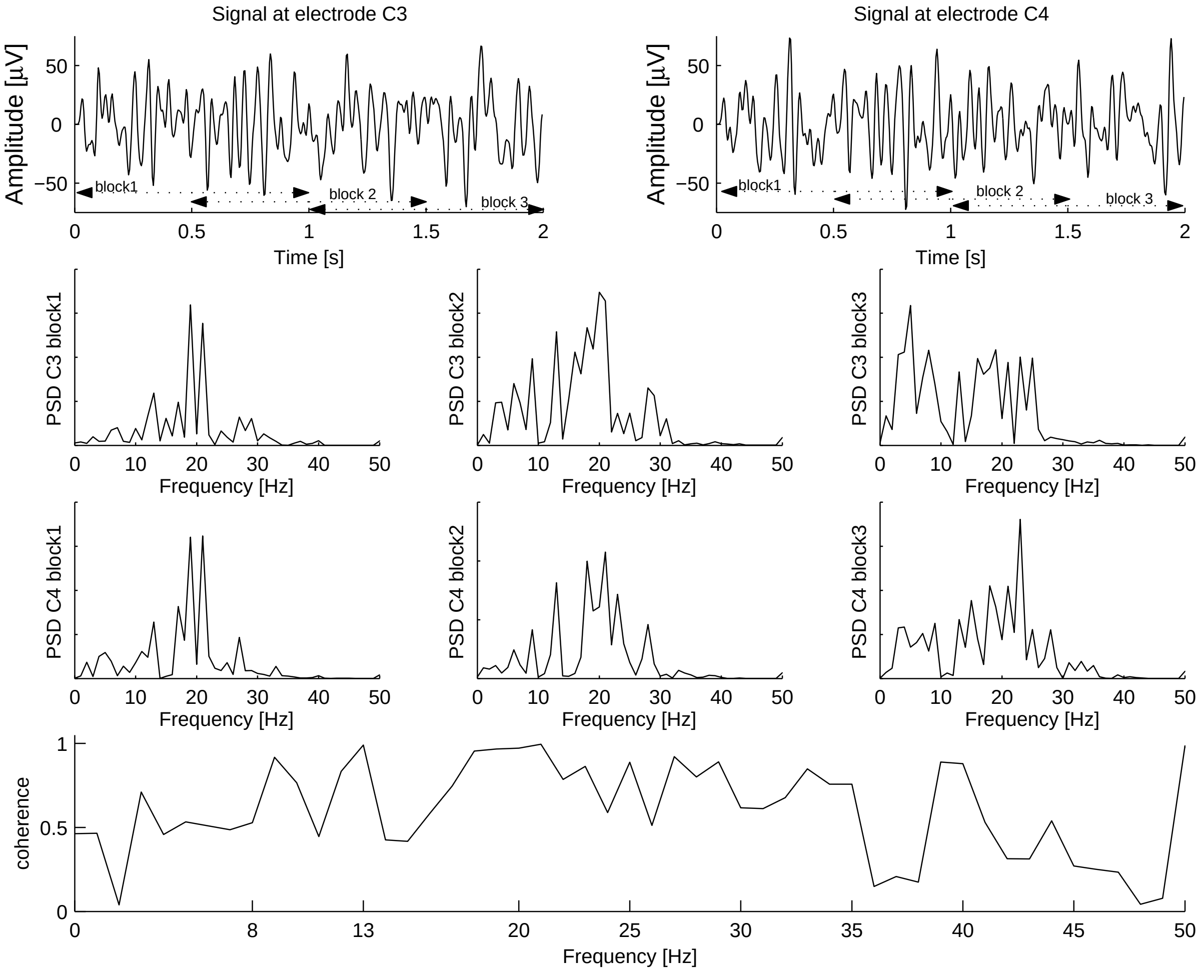

Figure 4.3. Estimation of the coherence function between two EEG channels. In this example, the coherence function is estimated by segmenting both signals into blocks of one second duration, computing the PSDs of each block and using (4.56).

The signals under study were recorded at electrodes C3 (top left) and C4 (top right) while the subject was imagining the movement of his left index finger.

The coefficients: $a_m(1), \ldots, a_m(Q_m)$ can be determined by minimizing the averaged squared prediction error, i.e. the prediction error power:

$$\mathcal{E}(Q_m) = \frac{1}{\mathrm{N_{spt}}} \sum_{n=0}^{\mathrm{N_{spt}}-1} e_m^2(n) = \frac{1}{\mathrm{N_{spt}}} \sum_{n=0}^{\mathrm{N_{spt}}-1} \left(s_m(n) + \sum_{i=1}^{Q_m} a_m(i) s(n-i) \right)^2 \tag{4.60}$$

in this relation, the samples prior to $s_m(0)$ are assumed to be zero.

Taking the derivatives of $\mathcal{E}(Q_m)$ with respect to $a_m(1), \ldots, a_m(Q_m)$ and setting them to zero, yields:

$$\frac{1}{\mathrm{N_{spt}}} \sum_{n=0}^{\mathrm{N_{spt}}-1} s_m(n) s_m(n-i) + \sum_{i'=1}^{Q_m} a_m(i') \frac{1}{\mathrm{N_{spt}}} \sum_{n=0}^{\mathrm{N_{spt}}-1} s_m(n-i') s_m(n-i) = 0 \qquad i = 1, \ldots, Q_m \tag{4.61}$$

By replacing the ergodicity condition on the time correlation function (4.44) in (4.61), we obtain the so called Yule-Walker [170, 182] equations:

$$\sum_{i'=1}^{Q_m} a_m(i') \mathrm{R}_{s_m}(i-i') = -\mathrm{R}_{s_m}(i) \qquad i = 1, \ldots, Q_m \tag{4.62}$$

The AR coefficients can be efficiently found by solving (4.62) using the recursive Levinson-Durbin algorithm [42, 97].

Since $\mathbf{s}_m$ is ergodic, it can be shown [22] that $\mathbf{e}_m$ is an independent and identically distributed (IID) stochastic process with mean zero and finite variance $\mathcal{E}(Q_m)$. Therefore, the spectrum of $\mathbf{e}_m$ is [22]: $\hat{\mathbf{e}}_m(\mathrm{f}) = \frac{\mathcal{E}(Q_m)}{\mathrm{f}_s}$ where f_s is the sampling frequency.

Taking the z-transform [135] on both sides in (4.59) yields:

$$Z_{s_m}(z) = \frac{Z_{e_m}(z)}{1 + \sum\limits_{i=1}^{Q_m} a_m(i) z^{-i}} \tag{4.63}$$

where $Z_{s_m}(z)$ and $Z_{s_m}(z)$ are the z-transforms of s_m and e_m respectively. The spectrum of s_m is obtained by evaluating (4.63) along the unit circle in the z-plane, i.e. $z = \exp\left(j\frac{2\pi\mathrm{f}}{\mathrm{f}_z}\right)$:

$$\hat{s}_m(\mathrm{f}) = \frac{\mathcal{E}(Q_m)}{\mathrm{f}_s \left(1 + \sum\limits_{i=1}^{Q_m} a_m(i) \exp\left(-j\frac{2\pi\mathrm{f}i}{\mathrm{f}_s}\right)\right)} = \frac{\mathcal{E}(Q_m)}{\mathrm{f}_s} \mathcal{H}_m(\mathrm{f}) \tag{4.64}$$

where f_s is the sampling frequency.

The PSD of s_m is given by: $\left|\frac{\mathcal{E}(Q_m)}{\mathrm{f}_s}\mathcal{H}_m(\mathrm{f})\right|^2$. Thus, the flat spectrum of $\mathbf{e}_m$ is filtered by the all-pole filter $\mathcal{H}_m(\mathrm{f})$ (see Fig. 4.4) to produce an output spectrum which contains sharp peaks at certain frequencies corresponding to the zeros of the denominator of (4.64) (poles). This property makes the AR model particularly suited for EEG signals whose PSD is generally characterized by dominant frequencies (see Figs. 4.2 to 4.5) rather than by the absence of power at certain frequencies (notches).

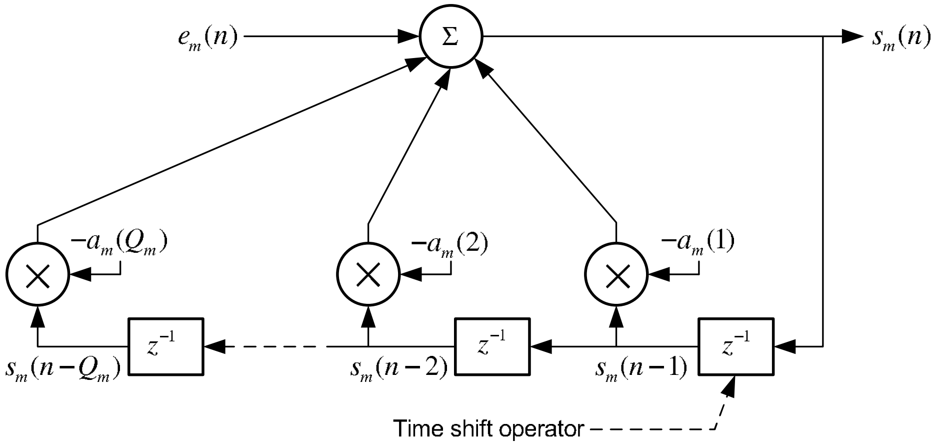

Figure 4.4. AR modelling of the m-th channel as an all-pole filter. The current output $s_m(n)$ depends on the Q_m most recent outputs, $s_m(n-1), \ldots, s_m(n-Q_m)$ and the current input, $e_m(n)$.

The AR order Q_m needs to be selected so as to appropriately approximate the signal PSD. Enough poles must be used to resolve all the peaks of the PSD with additional poles added to provide general spectral shaping and to approximate any notches in the PSD [178]. Too high a value of the AR model over-fits the signal and introduces spurious details such as false peaks into the PSD, whereas too low a value produces a PSD that is over-smoothed. Between these two extremes the minimum value of Q_m that adequately represents the signal being modelled is chosen. Determining this value is often based upon a goodness-of-fit term such as the prediction error power $\mathcal{E}(Q_m)$ combined with a cost function that penalizes some measure of the model complexity, i.e some monotonically growing function of Q_m. Indeed, since the fit of the model improves as Q_m increases, the prediction error power is a non-increasing function of Q_m (see Fig. 4.5) and the optimum order is rarely apparent form the inspection of $\mathcal{E}(Q_m)$ alone (Fig. 4.5). Order selection methods include the final prediction error (FPE) [2], the Akaike information (AIC) [3], and minimum description length (MDL) [140] criteria.

$$\mathrm{FPE}(Q_m) = \log\left(\mathcal{E}(Q_m)\right) + \log\left(\frac{\mathrm{N_{spt}} - Q_m - 1}{\mathrm{N_{spt}} + Q_m + 1}\right) \approx \log\left(\mathcal{E}(Q_m)\right) + \frac{2\left(Q_m+1\right)}{\mathrm{N_{spt}}} \tag{4.65}$$

$$\mathrm{AIC}(Q_m) = \log\left(\mathcal{E}(Q_m)\right) + \frac{2Q_m}{\mathrm{N_{spt}}} \tag{4.66}$$

$$\mathrm{MDL}(Q_m) = \log\left(\mathcal{E}(Q_m)\right) + Q_m\frac{\log(\mathrm{N_{spt}})}{\mathrm{N_{spt}}} \tag{4.67}$$

where the approximation (4.65) holds for $Q_m < \mathrm{N_{spt}}$. In practice, $\mathrm{N_{spt}}$ is at least ten times larger than Q_m [178].

Notice that each of the above mentioned criteria can be written in the form:

$$\log\left(\mathcal{E}(Q_m)\right) + \eta_{Q_m} Q_m$$

where η_{Q_m} is a penalization factor. Since the penalization factor associated with the MDL criterion is the largest, this criterion gives the smallest AR order (Fig. 4.5). In practice, the

MDL criterion is generally preferred [9]. In the framework of linear prediction, the MDL criterion takes the general form: log(error) + number of parameters × $\frac{\text{log(number of samples)}}{\text{number of samples}}$

The feature vector $\psi_{\text{AR}}(S)$ is composed of the AR coefficients associated to each channel:

$$\psi_{\text{AR}}(S) = \begin{pmatrix} a_1(1) & \dots & a_1(Q_1) & \dots & a_{\text{N}_e}(1) & \dots & a_{\text{N}_e}(Q_{\text{N}_e}) \end{pmatrix}^{\text{t}} \in \mathbb{R}^{\sum_m Q_m}$$

where $a_m(i)$ is the i-th AR coefficient and Q_m is the AR order associated to the m-th channel.

The AR mapping does not require to select a set of frequency bands and can lead to feature vectors whose dimensionality is smaller than that of the previous mappings. However, in BCI applications, the AR mapping presents an inconvenience residing in the fact that a direct connection between the AR coefficients and the power in a given frequency band is not evident [167]. Instead, this power is an intricate non-linear function of the AR coefficients. This makes difficult to explain the physiological mechanism that the subject actually uses to control an AR coefficient based BCI [175].

4.6 Non-stationary autoregressive mapping

The non-stationary autoregressive (NAR) mapping, denoted as ψ_{NAR} is built on the hypotheses of absence of coupling between the univariate components of EEG, and existence of a linear prediction model. As in the AR mapping, the hypothesis of absence of coupling permits to split the model in (4.50), into linear prediction models for each univariate component:

$$s_m(n) = -\sum_{i=1}^{Q_m} a_m(n,i) s_m(n-i) + \mathbf{e}_m(n) \tag{4.68}$$

where Q_m is the model order of the m-th channel.

When s_m is stationary (Section 4.5), the coefficients $a_m(n,i)$ are time independent and the prediction model can be interpreted as an all-pole (time-invariant) filter in which the filter output $s(n)$ depends on the weighted sum of its time-shifted versions (Fig. 4.4). In the non-stationary case, the linear prediction model can be interpreted as a time-varying filter which introduces frequency shifts in addition to time-shifts [77, 100, 136]. Hence, the linear prediction model in (4.68) becomes (see Fig. 4.6):

$$s_m(n) = -\sum_{i=1}^{Q_m} \sum_{u=-U_m}^{U_m} \tilde{a}_m(i,u) \exp\left(-j\frac{2\pi un}{\text{N}_{\text{spt}}}\right) s_m(n-i) + \mathbf{e}_m(n) \tag{4.69}$$

where U_m is the spectral order associated with s_m. Clearly, this equation is equivalent to (4.68) with:

$$a_m(n,i) = \sum_{u=-U_m}^{U_m} \tilde{a}_m(i,u) \exp\left(-j\frac{2\pi un}{\text{N}_{\text{spt}}}\right) \tag{4.70}$$

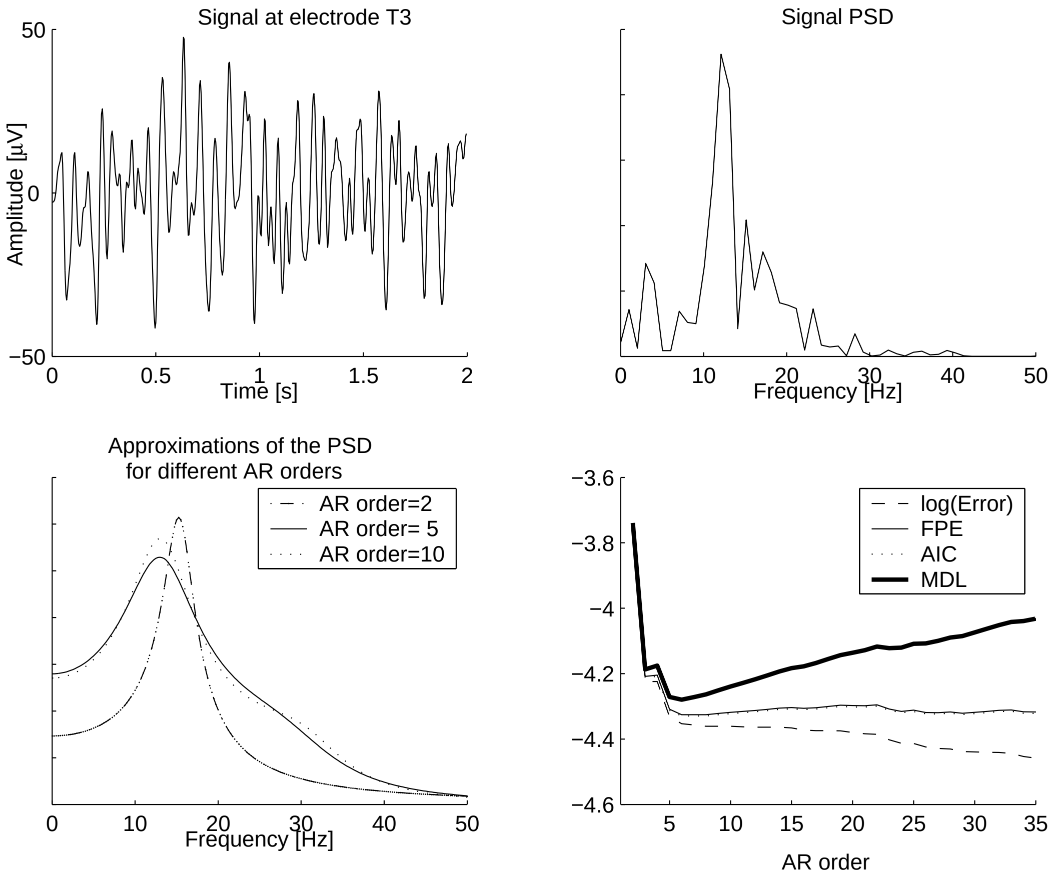

Figure 4.5. Autoregressive estimation of the PSD. *Top left*: Signal under study (the same as the one in Fig. 4.2). *Top right*: estimated PSD computed using the Welch method (Section 4.3). *Bottom left*: PSD approximations for different AR orders. As it can be seen, an AR order of 2 leads to the PSD over-smoothing, for AR orders equal to 5 and 10, the PSD is relatively well approximated. *Bottom right*: Logarithm of the prediction error power and the three AR order selection, i.e. final prediction error (FPE), Akaike information (AIC) and minimum description length (MDL), criteria are represented. The prediction error slowly decreases as the AR increases making this parameter, considered alone, not suitable to adequately chose the AR order. Because of the penalization of too large values of the AR order, the order selection criteria present an optimum which is more evident in the MDL criterion as it has the largest penalization factor (see Equations 4.65 to 4.67).

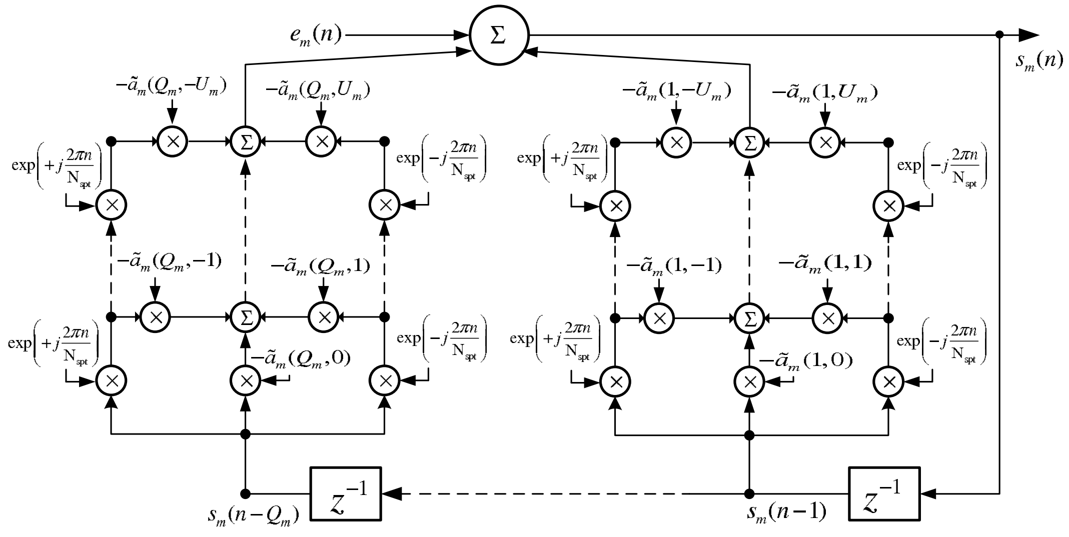

Figure 4.6. Block diagram of the non-stationary autoregressive model of the m-th channel. The current output $s_m(n)$ depends on frequency shifted (frequency shifts are introduced via the products by: $\exp\left(\pm j\frac{2\pi n}{\mathrm{N_{spt}}}\right)$) versions of the Q_m most recent outputs, $s_m(n-1),\ldots,s_m(n-Q_m)$ and the current input $e_m(n)$.

The coefficients $\tilde{a}_m(i,u)$ (NAR coefficients) can be determined by minimizing the prediction error power:

$$\begin{aligned}\mathcal{E}(Q_m,U_m) &= \frac{1}{\mathrm{N_{spt}}}\sum_{n=0}^{\mathrm{N_{spt}}-1}\mathbf{e}_m^2(n) \qquad (4.71)\\ &= \frac{1}{\mathrm{N_{spt}}}\sum_{n=0}^{\mathrm{N_{spt}}-1}\left(s_m(n)+\sum_{i=1}^{Q_m}\sum_{u=-U_m}^{U_m}\tilde{a}_m(i,u)\exp\left(-j\frac{2\pi un}{\mathrm{N_{spt}}}\right)s_m(n-i)\right)^2\end{aligned}$$

Taking the derivatives of $\mathcal{E}(Q_m,U_m)$ with respect to the NAR coefficients and setting them to zero, yields:

$$\begin{aligned}&\sum_{i'=1}^{Q_m}\sum_{u'=-U_m}^{U_m}\tilde{a}_m(i',u')\frac{1}{\mathrm{N_{spt}}}\sum_{n=0}^{\mathrm{N_{spt}}-1}s_m(n-i')s_m(n-i)\exp\left(-j\frac{2\pi(u+u')n}{\mathrm{N_{spt}}}\right)\\ &\qquad+\frac{1}{\mathrm{N_{spt}}}\sum_{n=0}^{\mathrm{N_{spt}}-1}s_m(n)s_m(n-i)\exp\left(-j\frac{2\pi un}{\mathrm{N_{spt}}}\right)=0 \text{ for }\left\{\begin{array}{c}1\leqslant i\leqslant Q_m\\ -U_m\leqslant u\leqslant U_m\end{array}\right.\end{aligned}$$

Taking expectations on both sides in the above equation and using the definition of expected ambiguity function (4.29) yields:

$$\sum_{i'=1}^{Q_m}\sum_{u'=-U_m}^{U_m}\tilde{a}_m(i',u')\mathrm{E}_{\mathbf{s}_m}\left[A_{\mathbf{s}_m}(i-i',u-u')\right]=-\mathrm{E}_{\mathbf{s}_m}\left[A_{\mathbf{s}_m}(i,u)\right] \qquad (4.72)$$

This set of linear equations generalize the Yule-Walker ones (4.62) to the non-stationary case. The total number of NAR coefficients is equal to: $Q_m\,(2U_m+1)$. For slowly time-varying $a_m(n,i)$, a small U_m suffices to characterize the frequency shifts [77].

The model and spectral orders Q_m and U_m can be selected similarly to the AR order in Section 4.5. Namely, choosing those values that make the MDL criterion (4.67) minimum:

$$(Q_m, U_m) = \underset{Q_m^*, U_m^*}{\operatorname{argmin}} \left(\log \left(\mathcal{E}(Q_m^*, U_m^*) \right) + Q_m^* \left(2U_m^* + 1 \right) \frac{\log(\mathrm{N_{spt}})}{\mathrm{N_{spt}}} \right) \tag{4.73}$$

The estimation of $\mathrm{E}_{\mathbf{s}_m}\left[A_{\mathbf{s}_m}(i,u)\right]$ is obtained by segmenting s_m into N_β (possibly) overlapping N-length blocks and computing the average of the blocks ambiguity functions:

$$\mathrm{E}_{\mathbf{s}_m}\left[A_{\mathbf{s}_m}(i,u)\right] = \frac{1}{\mathrm{N}_\beta} \sum_{\beta=1}^{\mathrm{N}_\beta} \frac{1}{N} \sum_{n=0}^{N-1} s_{m,\beta}(n-i) s_{m,\beta}(n) \exp\left(-j\frac{2\pi n u}{N}\right) \tag{4.74}$$

where $s_{m,\beta}(n)$ is the n-th sample of the β-th block.

The feature vector $\psi_{\mathrm{NAR}}(S)$ is composed of the NAR coefficients associated to each channel:

$$\psi_{\mathrm{NAR}}(S) = \begin{pmatrix} \tilde{a}_1(1,-U_1) & \ldots & \tilde{a}_m(i,u) & \ldots & \tilde{a}_{\mathrm{N}_e}(1,-U_{\mathrm{N}_e}) & \ldots & \tilde{a}_{\mathrm{N}_e}(Q_{\mathrm{N}_e}, U_{\mathrm{N}_e}) \end{pmatrix}^{\mathrm{t}}$$

$$\psi_{\mathrm{NAR}}(S) \in \mathbb{R}^{\sum_m Q_m(2U_m+1)}$$

Since for EEG signals, the parameters $a_m(n,i)$ of the general linear prediction model (4.68) slowly change in time [129, 146] the spectral orders: $U_1, \ldots, U_{\mathrm{N}_e}$ are relatively small (up to three, see Chapter 6). This makes the NAR particularly well suited for BCI applications. However, alike the AR mapping (Section 4.5), the physiological interpretation of the NAR coefficients is difficult since there is no direct link between them and the observed signals.

4.7 Multivariate autoregressive mapping

The multivariate autoregressive (MVAR) mapping, denoted as ψ_{MVAR}, is built on the hypotheses of ergodicity and existence of a linear prediction model. Because of the ergodicity, the matrices $\mathbf{A}(n,i)$ in the general prediction model in (4.50) are time independent. Thus, the EEG-trial S can be characterized using the MVAR model:

$$S\left(n\right) = -\sum_{i=1}^{Q} \mathbf{A}(i) S(n-i) + \mathbf{e}(n) \tag{4.75}$$

where the $\mathbf{A}(i)$ are $\mathrm{N}_e \times \mathrm{N}_e$ matrices, $S(n) = \begin{pmatrix} s_1(n) & \cdots & s_{\mathrm{N}_e}(n) \end{pmatrix}^{\mathrm{t}}$, and $\mathbf{e}(n) = \begin{pmatrix} e_1(n) & \cdots & e_{\mathrm{N}_e}(n) \end{pmatrix}^{\mathrm{t}}$ is the prediction error vector at time n.

As in the two previous mappings, the elements of matrices $\mathbf{A}(i)$ are determined by minimizing the prediction power error:

$$\begin{aligned} \mathcal{E}(Q) &= \frac{1}{\mathrm{N_{spt}}} \sum_{n=0}^{\mathrm{N_{spt}}-1} \mathbf{e}^{\mathrm{t}}(n)\mathbf{e}(n) \qquad (4.76) \\ &= \frac{1}{\mathrm{N_{spt}}} \sum_{n=0}^{\mathrm{N_{spt}}-1} \left(S^{\mathrm{t}}(n) + \sum_{i=1}^{Q} S^{\mathrm{t}}(n-i)\mathbf{A}^{\mathrm{t}}(i) \right) \left(S(n) + \sum_{i=1}^{Q} \mathbf{A}(i) S(n-i) \right) \end{aligned}$$

Taking the derivatives of $\mathcal{E}(Q)$ with respect to the elements of the matrices $\mathbf{A}(i)$ yields the multivariate Yule-Walker equations [85, 134]:

$$-\left(\begin{array}{ccc}\mathbf{K}(1) & \ldots & \mathbf{K}(Q)\end{array}\right)=\left(\begin{array}{ccc}\mathbf{A}(1) & \ldots & \mathbf{A}(Q)\end{array}\right)\tilde{\mathbf{K}} \tag{4.77}$$

where $\mathbf{K}(\tau)=\sum_{n=0}^{\mathrm{N_{spt}}-1} S(n-\tau)S^{\mathrm{t}}(n)$ and

$$\tilde{\mathbf{K}}=\left(\begin{array}{cccc}\mathbf{K}(0) & \mathbf{K}(1) & \cdots & \mathbf{K}(Q-1)\\ \mathbf{K}^{\mathrm{t}}(1) & \mathbf{K}(0) & \cdots & \mathbf{K}(Q-2)\\ \vdots & \vdots & & \vdots\\ \mathbf{K}^{\mathrm{t}}(Q-1) & \mathbf{K}^{\mathrm{t}}(Q-2) & \cdots & \mathbf{K}(0)\end{array}\right)$$

The matrices $\mathbf{A}(1),\ldots,\mathbf{A}(Q)$ can be found by inverting the $Q\mathrm{N}_e \times Q\mathrm{N}_e$ matrix $\tilde{\mathbf{K}}$ in (4.77). However, (4.77) can be more efficiently solved by applying a generalized version of the Levinson recursions [173].

The frequency domain form of the MVAR model is obtained by taking the Z-transform on both sides in (4.75) and evaluating it in the unit circle in the z-plane, i.e. at $z=\exp\left(j\frac{2\pi\mathrm{f}}{\mathrm{f}_z}\right)$ where f_s is the sampling frequency. Thus, the MVAR frequency-domain model is:

$$\left(\begin{array}{c}\hat{s}_1(\mathrm{f})\\ \vdots\\ \hat{s}_{\mathrm{N}_e}(\mathrm{f})\end{array}\right)=\left(\begin{array}{ccc}H_{1,1}(\mathrm{f}) & \cdots & H_{1,\mathrm{N}_e}(\mathrm{f})\\ \vdots & & \vdots\\ H_{\mathrm{N}_e,1}(\mathrm{f}) & \cdots & H_{\mathrm{N}_e,\mathrm{N}_e}(\mathrm{f})\end{array}\right)\left(\begin{array}{c}\hat{e}_1(\mathrm{f})\\ \vdots\\ \hat{e}_{\mathrm{N}_e}(\mathrm{f})\end{array}\right) \tag{4.78}$$

where $\hat{s}_m(\mathrm{f})$ is the spectrum of the m-th channel and $\hat{e}_m(\mathrm{f})$ is the spectrum of the m-th prediction error. The $\hat{e}_1(\mathrm{f}),\ldots,\hat{e}_{\mathrm{N}_e}(\mathrm{f})$ can be thought of as the input spectra which are filtered by the transfer functions $H_{m1,m2}(\mathrm{f})$ to produce the outputs $\hat{s}_1(\mathrm{f}),\ldots,\hat{s}_{\mathrm{N}_e}(\mathrm{f})$ (see Fig. 4.7). Since $H_{m1,m2(\mathrm{f})}$ is different from $H_{m2,m1}(\mathrm{f})$, the transfer function: $H_{m1,m2(\mathrm{f})}$ is a sort of "directed" intra-spectrum from the $m2$-th channel to the $m1$-th one [145].

The model order Q determines the shape of the transfer functions $H_{m1,m2}(\mathrm{f})$. In fact, higher orders imply more peaks in the transfer functions (see Fig. 4.8). To determine the optimal order Q we use the MDL criterion (Section 4.5). Thus, the optimal Q is selected as:

$$Q=\underset{Q*}{\operatorname{argmin}}\left(\log((Q))+Q\mathrm{N}_e\frac{\log(\mathrm{N_{spt}}\mathrm{N}_e)}{\mathrm{N_{spt}}}\right) \tag{4.79}$$

The feature vector ψ_{MVAR} is composed of the elements in matrices $\mathbf{A}(1),\ldots,\mathbf{A}(Q)$:

$$\psi_{\mathrm{MVAR}}(S)=\left(\begin{array}{ccc}\ddot{\mathbf{A}}(1) & \ldots & \ddot{\mathbf{A}}(Q)\end{array}\right)^{\mathrm{t}}\in\mathbb{R}^{Q\mathrm{N}_e^2}$$

where the notation $\ddot{\mathbf{A}}$ indicates that the elements in $\mathbf{A}$ are taken column-wise and rearranged in a single row.

The MVAR mapping was used to determine the spreading of brain activity in a defined frequency band by exploiting the concept of "directed" intra-spectrum [145] that we mentioned earlier. As in the two previous mappings a direct connection between physiological

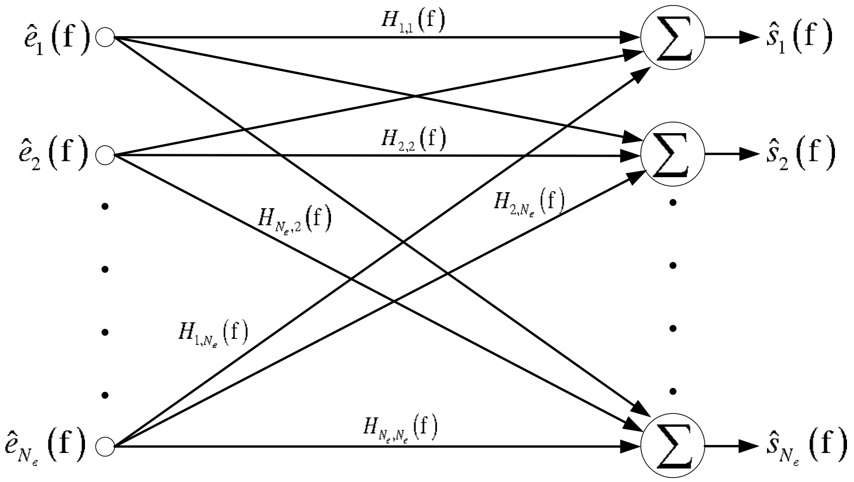

Figure 4.7. Frequency domain interpretation of a multivariate autoregressive model. In this interpretation, the inputs are the source spectra: $\hat{e}_1(\mathrm{f}),\ldots,\hat{e}_{\mathrm{N_e}}(\mathrm{f})$, the outputs the observed signal spectra: $\hat{s}_1(\mathrm{f}),\cdots,\hat{s}_{\mathrm{N_e}}(\mathrm{f})$ and the connection weights are filters that determine the contribution of each source to the observed signals. Generally $H_{m1,m2}(\mathrm{f}) \neq H_{m2,m1}(\mathrm{f})$ then, the transfer function $H_{m1,m2}(\mathrm{f})$ can be interpreted as the "directed" cross-spectrum from the $m2$-th channel to the $m1$-th one.

concepts and the MVAR coefficients does not exist. However, the representation of the transfer functions as in Fig. 4.8 allows us to evaluate the interaction between channels at different frequencies which appears to be non-symmetrical.

4.8 Synchronization mapping

The synchronization mapping, denoted as ψ_{Y} is built on the hypothesis of weak coupling. Under this hypothesis an EEG-trial is characterized by the degree of synchronization, in narrow frequency bands, between its channels (see Section 4.2.7).

The synchronization between two EEG channels s_{m1} and s_{m2}, in a frequency band $\mathrm{B} = [\mathrm{f}_1;\mathrm{f}_2]$, is determined by the phase locking (4.51) between $s_{m1}^{(\mathrm{B})}$ and $s_{m2}^{(\mathrm{B})}$ which are respectively, the signals resulting from the filtering of s_{m1} and s_{m2} in B.

The phase of a signal s can be determined by means of its complex analytic form [82] defined as:

$$\bar{s}(n) = s(n) + jH_s(n) \tag{4.80}$$

where H_s is the Hilbert transform[1] of s. The analytic form can be further decomposed as: $\bar{s}(n) = \mathcal{A}_s(n)\exp\left(j\varphi_s(n)\right)$, where $\mathcal{A}_s(n)$ is the instantaneous amplitude and $\varphi_s(n)$ the instantaneous phase of s.

The degree of phase locking between s_{m1} and s_{m2} in the frequency band B is given by

[1] The Hilbert transform can be determined using standard methods as presented in [82]

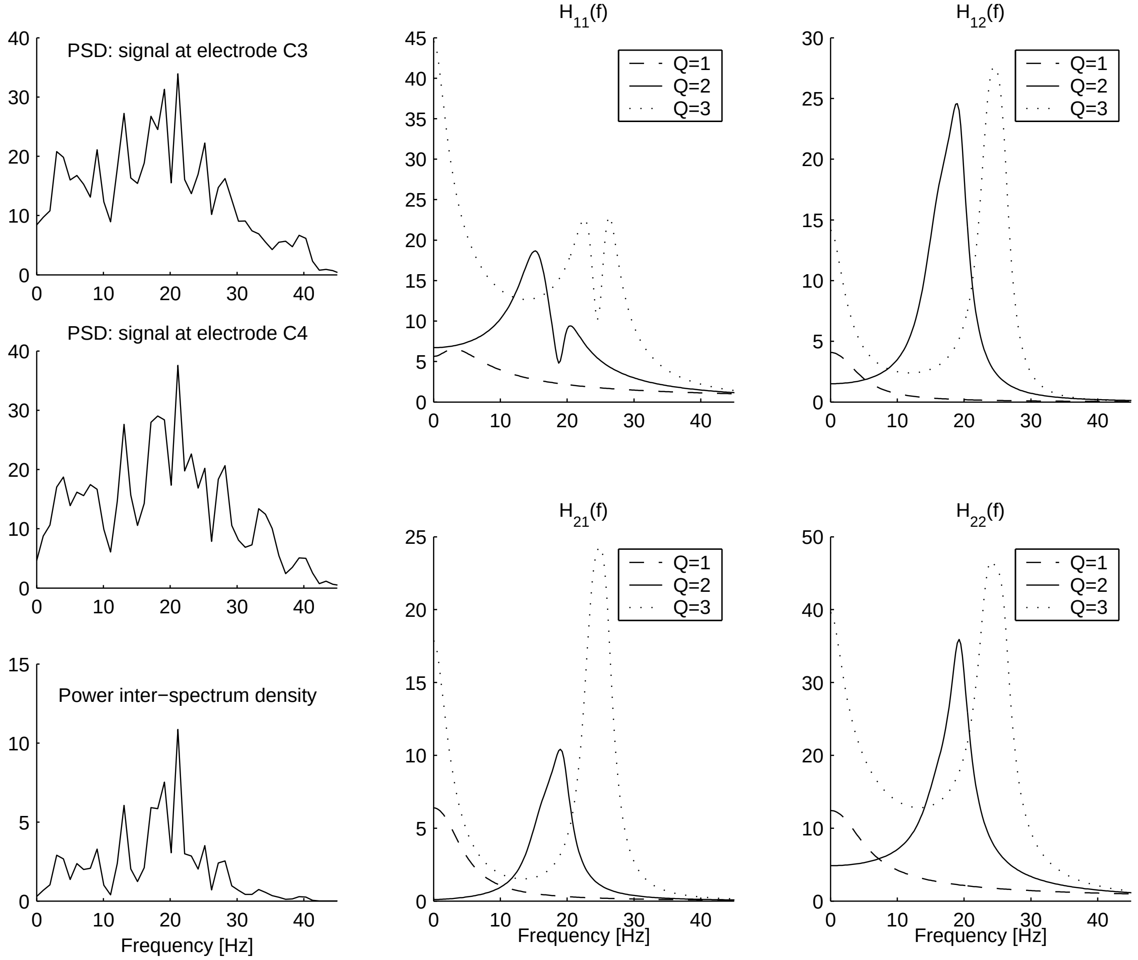

Figure 4.8. Multivariate autoregressive model for two EEG channels (C3 and C4). On the left, the power spectral and inter-spectral densities and on the right, the corresponding transfer functions. As it can be seen the higher Q the more peaks appear in the transfer function. The optimal order Q is determined using the MDL criterion (4.79)
The signals under study are the same than that in Fig. 4.3

the modulus of the average of the set of complex relative phases [111]:

$$\left\{ \exp\left(j\varphi^{(\mathrm{B})}_{s_{m1},s_{m2}}(n)\right) \middle| \, n = 0, \ldots, \mathrm{N}_{\mathrm{spt}} - 1 \right\}$$

where $\varphi^{(\mathrm{B})}_{s_{m1},s_{m2}}(n) = \varphi^{(\mathrm{B})}_{s_{m1}} - \varphi^{(\mathrm{B})}_{s_{m2}}$ is the relative phase of $s^{(\mathrm{B})}_{m1}$ and $s^{(\mathrm{B})}_{m2}$.

The synchronization between s_{m1} and s_{m2} in the frequency band B is (see Fig. 4.9):

$$\mathrm{Y}(m1, m2, \mathrm{B}) = \left| \frac{1}{\mathrm{N}_{\mathrm{spt}}} \sum_{n=0}^{\mathrm{N}_{\mathrm{spt}}-1} \exp\left(j\varphi^{(\mathrm{B})}_{s_{m1},s_{m2}}(n)\right) \right| \tag{4.81}$$

One can easily verify that $\mathrm{Y}(m1, m2, \mathrm{B})$ varies from zero, when the complex relative phases are uniformly distributed in the complex unit circle, to one when the complex relative phases are all equal.

The feature vector ψ_{Y} is determined by selecting N_{B} frequency bands $\{\mathrm{B}_1, \ldots, \mathrm{B}_{\mathrm{N}_{\mathrm{B}}}\}$, computing the synchronization for each frequency band and pair of EEG channels and grouping the results into an $\frac{\mathrm{N}_e(\mathrm{N}_e-1)}{2}\mathrm{N}_{\mathrm{B}}$ dimensional vector:

$$\psi_{\mathrm{Y}}(S) = \begin{pmatrix} \mathrm{Y}(1,2,\mathrm{B}_1) & \ldots & \mathrm{Y}(m1,m2,\mathrm{B}_i) & \ldots & \mathrm{Y}(\mathrm{N}_e-1,\mathrm{N}_e,\mathrm{B}_{\mathrm{N}_{\mathrm{B}}}) \end{pmatrix}^{\mathrm{t}} \in \mathbb{R}^{\frac{\mathrm{N}_e(\mathrm{N}_e-1)}{2}\mathrm{N}_{\mathrm{B}}}$$

Synchronization appears to be a basic mechanism for neuronal information processing within a brain area as well as for communication between different brain areas. The identification of phase locking between two EEG channels can provide useful insight into the cooperation mechanisms between the underlying neuronal groups during the execution of mental activities [16, 96, 98]. It is worth mentioning that synchronization, in contrast to coherence, can be high even if the amplitudes are uncorrelated [16, 96].

4.9 Summary

The characterization of EEG is based on the analysis of the generalized interactions between the EEG channels. By assuming some hypotheses on the properties of EEG signals, we derived different mappings from the EEG-trial set into feature spaces whose characteristics are determined by the hypotheses that define the corresponding mapping. Since a single mapping appears to be insufficient for the recognition of all the MAs that are used to control the BCI (see Chapter 6), an optimal association between an MA and a mapping should be established.

The following hypotheses were used: stationarity and ergodicity, absence of coupling between the EEG channels, existence of a linear prediction model and weak coupling between the EEG channels. The way in which these hypotheses are combined to obtain the stationary PSD, coherence, autoregressive, non-stationary autoregressive, multivariate autoregressive and synchronization mappings is depicted in Fig. 4.10.

The dimensionality of the feature vector spaces associated to each mapping is reported in Table 4.1. In general, the mappings built on the hypothesis of existence of a linear prediction model generate feature vectors with lower dimensionality. Furthermore such mappings do

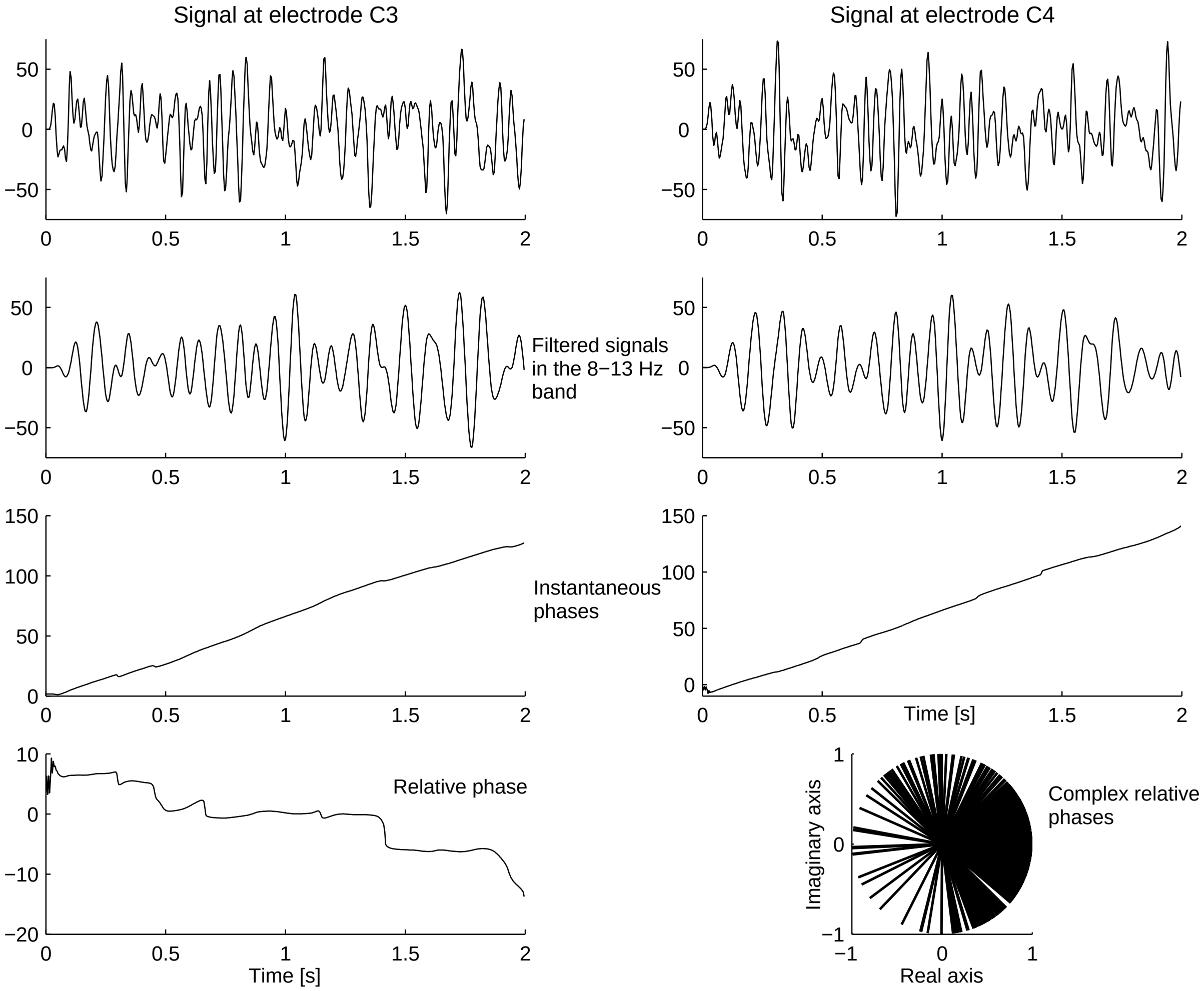

Figure 4.9. Estimation of the synchronization between two EEG channels in the alpha band. The signals are first filtered in the alpha band, then their instantaneous phases and relative phase are computed. In the lower right panel we represent the complex relative phases in the complex unit circle. The value of the synchronization determined using (4.81) is: 0.759. The signals under study are those in Fig. 4.3.

not require the choice of given frequency bands. However, the features that compose them are not directly connected to specific brain events (e.g. the power in a given frequency band or morphological signal properties). Thus, in a BCI based on feature vectors based on general autoregressive features it can be difficult to understand the type of physiological mechanisms that are actually used to control the BCI.

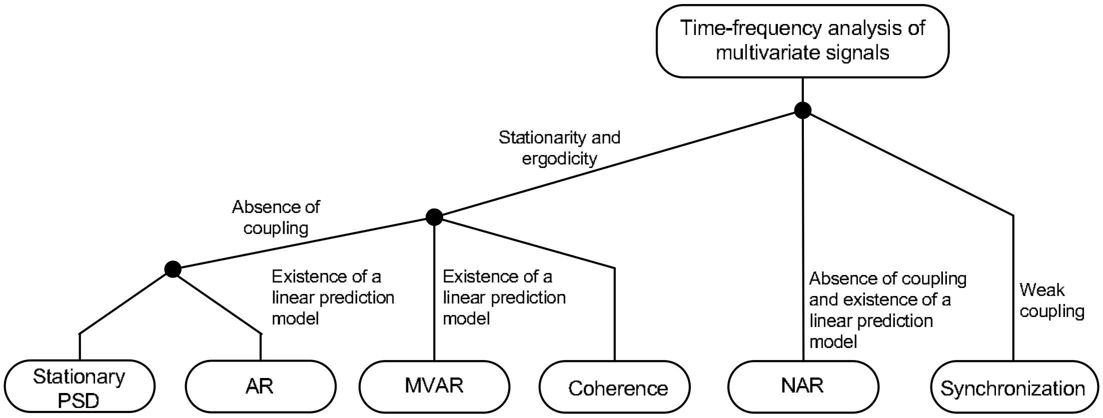
Time-frequency analysis of multivariate signals
Stationarity and ergodicity
Absence of coupling
Existence of a linear prediction model
Existence of a linear prediction model
Absence of coupling and existence of a linear prediction model
Weak coupling
Stationary PSD
AR
MVAR
Coherence
NAR
Synchronization

Figure 4.10. Derivation of the mappings from hypotheses on the nature of EEG trials.

Mapping	**Dimension of feature vectors**	**Typical dimension** *
Stationary PSD (ψ_{P})	$\mathrm{N}_e\mathrm{N}_\mathrm{B}$	160
Coherence (ψ_{C})	$\frac{\mathrm{N}_e(\mathrm{N}_e-1)}{2}\mathrm{N}_\mathrm{B}$	1200
Autoregressive (ψ_{AR})	$\sum_{m=1}^{\mathrm{N}_e} Q_m$	32
Non-stationary autoregressive (ψ_{NAR})	$\sum_{m=1}^{\mathrm{N}_e} Q_m\,(2U_m+1)$	160
Synchronization (ψ_{Y})	$\frac{\mathrm{N}_e(\mathrm{N}_e-1)}{2}\mathrm{N}_\mathrm{B}$	1200
Multivariate autoregressive (ψ_{MVAR})	$Q\mathrm{N}_e^2$	512

Table 4.1. Dimensionality of the feature vectors associated to each mapping. (*) Typical values for the parameters in the third column are (see Chapter 6): $\mathrm{N}_e = 16, \mathrm{N}_\mathrm{B} = 10, Q_{m=1,\ldots,\mathrm{N}_e} = 2$, $U_{m=1,\ldots,\mathrm{N}_e} = 2$, and $Q = 2$. Using such values it appears that the coherence and synchronization mappings produce the highest dimensional feature vectors whereas the autoregressive mapping produces the lowest dimensional one. Thus, unless a small number of frequency bands is considered, the mappings built on the hypothesis of existence of a linear prediction model produce the feature vectors with the smallest number of elements.

Pattern recognition 5

"The purpose of models is not to fit the data
but to sharpen the questions"
Samuel Karlin

5.1 Introduction

In the previous chapter we presented different mappings from the EEG-trial set into a feature vector space $\mathcal{X}$ that is suitable for the recognition of a given MA in the controlling set. We pointed out that the choice of the optimal mapping (which determines $\mathcal{X}$) depends on the subject and the mental activity. In this chapter we assume that the choice of the optimal mapping is done according to an optimality criterion (see Chapter 6) and concentrate on the recognition process.

Let Ω be the set of all possible EEG-trials and Ω_k the set of EEG-trials produced during the performance of mental activity MA_k. The optimal mapping for the recognition of MA_k, denoted as $\psi^{(k)}$ maps Ω and Ω_k into the feature vector space (induced by $\psi^{(k)}$) $\mathcal{X}_k$ and the target set X_k respectively (see Fig. 5.1). Our goal is to estimate a measure of the likelihood, denoted as $f_k(x)$ that a feature vector $x \in \mathcal{X}_k$ belongs to X_k. We call $f_k(\cdot)$ the membership function associated with the mental activity MA_k.

As shown in Fig. 5.2, the feature extraction module delivers to the pattern recognition one, N_{MA} feature vectors denoted as $x^{(1)}, \ldots, x^{(\text{N}_{\text{MA}})}$, which are computed by applying the optimal mappings $\psi^{(1)}, \ldots, \psi^{(\text{N}_{\text{MA}})}$ to an EEG-trial $S \in \Omega$. The pattern recognition module in turn computes N_{MA} membership functions: $f_1(x^{(1)}), \ldots, f_{\text{N}_{\text{MA}}}(x^{(\text{N}_{\text{MA}})})$ that are grouped into a vector of memberships $\vec{f}$ that is sent to the action generation module which decides on the action that the BCI executes (see Chapter 6).

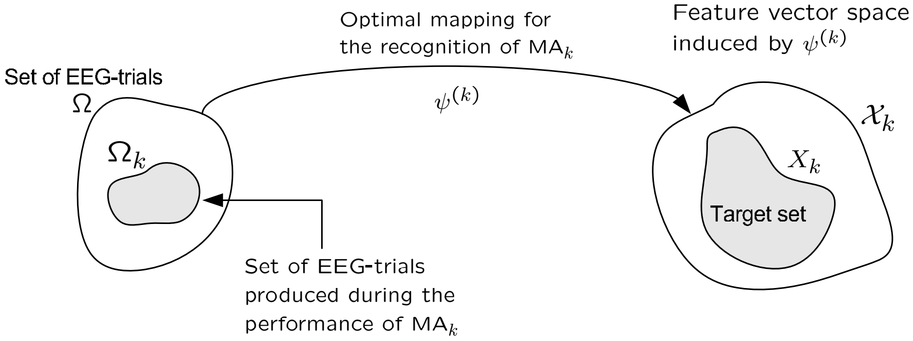

Figure 5.1. The optimal mapping for the recognition of MA$_k$, $\psi^{(k)}$ maps the set of EEG-trials Ω into a feature vector space $\mathcal{X}_k$. In $\mathcal{X}_k$ a feature vector is characterized with respect to its membership to the target set X_k (i.e. the set of feature vectors produced during the performance of MA$_k$)

Each membership function f_k is learned in a supervised way, i.e. the exact membership of a given set (the training set) of feature vectors belonging to $\mathcal{X}_k$ is known and f_k is estimated so as to minimize the discrepancy between the memberships computed by f_k and the real ones. Note that the exact membership of an element in the training set can only take two values: belongs or not to the target set. In contrast, the range of $f_k(\cdot)$ is in the real numbers, i.e. different degrees of membership exist.

The shape of target sets can change over time as a consequence of environmental factors or the subject's state of mind (fatigue, stress, etc [29]). Moreover, as the subject acquires more experience in using the BCI his brain dynamics may exhibit some changes resulting from his adaptation to the BCI [36]. Such adaptation induces changes on the target sets.

Thus, a static learning approach, in which the membership functions remain constant is clearly suboptimal. Instead, they need to be continuously adapted according to a dynamical learning strategy in which they are updated as new training data become available while progressively forgetting the contribution of old data.

In the following we present the methods to learn and dynamically update the membership functions. These methods are based on the statistical learning theory [166], kernel methods [147] and support vector machine learning algorithms [23]. Instead of introducing the support vector machine learning concepts using the classical large margin classifier approach [23, 113, 147, 165] we focus on the concept of loss and risk to derive the learning and dynamical updating algorithms from the same framework.

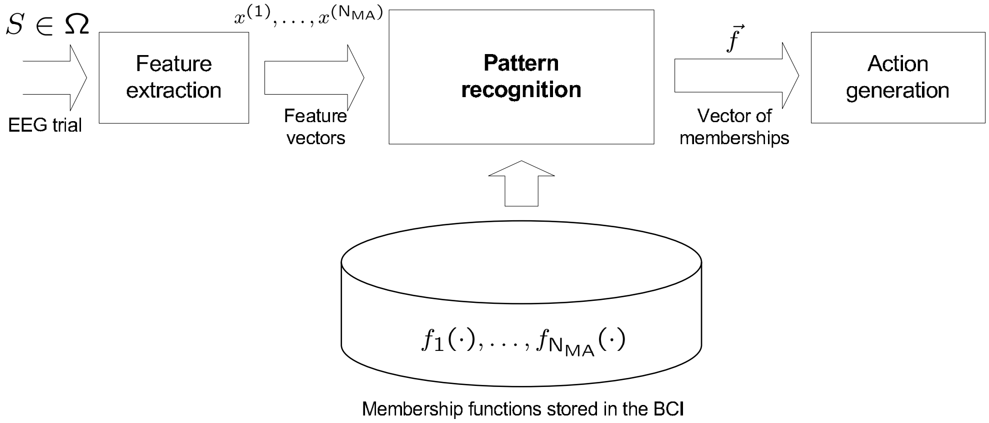

Figure 5.2. The feature extraction module delivers to the pattern recognition one, N_{MA} feature vectors: $x^{(1)}, \ldots, x^{(N_{MA})}$, where $x^{(k)} = \psi^{(k)}(S)$, S is the current EEG-trial, and $\psi^{(k)}$ is the optimal mapping for the recognition of mental activity MA_k.
The pattern recognition module computes N_{MA} membership functions: $f_1(x^{(1)}), \ldots, f_{N_{MA}}(x^{(N_{MA})})$ which are grouped into a vector of memberships $\vec{f}$, that is sent to the action generation module which decides on the action that the BCI executes.
The membership functions can be thought of as comparison models for the mental activities that are used to control the BCI. Such models are subject dependent and continuously updated.

5.2 Membership functions

The vector of memberships $\vec{f}$ that is sent to the action generation module (see Fig. 5.2) is a vector field that maps $\mathcal{X}_1 \times \ldots \times \mathcal{X}_{N_{MA}}$ into $\mathbb{R}^{N_{MA}}$ and is defined as follows

$$\vec{f}(x_1, \ldots, x_{N_{MA}}) = \begin{pmatrix} f_1(x^{(1)}) \\ \vdots \\ f_k(x^{(k)}) \\ \vdots \\ f_{N_{MA}}(x^{(N_{MA})}) \end{pmatrix}$$

where $x^{(k)} \in \mathcal{X}_k$ and $f_k(x^{(k)})$ is the membership function associated with MA_k. The ideal f_k (i.e. the error free membership function) is such that:

$$\begin{cases} f_k(x^{(k)}) + b_k \geqslant \rho_k & \text{if } x^{(k)} \in X_k \\ f_k(x^{(k)}) + b_k \leqslant -\rho_k & \text{if } x^{(k)} \notin X_k \end{cases} \tag{5.1}$$

where $\rho_k \geqslant 0$ and $b_k \in \mathbb{R}$ are the threshold and the offset of $f_k(\cdot)$ respectively (see Fig. 5.3). We call f_k, ρ_k, and b_k the membership parameters associated with MA_k whose estimation from observed data is the object of next section.

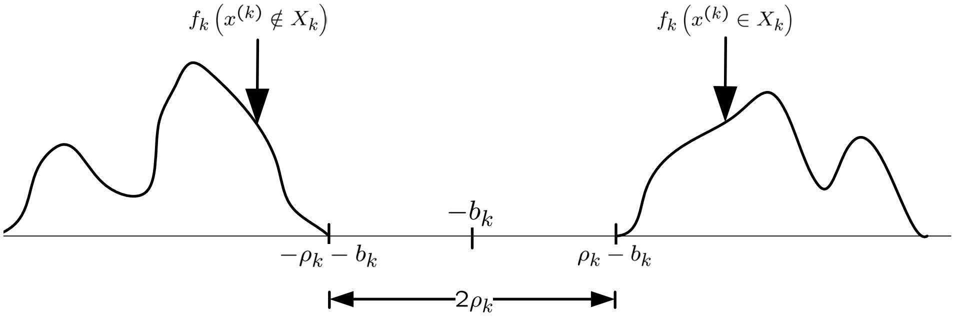
$f_k\left(x^{(k)} \notin X_k\right)$
$f_k\left(x^{(k)} \in X_k\right)$
$-b_k$
$-\rho_k - b_k$
$\rho_k - b_k$
$2\rho_k$

Figure 5.3. Distribution of the ideal membership function, f_k with respect to its target set X_k. According to (5.1) the membership values of the feature vectors in X_k are located right from: $\rho_k - b_k$ and that of feature vectors not belonging to X_k are located left from: $-\rho_k - b_k$.

In the practical implementation (see Chapter 6) it is more advantageous to consider a normalized form of the membership function. The normalized membership is defined as:

$$\zeta_k(x^{(k)}) = \frac{f_k(x^{(k)}) + b_k}{\rho_k} \tag{5.2}$$

one can easily verify that: $\zeta_k(x^{(k)}) \geqslant 1$ if $x^{(k)} \in X_k$, and $\zeta_k(x^{(k)}) \leqslant -1$ if $x^{(k)} \notin X_k$. In this chapter, as we seek to determine the membership parameters we consider the form in (5.1).

5.3 Estimation of the membership parameters

The membership parameters are estimated in a supervised way, i.e. using a set of feature vectors (training vectors) for which the exact membership is known. The training set is composed of the training vectors and their respective membership values. We denote as $\mathcal{S}_{\text{tr-k}}$ the training set for the estimation of the membership parameters associated with MA$_k$.

$$\mathcal{S}_{\text{tr-k}} = \left\{ \left(x_l^{(k)}, y_l^{(k)}\right) \middle| \, x_l^{(k)} \in \mathcal{X}_k,\ y_l^{(k)} \in \{-1, +1\},\ \text{and } l = 1, 2, \ldots, \mathrm{L} \right\}$$

where the membership value (or label) $y_l^{(k)}$ of $x_l^{(k)}$ is defined as:

$$y_l^{(k)} = \begin{cases} +1 & \text{if } x_l^{(k)} \in X_k \\ -1 & \text{otherwise} \end{cases}$$

We assume that the training set was independently drawn from a probability density function $p_k(x^{(k)}, y^{(k)})$.

From the definition of the ideal membership function (5.1) and Fig. 5.3 it comes out that the ideal distribution of the product $y^{(k)}\left(f_k(x^{(k)}) + b_k\right)$ (we call it product-distribution) should be concentrated right from ρ_k. However, the product-distributions corresponding to estimates of the membership parameters can spread left from ρ_k. Thus, the quality of an estimation is characterized by the deviation of its product-distribution from the ideal one. Such deviation is given by the risk functional presented in Section 5.3.2.

Henceforth, we adopt the following notation conventions. First, in order to simplify the notation we remove the index k from every parameter. Indeed, the concepts behind the estimation of ρ_k, b_k, f_k are identical for every MA_k (thus, the feature vector space, the target set and the training set are denoted as $\mathcal{X}$, X, and $\mathcal{S}_{\mathrm{tr}}$ respectively). Second, we use the $^{\mathrm{ideal}}$ superscript to denote the ideal value of the membership function, i.e. f should be interpreted as an estimate of f^{ideal}.

5.3.1 Loss function

The loss function associated with an estimation (ρ, b, f) of the membership parameters, maps $\mathcal{X} \times \{-1; +1\} \times \mathbb{R}$ into $\mathbb{R}$, and is defined as:

$$c(x, y, f(x)) = \begin{cases} -\nu\rho & y\,(f(x) + b) \geqslant \rho \\ \gamma(y\,(f(x) + b)) & y\,(f(x) + b) < \rho \end{cases} \tag{5.3}$$

where $\gamma(\cdot)$ is a derivable monotonically decreasing function in $]-\infty; \rho[$ such that: $\gamma(\rho) = -\nu\rho$ in order to ensure the continuity of the loss function.

A non-zero constant loss of $-\nu\rho$ is assigned to the (non-penalized) zone located right from the straight line: $yf(x) + yb = \rho$. The reason for this is that ν permits to establish a bound on the membership errors (or recognition errors) in the training set (see Section 5.3.5).

As shown in Fig. 5.4, the penalized zone, located left from $yf(x) + yb = \rho$ is penalized by the function $\gamma(\cdot)$. We consider $\gamma(\cdot)$ as a polynomial function of degree $q \geqslant 1$ defined as:

$$\gamma(u) = (\rho - u)^q - \nu\rho \tag{5.4}$$

As shown in Sections 5.3.3 and 5.6, the penalty degree q plays an important role in the dynamical updating of the membership parameters.

5.3.2 Risk functional

The risk functional $R\,[\cdot]$ is defined as the mathematical expectation of the loss function with respect to the probability density function $p(x, y)$ from which the training set was drawn.

$$R\,[f] = E_p\,[c(x, y, f(x))] = \int c(x, y, f(x))dP(x, y) \tag{5.5}$$

where $P(x, y)$ is the cumulative distribution function of $p(x, y)$.

The smaller the risk associated with an estimation of the membership parameters, the better the quality of the estimation. Indeed, a small value of the risk indicates that the estimation's product-distribution is concentrated right from ρ (i.e. in the non-penalized zone). Thus, the membership parameters can be estimated by taking the values that make

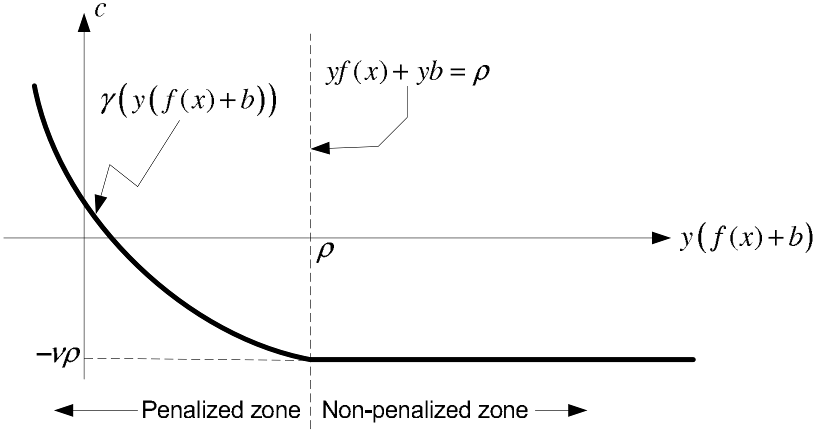

Figure 5.4. The loss function penalizes the values of the product $y\,(f(x)+b)$ that are smaller than ρ. The penalization function $\gamma(\cdot)$ is a polynomial of degree q that ensures the continuity of the loss function at the limit between the penalized and the non-penalized zones. The non-penalized zone corresponds to a constant loss of $-\nu\rho$. The reason for having a non-zero loss in the non-penalized zone is because ν permits to control the fraction of membership errors in the training set (see Section 5.3.5).

the risk functional minimum. However, the probability density function $p(x,y)$ is generally unknown in practical applications. An empirical estimate of $p(x,y)$ can be obtained from training data as follows.

$$p_{\mathrm{emp}}(x,y) = \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \delta_{\mathrm{a}}(x - x_l)\delta_{\mathrm{a}}(y - y_l) \tag{5.6}$$

where $(x_l, y_l) \in \mathcal{S}_{\mathrm{tr}}$ and $\delta_{\mathrm{a}}(\cdot)$ is the analog Dirac's delta function.

By replacing $p_{\mathrm{emp}}(x,y)$ into the definition of the risk functional, we obtain the empirical risk functional:

$$R_{\mathrm{emp}}\,[f] = \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} c(x_l, y_l, f(x_l)) \tag{5.7}$$

Direct minimization of the empirical risk to obtain ρ, b and f is an ill conditioned problem [147, 165], i.e. small changes in the training set may induce large changes in the estimated parameters. Furthermore, the resulting estimation is biased [72] because the risk functional is an ensemble statistic independent from any particular pair (x_l, y_l) whereas the empirical risk depends on the training set only.

Ill posed problems can be effectively solved by adding a regularization term [27, 166] to the original (ill posed) problem. In order to regularize the minimization of the empirical risk we introduce a functional space H to which f belongs. We then obtain the regularized risk functional $R_{\mathrm{reg}}\,[f]$ as follows.

$$R_{\mathrm{reg}}\,[f] = R_{\mathrm{emp}}\,[f] + \frac{\ell}{2}\,\langle f, f\rangle_{\mathrm{H}} \tag{5.8}$$

where $\langle\cdot,\cdot\rangle_{\mathrm{H}}$ is the inner product in H, $\ell \in \mathbb{R}^+$ is the regularization constant, and $\frac{\ell}{2}\,\langle f, f\rangle_{\mathrm{H}}$ is the regularization term.

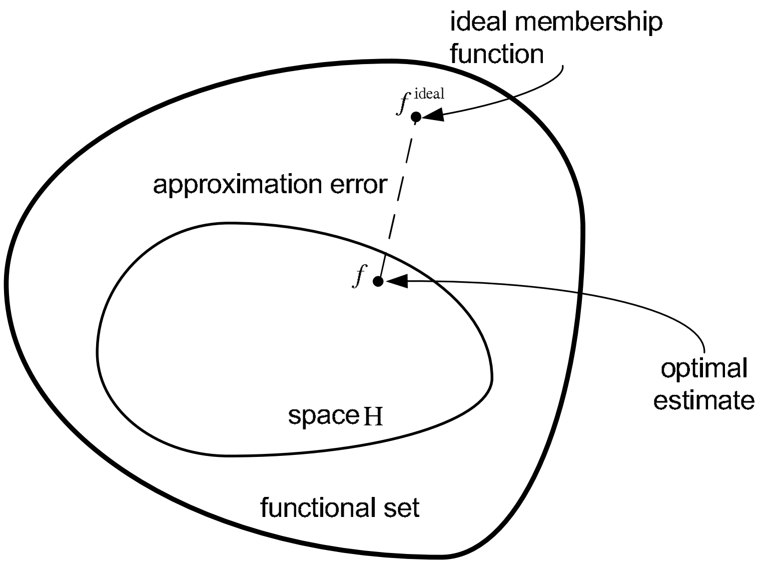

Figure 5.5. The ideal membership function f^{ideal} is in a general functional set which is not necessarily a functional space. The minimizer of the regularized risk (optimal estimate) lies in the functional space H. The (functional) distance between the optimal solution and the true solution is the approximation error which depends on the choice of H.

5.3.3 Nature of the functional space H

In addition to have made the risk minimization better conditioned, the functional space provides f with a structure that facilitates its estimation (see Equation 5.11). Nonetheless, the ideal membership function f^{ideal} does not necessarily belong to H. The element in H which minimizes the regularized risk is called the optimal estimate and its distance from f^{ideal} is the approximation error (Fig. 5.5). Provided that H is a reproducing kernel Hilbert space (RKHS), it is possible to show that as the training set grows, the optimal solution converges to the true one [147].

Recent works on statistical learning theory have proved that when the minimizer of the regularized risk belongs to an RKHS it has good generalization capabilities and is able to capture complex data structures [147, 165] in a theoretically well founded and elegant way. In addition, in [93] and [147] an online framework to estimate the minimizer of R_{reg} is presented. By virtue of these considerations we choose H to be an RKHS. Before defining an RKHS we recall the definition of a Hilbert space.

Definition 5.1. Hilbert space [34]
$\mathcal{H}$ is a Hilbert space if it is complete with respect to its norm $\left(\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}; f \in \mathcal{H}\right)$ Completeness in this context means that any Cauchy sequence of elements of the space converges to an element in the space, in the sense that the norm of differences approaches zero.

All finite dimensional inner product spaces (such as Euclidean space with the ordinary dot product) are Hilbert spaces. However, the infinite dimensional Hilbert spaces such as the spaces $\mathrm{L}^2(\mathbb{R}^n)$ of square-Lebesgue-integrable functions with values in $\mathbb{R}$ or $\mathbb{C}$ are more important in the applications. The inner product allows to perform many "geometrical" constructions familiar from finite dimensions also in the infinite dimensional settings. Of all the infinite dimensional topological vector spaces, the Hilbert spaces properties are the closest to those of the finite dimensional spaces.

Definition 5.2. Reproducing Kernel Hilbert Space (RKHS) [93, 147, 169]
H is a RKHS if it is a Hilbert space and the following properties are satisfied

- *There exists a map K from $\mathcal{X} \times \mathcal{X}$ into $\mathbb{R}$ such that $K(x,\cdot) \in H$ and $\langle h(\cdot), K(x,\cdot)\rangle_H = h(x)$ for all $x \in \mathcal{X}$ and $h \in H$. This is the reproducing property and the mapping $K(\cdot,\cdot)$ is known as the kernel function which generates H. This property implies*

$$\langle K(x_{l1},\cdot), K(x_{l2},\cdot)\rangle_H = K(x_{l1}, x_{l2}) \quad \forall x_{l1}, x_{l2} \in \mathcal{X} \tag{5.9}$$

- *$\forall x_1,\ldots,x_N \in \mathcal{X}$ and $\forall a_1,\ldots,a_N \in \mathbb{R}$, the sum: $\sum\limits_{l,m} a_l a_m K(x_l,x_m)$ is positive or zero.*

  This property implies that any matrix $\mathcal{K}$ with elements $\mathcal{K}_{mn} = K(x_l,x_m)$ is positive semi-definite.

- *$H = span\,\{\, K(x,\cdot)|\, x \in \mathcal{X}\}$*

  According to this property, any $h \in H$ can be written as a linear combination of the functions $K(x,\cdot)$

$$h = \sum_{x\in\mathcal{X}} \alpha_x K(x,\cdot)$$

  The set $\mathcal{X}$ can therefore be considered as an index set for the RKHS. The dimension of H depends on the kernel function $K(\cdot,\cdot)$. In particular, when K is a Gaussian kernel (see Section 5.4.2), H is infinite dimensional.

Using the fact that H is equal to the span of functions $K(x \in \mathcal{X},\cdot)$, the minimizer of the regularized risk (5.8) can be decomposed into a part contained in the span of the elements in the training set and one in the orthogonal complement. This yields:

$$f(\cdot) = \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l,\cdot) + f_\perp(\cdot) \tag{5.10}$$

where (x_l, y_l) belongs to the training set, $\alpha_l \in \mathbb{R}$, and the function $f_\perp \in \mathrm{H}$ is such that: $\langle f_\perp, K(x_l,\cdot)\rangle_{\mathrm{H}} = 0$ for $l = 1,\ldots,\mathrm{L}$.

By replacing (5.10) into the regularized risk (5.8), one gets

$$\begin{aligned} R_{\text{reg}}[f] &= R_{\text{emp}}[f] + \frac{\ell}{2}\left(\left\langle \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l,\cdot), \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l,\cdot)\right\rangle_{\mathrm{H}} + \langle f_\perp, f_\perp\rangle\right) \\ &\geqslant R_{\text{emp}}[f] + \frac{\ell}{2}\left\langle \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l,\cdot), \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l,\cdot)\right\rangle_{\mathrm{H}} \end{aligned}$$

Thus, for any fixed $\alpha_1, \ldots, \alpha_{\mathrm{L}} \in \mathbb{R}$ the regularized risk is minimized for $f_\perp = 0$. Therefore, f is a linear combination of $\{\, K(x_l, \cdot) | \, (x_l, y_l) \in \mathcal{S}_{\mathrm{tr}} \}$.

$$f(\cdot) = \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l, \cdot) \tag{5.11}$$

The coefficient α_l is called the expansion coefficient associated with x_l. Thus, to estimate f amounts to estimate the expansion coefficients of the training vectors. This result is a particular case of the representer theorem of Kimeldorf and Wahba [92].

5.3.4 Geometrical interpretation

We mentioned that Hilbert spaces permit to generalize the geometrical constructions in finite dimensional vector spaces endowed by the classical Euclidean inner product to general spaces (possibly infinite dimensional) such as functional ones.

To obtain a geometrical interpretation of the regularized risk minimization we start by defining a map ϕ from the feature vector space $\mathcal{X}$ into the functional space H as:
$\phi(x) = K(x, \cdot)$ where $x \in \mathcal{X}$.

The reproducing property (5.9) implies that for all $x, x' \in \mathcal{X}$ the following equation holds

$$\left\langle \phi(x), \phi(x') \right\rangle_{\mathrm{H}} = \left\langle K(x, \cdot), K(x', \cdot) \right\rangle_{\mathrm{H}} = K(x, x') \tag{5.12}$$

This relation states that the inner product of two elements in H, namely $\phi(x)$ and $\phi(x')$ can be simply calculated by applying the kernel function on the pre-images of those elements (i.e. x and x'). This constitutes the essence of the well known ”kernel trick” (see Chapter 3, Section 3.4.2 and [1]).

By computing the membership function (5.11) of an $x \in \mathcal{X}$, using the reproducing property (5.9) and the mapping ϕ we get

$$\begin{aligned} f(x) &= \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l, x) && (5.13) \\ &= \sum_{l=1}^{\mathrm{L}} \alpha_l \left\langle K(x_l, \cdot), K(x, \cdot) \right\rangle_{\mathrm{H}} && (5.14) \\ &= \left\langle w, \phi(x) \right\rangle_{\mathrm{H}} && (5.15) \\ \text{where} & \\ w &= \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l, \cdot) = \sum_{l=1}^{\mathrm{L}} \alpha_l \phi(x_l) && (5.16) \end{aligned}$$

Thus, $f(x) + b = 0$ represents a hyperplane in H that is normal to w and has an offset b. Note that the norm $\langle f, f \rangle_{\mathrm{H}}$ in the regularization term is equal to the norm of w.

$$\langle f, f \rangle_{\mathrm{H}} = \langle w, w \rangle_{\mathrm{H}} = \|w\|_{\mathrm{H}}^2 \tag{5.17}$$

Therefore, the membership of $\tilde{x} \in \mathcal{X}$ (decided by f) depends on the position of $\phi(\tilde{x})$ with respect to the hyperplanes $\langle w, \phi(x)\rangle_{\mathrm{H}} + b = \rho$ and $\langle w, \phi(x)\rangle_{\mathrm{H}} + b = -\rho$ (these hyperplanes are called separating margins). As shown in Fig. 5.6 we have:

$$\text{if} \begin{cases} & \langle w, \phi(\tilde{x})\rangle_{\mathrm{H}} + b & \geqslant & \rho & \tilde{x} \text{ is recognized as belonging to } X \\ -\rho & < \langle w, \phi(\tilde{x})\rangle_{\mathrm{H}} + b & < & \rho & \tilde{x} \text{ is a margin error} \\ & \langle w, \phi(\tilde{x})\rangle_{\mathrm{H}} + b & \leqslant & -\rho & \tilde{x} \text{ is recognized as not belonging to } X \end{cases}$$

From the above considerations, it can be said that the map ϕ makes the sets:

$$\begin{aligned} \Phi^{+} &= \{\phi(x_l) |\, (x_l, y_l) \in \mathcal{S}_{tr} \text{ and } y_l = +1\} \\ \Phi^{-} &= \{\phi_k(x_l) |\, (x_l, y_l) \in \mathcal{S}_{tr} \text{ and } y_l = -1\} \end{aligned}$$

linearly separable by f with margin $\mathcal{M}$ (see Fig. 5.6) defined as the distance between the separating margins.

$$2\mathcal{M} = \frac{2\rho}{\|w\|_{\mathrm{H}}} \tag{5.18}$$

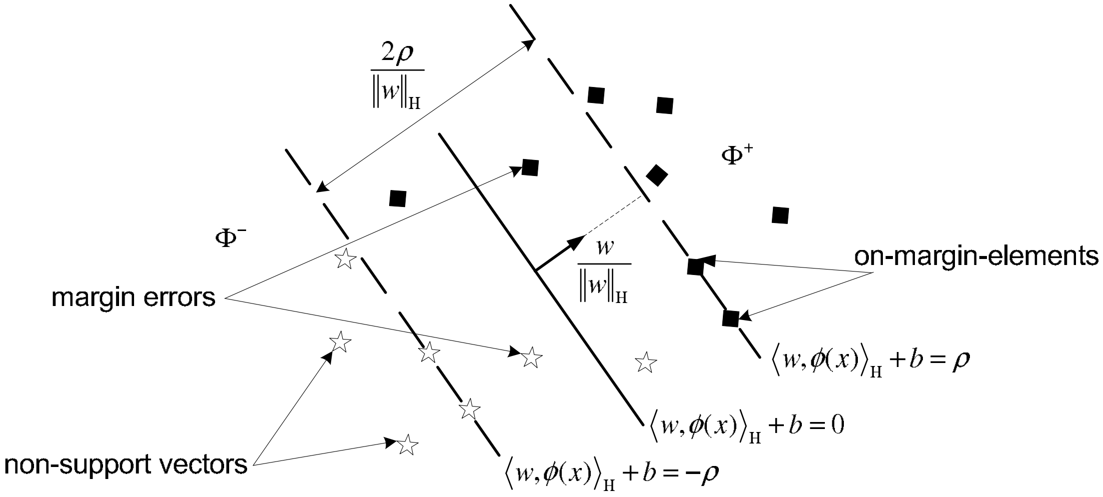
$\frac{2\rho}{\|w\|_{\mathrm{H}}}$
Φ^{+}
Φ^{-}
on-margin-elements
margin errors
$\frac{w}{\|w\|_{\mathrm{H}}}$
$\langle w,\phi(x)\rangle_{\mathrm{H}} + b = \rho$
$\langle w,\phi(x)\rangle_{\mathrm{H}} + b = 0$
non-support vectors
$\langle w,\phi(x)\rangle_{\mathrm{H}} + b = -\rho$

Figure 5.6. Separating margins $\langle w, \phi(x)\rangle_{\mathrm{H}} + b = \pm\rho$ in H. The filled squares represent the $\phi(x_l)$ for which $y_l = +1$ and the stars those for which $y_l = -1$ (the elements (x_l, y_l) belong to the training set). The $\phi(x_l)$ that are located between the separating margins are called margin errors (they have positive expansion coefficients equal to $\frac{1}{\mathrm{L}}$). The on-margin-elements have their positive expansion coefficients in $]0; \frac{1}{\mathrm{L}}[$.
Those $\phi(x_l)$ whose membership is correctly decided by f and are not on the margins are called non-support vectors and their expansion coefficients are equal to zero.

5.3.5 Regularized risk minimization

To determine the membership parameters, the regularized risk (5.8) needs to be minimized. We assume for the moment that the kernel function is given (the kernel choice is described in Section 5.4). By replacing $\langle f,f\rangle_{\mathrm{H}}$ by $\|w\|_{\mathrm{H}}^2$, the regularized risk becomes

$$R_{\mathrm{reg}}\left[f\right] = \mathrm{R}_{\mathrm{emp}}\left[f\right] + \frac{\ell}{2}\left\|w\right\|_{\mathrm{H}}^2 \tag{5.19}$$

The membership parameters are then given by

$$(f,\rho,b) = \underset{f,\rho,b}{\arg\min}\{R_{\mathrm{reg}}\left[f\right]\} \tag{5.20}$$

To solve this optimization problem, we make first the following hypothesis

Hypothesis 5.3. $\forall\,(x_l,y_l)\in\mathcal{S}_{tr}$, $y_l\,(f(x_l)+b)\geqslant\rho$, *i.e. the training data lie in the non-penalized zone of the loss function. In other words the membership of each element in* $\mathcal{S}_{tr}$ *is correctly decided by* f.

Under this hypothesis, the loss function of each element in the training set is equal to: $-\nu\rho$ (see Section 5.3.1). The empirical risk (5.7) being the average of the loss function in the training set, is also equal to $-\nu\rho$. Thus, the regularized risk becomes

$$R_{\mathrm{reg}}\left[f\right] = \frac{\ell}{2}\left\|w\right\|_{\mathrm{H}}^2 - \nu\rho \tag{5.21}$$

Since f and w are equivalent (5.14), we can obtain the first from the latter. Thus, the membership parameters are obtained by solving the optimization problem[1]:

$$(f,\rho,b) \quad = \quad \underset{w,\rho,b}{\arg\min}\left\{\frac{\ell}{2}\left\|w\right\|_{\mathrm{H}}^2 - \nu\rho\right\} \tag{5.22}$$

constrained to

$$y_l\left(\langle w,\phi(x_l)\rangle_{\mathrm{H}}+b\right) \quad \geqslant \quad \rho \text{ for } l=1,\ldots,\mathrm{L} \tag{5.23}$$

$$\rho \quad \geqslant \quad 0 \tag{5.24}$$

Geometrically, to minimize (5.21) amounts to maximize the distance between the separating margins $\mathcal{M}$ (5.18). This concept is central (and usually the starting point) to support vector machines learning algorithms which are considered as large margin classifiers [10].

The minimization of R_{reg} subject to constraints (5.23) and (5.24) is called "hard margin" optimization because no membership errors in the training set are allowed. However, a small number of membership errors in the training set (training error) does not necessarily lead to good predictions of the membership of feature vectors that were not used in the membership

[1] By abuse of notation (since f and w are equivalent (5.14)) we write $(f,\rho,b)=\underset{w,\rho,b}{\arg\min}\{\ldots\}$ for $(w,\rho,b)=\underset{w,\rho,b}{\arg\min}\{\ldots\}$.

parameters estimation (unseen feature vectors). The error incurred by f in predicting the membership of unseen feature vectors is called the generalization error. As it is pointed out in Section A.1 in the appendix, a too small training error leads to over-fitting, i.e. small training error and large generalization error.

To control the over-fitting, we modify our initial hypothesis (Hyp. 5.3) by relaxing the constraints (5.23). Such relaxation is carried out by introducing positive slack variables $\xi_1, \dots, \xi_{\mathrm{L}}$ so that the new constraints become: $y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) \geqslant \rho - \xi_l$. The slack variables have the effect of bringing the training data into the penalized zone of the loss function (see Section 5.3.1), i.e. $c(x_l, y_l, f(x_l)) = -\nu\rho + \xi_l^q$ for $l = 1, \dots, \mathrm{L}$. The empirical risk (5.7) associated with the training data under the relaxed constraints is:

$$R_{\mathrm{emp}} = -\nu\rho + \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \xi_l^q \tag{5.25}$$

By replacing (5.25) into (5.19) we obtain the relaxed risk

$$R_{\xi}\left[f\right] = \frac{\ell}{2} \left\| w \right\|_{\mathrm{H}}^2 - \nu\rho + \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \xi_l^q \tag{5.26}$$

Therefore, the membership parameters are estimated by minimizing the relaxed risk under the relaxed constants, as follows:

$$(f, \rho, b) = \underset{w, \rho, b, \xi_1, \dots, \xi_{\mathrm{L}}}{\arg\min} \left(\frac{\ell}{2} \left\| w \right\|_{\mathrm{H}}^2 - \nu\rho + \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \xi_l^q \right) \tag{5.27}$$

constrained to

$$\begin{aligned} y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) &\geqslant \rho - \xi_l & (5.28) \\ \xi_l &\geqslant 0 & (5.29) \\ \rho &\geqslant 0 & (5.30) \end{aligned}$$

for $l = 1, \dots, L$.

It is worth noting that while the constraints are relaxed by the slack variables, the sum $\sum_{l=1}^{L} \xi_l^q$ prevents too many ξ_l becoming larger than zero. In this way, the slack variables determine the tradeoff between over-fitting and training error.

The relaxed optimization is handled by introducing positive Lagrange multipliers $\tilde{\alpha}_1, \dots, \tilde{\alpha}_L, \beta_1, \dots, \beta_L, \delta \geqslant 0$ and a Lagrangian Λ_P.

$$\begin{aligned} \Lambda_P = {} & \frac{\ell}{2} \| w \|_{\mathrm{H}}^2 - \nu\rho + \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \xi_l^q \\ & - \sum_{l=1}^{L} \left(\tilde{\alpha}_l \left(y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) - \rho + \xi_l \right) + \beta_l \xi_l \right) - \delta\rho \end{aligned} \tag{5.31}$$

The Lagrangian Λ_P has to be minimized with respect to the primal variables $w, \rho, b, \xi_1, \ldots, \xi_\mathrm{L}$ and maximized with respect to the dual variables $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\mathrm{L}, \beta_1, \ldots, \beta_\mathrm{L}, \delta$, i.e. a saddle point has to be found. Taking partial derivatives of Λ_P with respect to the primal variables and setting them to zero leads to the following results.

$$\partial_w \Lambda_P = 0 \quad \Rightarrow \quad w = \frac{1}{\ell} \sum_{l=1}^{\mathrm{L}} \tilde{\alpha}_l y_l \phi(x_l) \tag{5.32}$$

$$\partial_\rho \Lambda_P = 0 \quad \Rightarrow \quad \sum_{l=1}^{\mathrm{L}} \tilde{\alpha}_l - \delta = \nu \tag{5.33}$$

$$\partial_b \Lambda_P = 0 \quad \Rightarrow \quad \sum_{l=1}^{\mathrm{L}} y_l \tilde{\alpha}_l = 0 \tag{5.34}$$

$$\partial_{\xi_l} \Lambda_P = 0 \quad \Rightarrow \quad \tilde{\alpha}_l + \beta_l = \frac{q}{\mathrm{L}} \xi_l^{q-1} \tag{5.35}$$

By replacing (5.32) to (5.35) in the Lagrangian Λ_P and using:

$$\begin{aligned} \|w\|_\mathrm{H}^2 &= \frac{1}{\ell^2} \sum_{l,m=1}^{\mathrm{L}} \tilde{\alpha}_l \tilde{\alpha}_m y_l y_m K(x_l, x_m) \\ \sum_{l=1}^{\mathrm{L}} \tilde{\alpha}_l y_l \langle w, \phi(x_l) \rangle_\mathrm{H} &= \frac{1}{\ell} \sum_{l,m=1}^{\mathrm{L}} \tilde{\alpha}_l \tilde{\alpha}_m y_l y_m K(x_l, x_m) \end{aligned}$$

we get the dual Lagrangian:

$$\Lambda_D = -\frac{1}{2\ell} \sum_{l,m=1}^{\mathrm{L}} \tilde{\alpha}_l \tilde{\alpha}_m y_l y_m K(x_l, x_m) - \frac{q-1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \xi_l^q \tag{5.36}$$

which has to be maximized with respect to $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\mathrm{L}, \xi_1, \ldots, \xi_\mathrm{L}$.

Two parameters remain to be determined, namely the penalty degree q and the regularization constant ℓ (the kernel function is assumed fixed).

Let q be assumed as fixed. Then, the regularization constant determines the tradeoff between the margin (defined as the distance between the separating margins see Section 5.3.4) maximization and training error. Indeed, $\frac{1}{\ell}$ penalizes the first term on the right in (5.36) which is proportional to the inverse of the margin (5.18), i.e. the smaller ℓ the larger the margin and the larger the training error since more ξ_l's are allowed to become strictly positive. The optimal value of ℓ is data dependent and can be determined using cross-validation.

A similar argument can be used for q. Indeed, the larger q the smaller the training error and the smaller the margin. Again, its optimal value could be determined using cross-validation. However, the value of q is related to the uniqueness of the maximizing arguments of Λ_D [24] and (as described in Section 5.6) to the dynamic updating of the membership parameters.

In fact, one can show [24] that the maximizing arguments of Λ_D, i.e. $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\mathrm{L}, \xi_1, \ldots, \xi_\mathrm{L}$, are unique if $q > 1$ (because the sum of convex functions is convex). When $q = 1$ the maximizing arguments are non-unique only if the separating margins (see Fig. 5.6) can be rigidly translated without changing the optimum value of Λ_D [24]. This situation is highly exceptional in practice. Furthermore, in Section 5.6 we show that in order to dynamically update the membership parameters q must be set to one.

By setting $q = 1$, the second term on the right in (5.36) vanishes making the regularization constant ℓ no longer relevant. For convenience we set ℓ to one. Thus, the $\tilde{\alpha}_l$ are found by minimizing: $-\Lambda_D|_{\ell=1,q=1}$.

$$(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{\mathrm{L}}) = \underset{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_L}{\arg\min} \left(\frac{1}{2} \sum_{l,m=1}^{\mathrm{L}} \tilde{\alpha}_l \tilde{\alpha}_m y_l y_m K(x_l, x_m) \right) \tag{5.37}$$

constrained to:

$$0 \leqslant \tilde{\alpha}_l \leqslant \frac{1}{\mathrm{L}} \text{ (results from 5.35, } q = 1\text{, and } \beta_l \geqslant 0 \text{)} \tag{5.38}$$

$$\sum_{l=1}^{L} \tilde{\alpha}_l \geqslant \nu \text{ (results from (5.33) and } \delta \geqslant 0) \tag{5.39}$$

$$\sum_{l=1}^{L} y_l \tilde{\alpha}_l = 0 \text{ (results from (5.34))} \tag{5.40}$$

Standard quadratic programming techniques [168] can be used to solve the above optimization problem and find the optimum values of $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{\mathrm{L}}$. These coefficients completely determine the membership parameters as we explain later in the text. For ease of explanation we analyze the solution in function of the $\tilde{\alpha}_l$'s.

At the optimum, the Karush-Kuhn-Tucker (KKT) conditions [95] imply that the following relations hold.

$$\tilde{\alpha}_l \left(y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) - \rho + \xi_l \right) = 0 \tag{5.41}$$

$$\beta_l \xi_l = 0 \tag{5.42}$$

$$\delta \rho = 0 \tag{5.43}$$

The position of $\phi(x_l)$ with respect to the separating margins: $\langle w, \phi(x) \rangle_{\mathrm{H}} + b = \pm\rho$, depends on the $\tilde{\alpha}_l$. According to (5.38) three possibilities exist:

- If $\tilde{\alpha}_l = 0$ then, $y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) \geqslant \rho$. Therefore, the membership of x_l with respect to X is correctly determined (see Fig. 5.7a).

- If $0 < \tilde{\alpha}_l < \frac{1}{\mathrm{L}}$, then $y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) = \rho$. Again the membership of x_l with respect to X is correctly determined (Fig. 5.7b). In this case, $\phi(x_l)$ is an on-margin-element, i.e. it lies on the separating margin: $\langle w, \phi(x) \rangle_{\mathrm{H}} + b = y_l \rho$ (see Fig. 5.6).

- If $\tilde{\alpha}_l = \frac{1}{\mathrm{L}}$, depending on whether $\xi_l = 0$ or $\xi_l > 0$, $y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) = \rho$ or $y_l \left(\langle w, \phi(x_l) \rangle_{\mathrm{H}} + b \right) < \rho$ respectively. In the first case $\phi(x_l)$ is on the separating margin and in the latter case the membership of x_l is wrongly determined (see Fig. 5.7c).

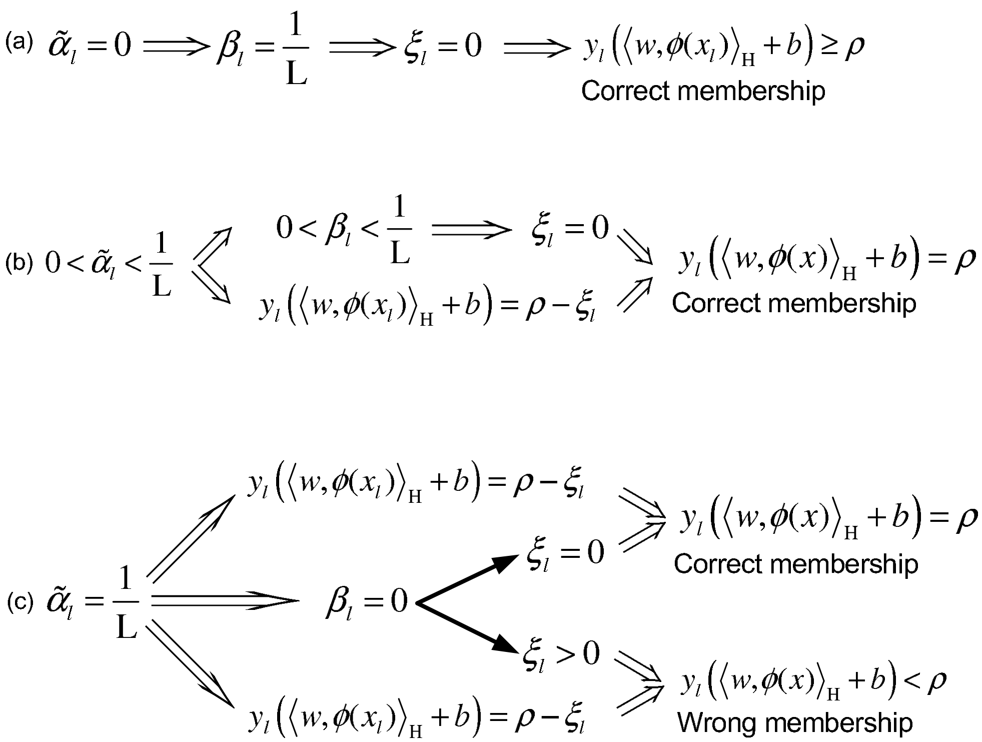
(a) $\tilde{\alpha}_l = 0 \Longrightarrow \beta_l = \frac{1}{\mathrm{L}} \Longrightarrow \xi_l = 0 \Longrightarrow y_l\left(\langle w, \phi(x_l)\rangle_{\mathrm{H}} + b\right) \geq \rho$
Correct membership
(b) $0 < \tilde{\alpha}_l < \frac{1}{\mathrm{L}}$
$0 < \beta_l < \frac{1}{\mathrm{L}} \Longrightarrow \xi_l = 0$
$y_l\left(\langle w, \phi(x_l)\rangle_{\mathrm{H}} + b\right) = \rho - \xi_l$
$y_l\left(\langle w, \phi(x)\rangle_{\mathrm{H}} + b\right) = \rho$
Correct membership
(c) $\tilde{\alpha}_l = \frac{1}{\mathrm{L}}$
$y_l\left(\langle w, \phi(x_l)\rangle_{\mathrm{H}} + b\right) = \rho - \xi_l$
$\beta_l = 0$
$\xi_l = 0$
$y_l\left(\langle w, \phi(x)\rangle_{\mathrm{H}} + b\right) = \rho$
Correct membership
$\xi_l > 0$
$y_l\left(\langle w, \phi(x)\rangle_{\mathrm{H}} + b\right) < \rho$
Wrong membership
$y_l\left(\langle w, \phi(x_l)\rangle_{\mathrm{H}} + b\right) = \rho - \xi_l$

Figure 5.7. Membership of x_l depending on the value of $\tilde{\alpha}_l$. *Top*: when $\tilde{\alpha}_l = 0$, the membership of x_l is correctly determined. *Middle*: when $0 < \tilde{\alpha}_l < \frac{1}{L}$, $\phi(x_l)$ is on the separating margin: $\langle w, \phi(x)\rangle_{\mathrm{H}} + b = y_l\rho$. The membership of x_l is correctly decided. *Bottom*: when $\tilde{\alpha}_l = \frac{1}{L}$, the membership is correctly determined if $\xi_l = 0$ and wrongly determined if $\xi_l > 0$.

For properly collected training data and an adequate choice of the kernel function (Sect. 5.4) one expects that most of the training elements satisfy the condition in Fig. 5.7a, i.e. the solution is expected to be sparse in the $\tilde{\alpha}_l$'s. As a matter of fact, the number of $\tilde{\alpha}_l$'s different from zero constitutes an indication on the expected generalization error [147, 165].

The role of ν

In Section 5.3.1 we assigned a constant loss of $-\nu\rho$ to the non-penalized zone and mentioned that the parameter ν permits to control the training error. To illustrate this, we consider the KKT condition on ρ (5.43). If ρ is strictly positive then, $\delta = 0$ which according to (5.33) imply: $\sum\limits_{l=1}^{\mathrm{L}} \tilde{\alpha}_l = \nu$. In particular, the sum of the $\tilde{\alpha}_l$'s for which the membership of their respective x_l is wrongly determined (i.e. $\xi_l > 0$) should satisfy:

$$\sum_{l|\tilde{\alpha}_l = \frac{1}{\mathrm{L}}\ ;\ \xi_l > 0} \tilde{\alpha}_l \leqslant \nu \tag{5.44}$$

From Fig. 5.7c it comes out:

$$\frac{1}{\mathrm{L}}\left|\{\,l|\,y_l(\langle w,\phi(x_l)\rangle_{\mathrm{H}}+b)<\rho\}\right| \quad \leqslant \quad \nu \tag{5.45}$$

where $|\{\,l|\,y_l(\langle w,\phi(x_l)\rangle_{\mathrm{H}}+b)<\rho\}|$ is the number of membership errors in the training set. This above inequality states that the fraction of training errors (FTE) is upper bounded by ν [147].

In addition to bounding the FTE, ν is a lower bound on the fraction of $\tilde{\alpha}_l$'s different from zero. Indeed, using $\tilde{\alpha}_l \leqslant \frac{1}{\mathrm{L}}$ and $\sum\limits_{l=1}^{\mathrm{L}} \tilde{\alpha}_l = \nu$ one can easily obtain:

$$\nu = \sum_{l|\tilde{\alpha}_l>0} \tilde{\alpha}_l \leqslant \frac{1}{\mathrm{L}}\left|\{\,l|\,\tilde{\alpha}_l>0\}\right| \tag{5.46}$$

The $\phi(x_l)$ (and by extension the respective x_l) for which $\alpha_l > 0$ are called support vectors because they completely determine the membership of any $x \in X$. Thus, from (5.46), ν lower bounds the fraction of support vectors (FSV). Combining (5.46) and (5.45) we obtain the ν inequality:

$$\mathrm{FTE} \leqslant \nu \leqslant \mathrm{FSV} \tag{5.47}$$

The smaller ν the smaller the FTE. However, a too small FTE will not generally lead to a small generalization error because of the over-fitting. Thus, ν needs to be adjusted in order to reach a compromise between the good generalization (i.e. small expected generalization error) and the FTE. As we show in the next section, the generalization error and the FTE depend also on the kernel function. An interdependent choice of ν and the kernel function is presented in Section 5.5.

So far, we have discussed the solution of the relaxed optimization in terms of the $\tilde{\alpha}_l$'s. We now turn to determining the membership parameters f, ρ, b from the $\tilde{\alpha}_l$'s.

The membership function f is completely determined by the expansion coefficients α_l which in turn, according to (5.16) and (5.32) satisfy

$$\alpha_l = y_l\tilde{\alpha}_l \quad \text{for } l = 1,\ldots,\mathrm{L} \tag{5.48}$$

The membership function of an $x \in \mathcal{X}$ is then given by:

$$f(x) = \sum_{l=1}^{\mathrm{L}} y_l\tilde{\alpha}_l K(x_l, x) \tag{5.49}$$

The decision threshold ρ and decision offset b are obtained by taking two elements x_{l1} and x_{l2} such that:

$$\begin{aligned} l1 &= \operatorname*{arg\,min}_{l1|y_{l1}=+1}\left[\mathrm{abs}\left(\frac{1}{2\mathrm{L}}-\tilde{\alpha}_{l1}\right)\right] \\ l2 &= \operatorname*{arg\,min}_{l2|y_{l2}=-1}\left[\mathrm{abs}\left(\frac{1}{2\mathrm{L}}-\tilde{\alpha}_{l2}\right)\right] \end{aligned}$$

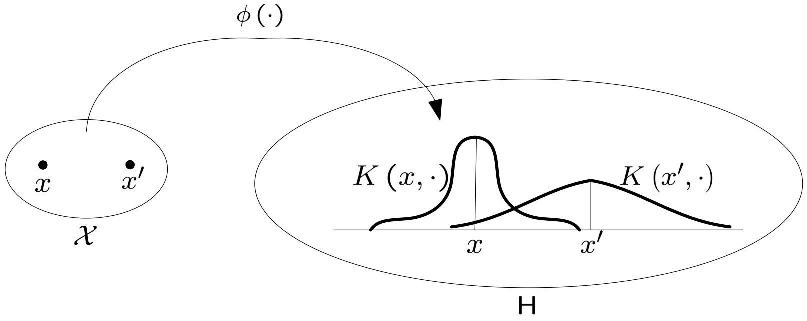

Figure 5.8. The functional map ϕ is such that it transforms an element $x \in \mathcal{X}$ into a function $\phi(x) = K(x, \cdot)$ which represents a measure of the similarity between x and all the other elements in $\mathcal{X}$. Thus, we expect $K(x, \cdot)$ to be centered around x and take its maximum value at this same point.

From Fig. 5.7(b) we know that $\phi(x_{l1})$ and $\phi(x_{l2})$ are on the margins, therefore:

$$\rho = \frac{1}{2}\left(\langle w, \phi(x_{l1})\rangle_{\mathrm{H}} - \langle w, \phi(x_{l2})\rangle_{\mathrm{H}}\right) \tag{5.50}$$

$$b = -\frac{1}{2}\left(\langle w, \phi(x_{l1})\rangle_{\mathrm{H}} + \langle w, \phi(x_{l2})\rangle_{\mathrm{H}}\right) \tag{5.51}$$

Theoretically any x_{l1}, x_{l2} whose corresponding $\tilde{\alpha}_{l1}, \tilde{\alpha}_{l2}$ are in $\left]0; \frac{1}{\mathrm{L}}\right[$ can be selected to estimate ρ and b. But, for numerical precision reasons, the choice of two elements such that their positive expansion coefficients are close to $\frac{\mathrm{L}}{2}$ allow us to improve the reliability of the result.

5.4 Kernel function

In the previous section we have shown how the map ϕ transforms the set $\mathcal{X}$ into a functional space H where the membership of the training data is determined by their position with respect to the separating hyperplanes. In other words, the training data become linearly classifiable, into classes defined by their membership, in H. The properties of H, namely its dimensionality and its ability to achieve the linear separation with a reasonably small FTE and good generalization depends on the kernel.

For classification purposes, the kernel function $K(x, x')$, where $x, x' \in \mathcal{X}$, is considered as a measure of the similarity between x and x'. Thus, $\phi(x)$ maps x into the function $K(x, \cdot)$ which represents a measure of the similarity between x and all other elements in $\mathcal{X}$ (see Fig. 5.8). In the classification framework the polynomial and Gaussian kernels are commonly used. In the next sections we describe these kernels and discuss their properties.

5.4.1 Polynomial kernel

The polynomial kernel similarity measure is the cosine of the angle between its arguments. For $x_1, x_2 \in \mathcal{X}$ and $d \in \mathbb{N}^*$ (the polynomial kernel grade) the polynomial kernel is defined as:

$$K_d(x_1, x_2) = (\langle x_1, x_2 \rangle_{\mathcal{X}})^d \tag{5.52}$$

If d is set to 1 we obtain the linear kernel which represents the inner product in $\mathcal{X}$. In this case the space H is equivalent to $\mathcal{X}$.

$$K_{d=1}(x_1, x_2) = \langle x_1, x_2 \rangle_{\mathcal{X}} \tag{5.53}$$

One can prove [139] that the larger the polynomial kernel grade the smaller the FTE. Yet, large generalization errors are associated with large values of d. A compromise can be reached through cross-validation. In a more elaborated approach [26], d depends on the minimization of a theoretical bound on the generalization error.

The disadvantage of the polynomial kernel resides in its sensitivity with respect to scaling factors [157]. Indeed, if the x's are not centered around the origin (usually they are far from the origin, especially when band power values are used to determine the feature vector space (see Chapter 4, Section 4.3), see Chapter 4) their norms are large and the angle between them is small. In this case the polynomial kernel becomes:

$$K_d(x_1, x_2) = \|x_1\|_{\mathcal{X}}^d \, \|x_2\|_{\mathcal{X}}^d \cos^d \measuredangle\,(x_1, x_2) \approx \|x_1\|_{\mathcal{X}}^d \, \|x_2\|_{\mathcal{X}}^d$$

Thus, K_d is determined only by the norm of its arguments, without taking into account the angle (i.e. the genuine dissimilarity factor). A possible way to handle it, would consist in normalizing the training data before estimating the membership parameters. The normalization process consists in making each component of the training vectors zero average and unit standard deviation. The normalization parameters are then stored so as to apply them on new data. Another, more principled way to prevent scaling problems consists in whitening the training data by diagonalizing their covariance matrix [20].

5.4.2 Gaussian kernel

The similarity measure of the Gaussian kernel is the Euclidean distance of its arguments (this makes the Gaussian kernel less sensitive to scaling factors). For $x_1, x_2 \in \mathcal{X}$ and (Gaussian kernel parameter) $\sigma \in \mathbb{R}^+$, the Gaussian kernel is defined as [148, 165]:

$$K_\sigma(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|_{\mathcal{X}}^2}{\sigma^2}\right) \tag{5.54}$$

The dimension of the functional space generated by K_σ is infinite [147, 157]. This can be intuitively understood because it is possible to find an infinite number of pairs $x_1, x_2 \in \mathcal{X}$ such that: $\langle \phi(x_1), \phi(x_2) \rangle_{\mathrm{H}} = K_\sigma(x_1, x_2) \approx 0$.

The Gaussian kernel constitutes an attractive choice for our application because of its outstanding classification performance in the EEG framework [59, 61, 62, 63], its relative insensibility to scaling factors and capacity to accurately approximate any classification surface [148].

Generally, the smaller σ the smaller the FTE (see Section A.1 in the appendix). However, a too small σ leads to over-fitting. The following proposition summarizes the influence of σ on the FTE and FSV.

Proposition 5.4. *If the membership parameters are estimated in a functional space H, generated by a Gaussian kernel with parameter σ and the membership threshold is such that $\rho > 0$ then, the fraction of training errors $FTE[K_\sigma, \nu]$ and the fraction of support vectors $FSV[K_\sigma, \nu]$ associated with a given choice of σ and ν satisfy:*

i) $FTE[K_\sigma, \nu] \xrightarrow[\sigma \to 0]{} 0$ *and* $FSV[K_\sigma, \nu] \xrightarrow[\sigma \to 0]{} 1$.
Too small values of σ over-fit the training data.

ii) $FTE[K_{\sigma \gg 1}, \nu] = FTE[K_{d=1}, \nu]$ *where $K_{d=1}$ is the linear kernel defined in* (5.53).
Too large values of σ under-fit the training data.

Proof

i) From the definition of Gaussian kernel (5.54) it is clear that

$$\lim_{\sigma \to 0} K_\sigma(x_l, x_m) = \begin{cases} 0 & \text{if } x_l \neq x_m \\ 1 & \text{otherwise} \end{cases} \tag{5.55}$$

By replacing (5.55) into (5.37), the coefficients $\tilde{\alpha}_l$ are then found by solving:

$$(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_L) = \underset{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_L}{\arg\min} \left(\frac{1}{2} \sum_{l=1}^{L} \tilde{\alpha}_l^2 \right) \tag{5.56}$$

Since $\rho > 0$, (5.39) implies $\sum\limits_{l=1}^{L} \tilde{\alpha}_l = \nu$. Then the solution of (5.56) is:

$$\tilde{\alpha}_1 = \ldots = \tilde{\alpha}_L = \frac{\nu}{\mathrm{L}} < \frac{1}{L}$$

Thus, $\phi(x_1), \ldots, \phi(x_L)$ are all on the separating margin (see Fig. 5.7). Consequently, the membership of all the training data is correctly determined, i.e. FTE $[K_{\sigma \to 0}, \nu] = 0$. Also, since all the $\tilde{\alpha}_l$'s are strictly positive FSV $[K_{\sigma \to 0}, \nu] = 1$.

ii) If $\sigma \gg 1$, the following approximation holds

$$K_{\sigma \gg 1}(x_l, x_m) \approx 1 - \frac{\|x_l - x_m\|_{\mathcal{X}}^2}{\sigma^2} = 1 - \frac{\|x_l\|_{\mathcal{X}}^2 + \|x_m\|_{\mathcal{X}}^2}{\sigma^2} + \frac{2}{\sigma^2} \langle x_l, x_m \rangle_{\mathcal{X}} \tag{5.57}$$

Using this approximation, the term to minimize in (5.37) becomes

$$\begin{aligned} \frac{1}{2} \sum_{l,m=1}^{L} \tilde{\alpha}_l \tilde{\alpha}_m y_l y_m K_{\sigma \gg 1}(x_l, x_m) \approx \frac{1}{2} \sum_{l,m=1}^{L} y_l y_m \tilde{\alpha}_l \tilde{\alpha}_m - \frac{1}{\sigma^2} \sum_{l=1}^{L} y_l \tilde{\alpha}_l \sum_{m=1}^{L} y_m \tilde{\alpha}_m \|x_m\|_{\mathcal{X}}^2 \\ + \frac{1}{\sigma^2} \sum_{l,m=1}^{L} y_l y_m \tilde{\alpha}_l \tilde{\alpha}_m \langle x_l, x_m \rangle_{\mathcal{X}} \end{aligned} \tag{5.58}$$

Since, according to (5.34), $\sum_{l=1}^{L} y_l \tilde{\alpha}_l = 0$ the first two terms on the right in (5.58) vanish. Using the definition of linear Kernel (5.53), the coefficients $\tilde{\alpha}_l$ are found by solving:

$$(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_L) = \underset{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_L}{\arg\min} \left(\frac{1}{\sigma^2} \sum_{l,m} y_l y_m \tilde{\alpha}_l \tilde{\alpha}_m K_{d=1}(x_l, x_m) \right) \tag{5.59}$$

The optimum $\tilde{\alpha}_l$ being identical for $K_{\sigma \gg 1}$ and $K_{d=1}$, $\text{FTE}\,[K_{\sigma \gg 1}] = \text{FTE}\,[K_{d=1}]$.

5.5 Choice of the parameters ν and σ

In Section 5.3.5 we discussed the role of ν as an upper bound on the FTE and as a lower bound on the FSV. We pointed out that a too small value of ν leads to over-fitting. On the other hand, according to Proposition 5.4 the FTE and the FSV depend on the kernel parameter σ as well.

Both σ and ν are data dependent, their optimal values need to be jointly selected so as to ensure a small generalization error. In this section we present two approaches to achieve this. The first approach is based on the well known cross-validation method and the second one on the minimization of a theoretical bound on the generalization error proposed in [165].

5.5.1 Cross-validation approach

In this approach, the generalization error associated with a given choice of ν and σ is estimated as follows. The training data is split into $\mathcal{P}$ parts of (approximately) equal size. Then, $\mathcal{P}$ estimations of the membership parameters are performed. Each estimation leaves out one of the parts to compute the fraction of membership errors on it (validation errors fraction), and estimates the membership parameters on the $(\mathcal{P} - 1)$ remaining parts. The average of the $\mathcal{P}$ validation errors fractions constitutes the generalization error estimate (GE). At the limit where each part contains a single training element one has the leave-one-out estimate of the generalization error which is almost unbiased [37, 53, 147]. Similarly, the average of the $\mathcal{P}$ fractions of training errors correspond to the FTE associated with the chosen ν and σ.

The parameters ν and σ are determined in two steps. First, an initial ν is used to obtain a relatively wide interval (we call it loose interval) in which the value of σ that makes the GE minimum lies. Second, ν is set equal to the maximum training error in the loose interval and finally the optimum σ in the loose interval is determined. In the following, we detail this procedure.

The initial choice of ν (denoted as ν_0) is based on the fact that this parameter constitutes an upper bound on the FTE for any kernel function (Gaussian or not, see Section 5.3.5). Then, a possible initial choice for ν would be the random classification threshold (i.e. 0.5). However, a better choice is given by $\text{FTE}\,[K_{d=1}, \nu = 0.5]$ which according to Prop. 5.4 and the fact that the fraction of training errors generally increase with σ (see [147] and

Section A.1 in the appendix) constitutes the maximum FTE for Gaussian kernels. In practice, FTE $[K_{d=1}, \nu = 0.5]$ can be larger than 0.5; the initial ν is therefore given by:

$$\nu_0 = \min\left(\frac{1}{2}, \text{FTE}\left[K_{d=1}, \nu = 0.5\right]\right) \tag{5.60}$$

To determine the loose interval for σ we need to sample the function GE $[K_\sigma, \nu_0]$, i.e. the generalization error estimate in function of the Gaussian kernel parameter. Between the small and large values of σ which respectively, over-fit and under-fit the training data (Proposition 5.4) a small value for GE $[K_\sigma, \nu_0]$ has to be found.

Since the Gaussian kernel dissimilarity measure is the Euclidean distance of its arguments, it makes sense to take, as extreme values for σ, the minimum Δ_{min} and the maximum Δ_{max} of the Euclidean distance in the training set.

$$\Delta^2_{\text{min}} = \min_{x_{l1}, x_{l2} \in \mathcal{S}_{\text{tr}}} \|x_{l1} - x_{l2}\|^2_{\mathcal{X}} \tag{5.61}$$

$$\Delta^2_{\text{max}} = \max_{x_{l1}, x_{l2} \in \mathcal{S}_{\text{tr}}} \|x_{l1} - x_{l2}\|^2_{\mathcal{X}} \tag{5.62}$$

The evolution of GE $[K_\sigma, \nu_0]$ for σ in $[\Delta_{\text{min}}, \Delta_{\text{max}}]$ can be efficiently covered (i.e. with relatively few values) by geometrically sampling in $[\Delta_{\text{min}}, \Delta_{\text{max}}]$. Thus, the set of σ values at which GE $[K_\sigma, \nu_0]$ is evaluated is:

$$\text{V}_\sigma = \left\{ \sigma_v = (\Delta_{\text{min}})^{\frac{N_\sigma - v}{N_\sigma - 1}} (\Delta_{\text{max}})^{\frac{v-1}{N_\sigma - 1}} \,\middle|\, v = 1, \ldots, N_\sigma \right\}$$

where N_σ is the number of samples. The sampling ratio is: $\left(\frac{\Delta_{\text{max}}}{\Delta_{\text{min}}}\right)^{\frac{1}{N_\sigma - 1}}$

Let σ_{v*} be the value for which the generalization error estimate in V_σ is minimum, namely, $v* = \underset{v=1,\ldots,N_\sigma}{\arg\min} \left(\text{GE}\left[K_{\sigma_v}, \nu_0\right]\right)$. Then, the loose interval for σ is:

$$\mathcal{I}_\sigma = [\sigma_{v*-1}; \sigma_{v*+1}]$$

We mentioned earlier that the FTE is a growing function of σ. This means that ν can be set to: FTE $\left[K_{\sigma_{v*+1}}, \nu_0\right]$, i.e. the training error associated with σ_{v*+1} and ν_0.

By linearly sampling in $\mathcal{I}_\sigma$, one can readily find an approximation of: $\underset{\sigma \in \mathcal{I}_\sigma}{\arg\min} \left(\text{GE}\left[K_\sigma, \nu\right]\right)$. The approximation accuracy depends on the sampling resolution.

The cross-validation approach constitutes a practical way to select σ and ν that more often than not leads to very good results [62, 147]. However, as explained in [147] this approach amounts to optimize the parameters on the same set as the one used for training, which can potentially lead to over-fitting.

In Fig. 5.9 we report a case (that appears often in practice) in which the location of the σ that makes GE $[\sigma, \nu]$ minimum is rather fuzzy. Indeed, there is almost no difference in choosing either of the σ values indicated by the vertical dashed lines. We can empirically remove the fuzziness by also considering the FSV which, according to [147] constitutes an upper bound on the generalization error and is also linked with the complexity of the decision surface (see Section A.1 in the appendix). Thus, the selection of σ can be modified

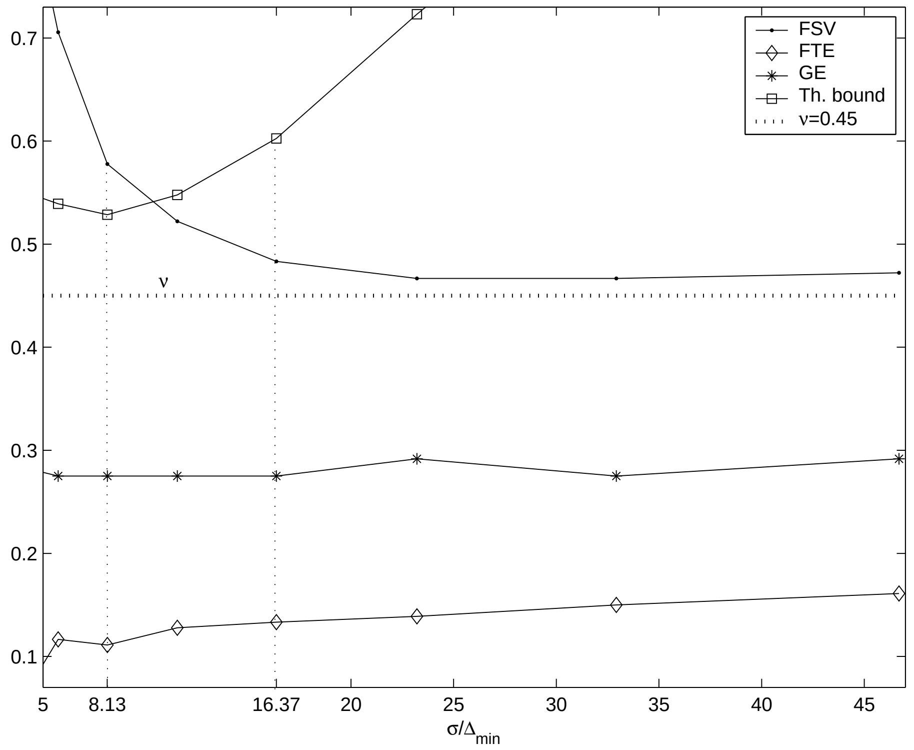

Figure 5.9. Evolution of the FSV, FTE, GE and the theoretical bound $\mathcal{B}$ in function of σ normalized to Δ_{min}. The dotted line marks the value of ν which upper bounds the FTE and lower bounds the FSV. In this example, the optimum value of σ is rather fuzzy. Indeed, if the GE only is considered there is almost no difference in choosing either of the values indicated by the vertical dashed lines. On the opposite, the theoretical bound $\mathcal{B}$ clearly indicates the most adequate choice.
The theoretical bound curve was conveniently scaled for visualization purposes.

by taking the minimum of an empirical aggregate criterion that takes into account the GE and a strictly growing function of the FSV. In fact, such criterion can be obtained using the minimum description length (MDL) general framework (see [140] and Chapter 4 ,Section 4.5). The MDL based choice for σ is:

$$\sigma = \arg\min_{\sigma} \left(\log\left(\mathrm{GE}\left[\sigma,\nu\right]\right) + \mathrm{FSV}\left[\sigma,\nu\right] \frac{\log \mathrm{L}}{\mathrm{L}} \right) \tag{5.63}$$

where ν is determined as explained above.

With the above example, we illustrated the fact that additional elements like the FSV are sometimes necessary to refine the choice of σ. However, the combination of the GE and the FSV is rather empirical. This lead us to consider another more theoretical approach that considers a bound on the generalization error.

5.5.2 Theoretical bound based approach

In [166] V. Vapnik proposes an upper bound, denoted as $\mathcal{B}$, on the GE associated with a separating hyperplane that passes through the origin. Although we consider separating hyperplanes that do not necessarily satisfy this condition (i.e. the membership offset, b can be non-zero) $\mathcal{B}$ can still be used to approximate [26, 61] the evolution of the GE in function of ν and σ.

The bound, $\mathcal{B}$ depends on the the radius R (see Section A.2 in the appendix, for the detailed computation of R) of the smallest sphere containing the set: $\{\,\phi(x_l)|\,(x_l, y_l) \in \mathcal{S}_{tr}\}$ and the distance between the separating margins $\mathcal{M}$ (5.18).

$$\begin{aligned} \mathcal{B} &= \frac{1}{\mathrm{L}}\frac{R^2}{\mathcal{M}^2} && (5.64)\\ (5.18) \Rightarrow \mathcal{B} &= \frac{1}{\mathrm{L}}\frac{R^2\,\|w\|_{\mathrm{H}}^2}{(\rho)^2} && (5.65)\end{aligned}$$

In Fig. 5.9 we report the evolution of $\mathcal{B}$ (conveniently scaled for visualization) in function of σ normalized to $\Delta_{\min}$. The position of the σ that makes $\mathcal{B}$ minimum is now far more clear than in the cross-validation approach. It is worth noting that while $\mathcal{B}$ appears to be a rough upper-bound on the GE, its minimum permits to select the optimal Gaussian kernel parameter. Moreover, the smooth evolution of $\mathcal{B}$ makes possible the use of gradient descent techniques to find its minimum.

The optimum kernel parameter, σ^* is the solution of:

$$\sigma^* \;=\; \arg\min_{\sigma}\left(\frac{1}{L}\frac{R^2\,\|w\|_{\mathrm{H}}^2}{(\rho)^2}\right) \qquad (5.66)$$

since $\mathcal{B}$ is differentiable, σ^* can be iteratively estimated using gradient descent as follows.

$$\sigma^* \leftarrow \sigma^* - \eta_\sigma \left.\frac{\partial \mathcal{B}}{\partial \sigma}\right|_{\sigma^*} \qquad (5.67)$$

where η_σ is the learning rate which has to be set so that it avoids oscillations around the optimum value while granting a reasonable speed of convergence. Generally, the value of η_σ is adjusted in the curse of the algorithm.

If the value of σ^* is sought by applying gradient descent (especially if a small η_σ is used) in the whole σ-range, i.e. between $\Delta_{\min}$ and $\Delta_{\max}$, this process can take a considerable amount of time. To overcome this problem we use the loose interval determined in the cross-validation approach (Section 5.5.1), i.e. $\mathcal{I}_\sigma = [\sigma_{v*-1}; \sigma_{v*+1}]$. From $\mathcal{I}_\sigma$, a narrow interval I_σ (in which gradient descent is applied) can be determined by considering the value of the derivative $\frac{\partial \mathcal{B}}{\partial \sigma}$ at σ_{v*}:

$$\text{if}\begin{cases} \frac{\partial \mathcal{B}}{\partial \sigma}\Big|_{\sigma_{v*}} < 0 & I_\sigma = [\sigma_{v*}; \sigma_{v*+1}] \\ \frac{\partial \mathcal{B}}{\partial \sigma}\Big|_{\sigma_{v*}} = 0 & \sigma^* = \sigma_{v*} \\ \frac{\partial \mathcal{B}}{\partial \sigma}\Big|_{\sigma_{v*}} > 0 & I_\sigma = [\sigma_{v*-1}; \sigma_{v*}] \end{cases} \qquad (5.68)$$

The computation of $\frac{\partial\mathcal{B}}{\partial\sigma}$ is detailed in Section A.3 in the appendix. In the following we describe the algorithm through which the optimum values for ν and σ (denoted as ν^* and σ^* respectively) are found.

Algorithm to determine ν^* and σ^*

```
Determine the loose interval $\mathcal{I}_\sigma = [\sigma_{v*-1}; \sigma_{v*+1}]$ using cross-validation.
Compute the value of $\frac{\partial\mathcal{B}}{\partial\sigma}\Big|_{\sigma_{v*}}$
**If** $\frac{\partial\mathcal{B}}{\partial\sigma}\Big|_{\sigma_{v*}} = 0$ **then** stop. The $\mathcal{B}$ minimum has been found. Thus, $\sigma^*$ is set to $\sigma_{v*}$ and
    $\nu^*$ is set to the initial value of the cross-validation approach $\nu_0$ (see Eq. 5.60).
**else** determine the narrow interval $I_\sigma = [\sigma_{\min}; \sigma_{\max}]$ using (5.68)
Initialization
    $\sigma^{*(0)} = \sigma_{\max}$
    $\nu^{*(0)} = \min\left(\frac{1}{2}, \mathrm{FTE}\left[K_{d=1}, \nu = 0.5\right]\right)$
    $\mathcal{B}^{(0)} = \mathcal{B}^{(0)}\left(\sigma^{*(0)}, \nu^{*(0)}\right)$
    $\mathrm{FTE}^{(0)} = \mathrm{FTE}\left[K_{\sigma^{*(0)}}, \nu^{*(0)}\right]$
    $n = 1$
**repeat**
    $\sigma^{*(n)} = \sigma^{*(n-1)} - \eta_\sigma\, \partial_\sigma \mathcal{B}^{(n-1)}\Big|_{\sigma=\sigma^{*(n-1)}}$
    $\nu^{*(n)} = \mathrm{FTE}^{(n-1)}$
    $\mathcal{B}^{(n)} = \mathcal{B}^{(n)}\left(\sigma^{*(n)}, \nu^{*(n)}\right)$
    $\mathrm{FTE}^{(n)} = \mathrm{FTE}\left[K_{\sigma^{*(n)}}, \nu^{*(n)}\right]$
    $n = n + 1$
**until** $\mathcal{B}^{(n-1)} > \mathcal{B}^{(n-2)}$ (i.e. $\mathcal{B}$ starts to increase). Thus, $\sigma^*$ and $\nu^*$ are set to $\sigma^{*(n-1)}$
    and $\nu^{*(n-1)}$ respectively.
```

The notation $\mathcal{B}^{(n)} = \mathcal{B}^{(n)}\left(\sigma^{*(n)}, \nu^{(n)}\right)$ means that $\mathcal{B}^{(n)}$ is computed using the values of the radius and the margin corresponding to $\sigma^{*(n)}$ and $\nu^{*(n)}$. Note that since ν upper bounds the FTE, its value at the n-th step is set to the previous FTE, namely $\mathrm{FTE}^{(n-1)}$.

5.6 Dynamic updating of the membership parameters

So far, we have discussed the estimation of the membership parameters, f, ρ, b by minimizing the regularized risk in a training set:

$$\mathcal{S}_{\mathrm{tr}} = \{\,(x_l, y_l)|\, x_l \in \mathcal{X},\, y_l \in \{-1, +1\}, \text{ and } l = 1, 2, \ldots, \mathrm{L}\}$$

In particular, we have shown that the membership function f can be written as: $f(\cdot) = \sum_{l=1}^{\mathrm{L}} \alpha_l K(x_l, \cdot)$ and is sparse in the the expansion coefficients α_l. The non-zero expansion coefficients determine which training elements need to be stored in order to compute the membership of new EEG vectors.

If we denote as $\mathcal{D}$ the set composed of the estimated membership parameters, we have

$$\mathcal{D} = \arg\min\left(R_{\text{reg}}\left[\mathcal{S}_{tr}\right]\right) \tag{5.69}$$

where $R_{\text{reg}}\left[\mathcal{S}_{tr}\right]$ is the regularized risk (5.8) associated with the training set.

When new feature vectors whose membership is known become available, the membership parameters need to be estimated again so as to adapt them to possible changes. Let $\{(x_{\text{L}+1}, y_{\text{L}+1}), \ldots, (x_{\text{L}+m}, y_{\text{L}+m})\}$ be the set of new training data, the new set of membership parameters $\mathcal{D}^{\text{L}+m}$ is therefore given by:

$$\mathcal{D}^{(\text{L}+m)} = \arg\min\left(R_{\text{reg}}\left[\mathcal{S}_{\text{tr}} \cup \{(x_{\text{L}+1}, y_{\text{L}+1}), \ldots, (x_{\text{L}+m}, y_{\text{L}+m})\}\right]\right) \tag{5.70}$$

Direct use of (5.70) to estimate $\mathcal{D}^{(\text{L}+m)}$ results in a non-scalable problem of growing complexity. Instead, we seek to determine an updating relation $\mathcal{R}$ such that:

$$\begin{aligned} \mathcal{D}^{(\text{L}+m)} &= \mathcal{R}\left(\mathcal{D}^{(\text{L}+m-1)}, (x_{\text{L}+m}, y_{\text{L}+m})\right) && (5.71)\\ \mathcal{D}^{(\text{L})} &= \mathcal{D} && (5.72) \end{aligned}$$

To determine $\mathcal{R}$ we apply the method presented in [93] according to which, the regularized risk (5.8) with $\ell = 1$, is locally approximated at $(x_{\text{L}+m}, y_{\text{L}+m})$ by the stochastic risk

$$\begin{aligned} R_{\text{stoch}}\left[f^{(\text{L}+m-1)}, \text{L}+m\right] = c&\left(x_{\text{L}+m}, y_{\text{L}+m}, f^{(\text{L}+m-1)}\left(x_{\text{L}+m}\right)\right) \\ &+ \frac{1}{2}\left\langle f^{(\text{L}+m-1)}, f^{(\text{L}+m-1)}\right\rangle_{\text{H}} \end{aligned} \tag{5.73}$$

where $f^{(\text{L}+m-1)}$ is given by

$$f^{(\text{L}+m-1)}\left(\cdot\right) = \sum_{l=1}^{\text{L}+m-1} \alpha_l^{(\text{L}+m-1)} K(x_l, \cdot) \tag{5.74}$$

and $c\left(x_{\text{L}+m}, y_{\text{L}+m}, f^{(\text{L}+m-1)}\left(x_{\text{L}+m}\right)\right)$ is the loss function function corresponding to the new training pair $(x_{\text{L}+m}, y_{\text{L}+m})$ and the previous membership function $f^{(\text{L}+m-1)}$

If we denote as θ any element in $\mathcal{D}$, its updating relation is:

$$\theta^{(\text{L}+m)} = \theta^{(\text{L}+m-1)} - \eta_m \frac{\partial R_{\text{stoch}}\left[f^{(\text{L}+m-1)}, \text{L}+m\right]}{\partial \theta^{(\text{L}+m-1)}} \tag{5.75}$$

where $\eta_m \in \mathbb{R}^+$ is the updating coefficient when the $(\text{L}+m)$-th training element becomes available. The membership parameters of index (L) are those estimated using $\mathcal{S}_{\text{tr}}$. This means: $\alpha_l^{(\text{L})} = \alpha_l$, $\rho^{(\text{L})} = \rho$ and $b^{(\text{L})} = b$.

5.6.1 Dynamic updating of the membership function

The membership function updating equation is obtained by replacing θ by f in (5.75). This yields:

$$f^{(\mathrm{L}+m)} = f^{(\mathrm{L}+m-1)} - \eta_m \frac{\partial R_{\mathrm{stoch}}\left[f^{(\mathrm{L}+m-1)}, \mathrm{L}+m\right]}{\partial f^{(\mathrm{L}+m-1)}} \tag{5.76}$$

Using the stochastic risk definition (5.73) and applying the chain rule to obtain $\frac{\partial c(\dots)}{\partial f}$ we get:

$$\begin{aligned} f^{(\mathrm{L}+m)} &= f^{(\mathrm{L}+m-1)} \\ &- \eta_m \left(\frac{\partial c\left(x_{\mathrm{L}+m}, y_{\mathrm{L}+m}, f^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right)\right)}{\partial f^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right)}\right)\left(\frac{\partial f^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right)}{\partial f^{(\mathrm{L}+m-1)}}\right) \\ &- \frac{\eta_m}{2}\left(\frac{\partial\left\langle f^{(\mathrm{L}+m-1)}, f^{\mathrm{L}+m-1}\right\rangle_{\mathrm{H}}}{\partial f^{(\mathrm{L}+m-1)}}\right) \end{aligned} \tag{5.77}$$

The functional derivatives: $\frac{\partial f^{(\mathrm{L}+m-1)}(x_{\mathrm{L}+m})}{\partial f^{(\mathrm{L}+m-1)}}$ and $\frac{\partial\left\langle f^{(\mathrm{L}+m-1)}, f^{\mathrm{L}+m-1}\right\rangle_{\mathrm{H}}}{\partial f^{(\mathrm{L}+m-1)}}$ are computed using the following definition.

Definition 5.5. Functional derivative
The derivative of the functional $\mathcal{F}[f]$ with respect to f in a functional space $\mathcal{H}$ is defined [34] by:

$$\frac{\partial \mathcal{F}[f]}{\partial f} = \lim_{\epsilon \to 0} \frac{\mathcal{F}\left[f(u) + \epsilon\delta(u-v)\right] - \mathcal{F}\left[f(u)\right]}{\epsilon}$$

In particular, for $g, f \in \mathcal{H}$, we have:

$$\frac{\partial \langle g, f\rangle_{\mathcal{H}}}{\partial f} = g \tag{5.78}$$

$$\frac{\partial \langle f, f\rangle_{\mathcal{H}}}{\partial f} = 2f \tag{5.79}$$

Using (5.78), (5.79), and the reproducing property (Def. 5.2); the functional derivatives in (5.76) are given by:

$$\begin{aligned} \frac{\partial f^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right)}{\partial f^{(\mathrm{L}+m-1)}} &= \frac{\partial\left\langle K_\sigma\left(x_{\mathrm{L}+m}, \cdot\right), f^{(\mathrm{L}+m-1)}\right\rangle_{\mathrm{H}}}{\partial f^{(\mathrm{L}+m-1)}} \\ &= K_\sigma\left(x_{\mathrm{L}+m}, \cdot\right) \end{aligned} \tag{5.80}$$

$$\frac{\partial\left\langle f^{(\mathrm{L}+m-1)}, f^{\mathrm{L}+m-1}\right\rangle_{\mathrm{H}}}{\partial f^{(\mathrm{L}+m-1)}} = 2 f^{(\mathrm{L}+m-1)} \tag{5.81}$$

Replacing the loss function definition (see Section 5.3.1), (5.80), and (5.81) in the membership function updating equation (5.77) yields:

$$\begin{aligned} f^{(\mathrm{L}+m)} &= \left(1-\eta_m\right) f^{(\mathrm{L}+m-1)} \\ &+ \eta_m y_{\mathrm{L}+m} q \left(g^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right)\right)^{q-1} \Theta\left[g^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right)\right] K_\sigma\left(x_{\mathrm{L}+m}, \cdot\right) \end{aligned} \tag{5.82}$$

where $g^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right) = -\rho^{(\mathrm{L}+m-1)} + y_{\mathrm{L}+m}\left(f^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right) + b^{(\mathrm{L}+m-1)}\right)$ and $\Theta\left[u\right]$ is such that

$$\Theta\left[u\right] = \begin{cases} 1 & u < 0 \\ 0 & u \geqslant 0 \end{cases}$$

It should be noted that if $g^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right)$ is positive or zero the membership of $x_{\mathrm{L}+m}$ is correctly determined by $f^{(\mathrm{L}+m-1)}$. In this case, (5.82) reduces to: $f^{(\mathrm{L}+m)} = \left(1-\eta_m\right) f^{(\mathrm{L}+m-1)}$

By replacing the membership functions by their respective linear expansions in terms of the kernel functions (5.74) we get the updating equations for the expansion coefficients α_l.

If the penalty degree q is larger than 1, the second term on the right side of (5.82) has multiplicative terms of the form:

$$\alpha_{l_1}^{(\mathrm{L}+m-1)} \cdots \alpha_{l_{q-1}}^{(\mathrm{L}+m-1)} K_\sigma\left(x_{\mathrm{L}+m},\cdot\right) K_\sigma\left(x_{l_1},\cdot\right) \cdots K_\sigma\left(x_{l_{q-1}},\cdot\right)$$

Since such multiplicative terms do not exist on the left side of (5.82) they have to be null, i.e. the penalty degree q should be set to 1. Thus, the update equations for the expansion coefficients are:

$$\alpha_l^{(\mathrm{L}+m)} = \left(1-\eta_m\right)\alpha_l^{(\mathrm{L}+m-1)} \text{ for } l = 1,\ldots,\mathrm{L}+m-1 \tag{5.83}$$

$$\alpha_{\mathrm{L}+m}^{(\mathrm{L}+m)} = \begin{cases} 0 & \text{if } g^{(\mathrm{L}+m-1)}\left(x_{\mathrm{L}+m}\right) \geqslant 0 \\ \eta_m y_{\mathrm{L}+m} & \text{otherwise} \end{cases} \tag{5.84}$$

If the membership of the most recent training element, namely $x_{\mathrm{L}+m}$ is correctly determined by $f^{(\mathrm{L}+m-1)}$, its corresponding expansion coefficient $\alpha_{\mathrm{L}+m}^{(\mathrm{L}+m)}$ is set to zero. Given the updating equation (5.83) it is clear that $\alpha_{\mathrm{L}+m}$ will remain equal to zero. Thus, $x_{\mathrm{L}+m}$ does not contribute to the decision function f and can be safely "forgotten".

On the other hand, if the membership of $x_{\mathrm{L}+m}$ is wrongly determined, its corresponding expansion coefficient is set to $\eta_m y_{\mathrm{L}+m}$, i.e. $x_{\mathrm{L}+m}$ becomes a support vector. In Section 5.3.5 we have seen that the expansion coefficient associated with a support vector whose membership is wrongly decided is equal to $\frac{1}{\mathrm{L}+m} y_{\mathrm{L}+m}$. The latter suggest that η_m should be set equal to $\frac{1}{\mathrm{L}+m}$. However, for m large enough η_m becomes closer to zero which makes the contribution of $x_{\mathrm{L}+m}$ insignificant. This is certainly not suitable as f would be determined by the first training elements only. To deal with this problem we consider the fact that only the support vectors determine the membership function. The effective number of training elements when $x_{\mathrm{L}+m}$ becomes available is then equal to the number of support vectors at the time just after the $(m-1)$-th updating is completed (this number is denoted as $\mathrm{NSV}^{(m-1)}$). Therefore, the coefficient η at the m-th updating is given by:

$$\eta_m = \frac{1}{\mathrm{NSV}^{(m-1)} + 1} \tag{5.85}$$

We now turn to assessing the evolution of a expansion coefficient which appeared at the m-th updating, after $\hat{m}$ steps. Using the updating equations (5.83) and (5.84) we have:

$$\begin{aligned} \alpha_{\mathrm{L}+m}^{(\mathrm{L}+m+\hat{m})} &= \eta_m y_{\mathrm{L}+m} \prod_{l=1}^{\hat{m}} (1-\eta_{m+l}) \\ &= \eta_m y_m \prod_{l=1}^{\hat{m}} \frac{\mathrm{NSV}^{(m+l-1)}}{\mathrm{NSV}^{(m+l-1)}+1} \end{aligned} \tag{5.86}$$

We consider two extreme cases:

- The memberships of $x_{\mathrm{L}+m+1}, \ldots, x_{\mathrm{L}+m+\hat{m}}$ are wrongly determined, i.e. the number of support vectors increases at each updating. Then, after the $\hat{m}$-th updating, $\alpha_{\mathrm{L}+m}$ becomes:

$$\begin{aligned} \alpha_{\mathrm{L}+m}^{(\mathrm{L}+m+\hat{m})} &= \eta_m y_{\mathrm{L}+m} \prod_{l=1}^{\hat{m}} \frac{l-1+\mathrm{NSV}^{(m)}}{l+\mathrm{NSV}^{(m)}} \\ &= \eta_m y_{\mathrm{L}+m} \frac{\mathrm{NSV}^{(m)}}{\hat{m}+\mathrm{NSV}^{(m)}} \end{aligned} \tag{5.87}$$

- The memberships of $x_{\mathrm{L}+m+1}, \ldots, x_{\mathrm{L}+m+\hat{m}}$ are correctly determined, i.e. the number of support vectors remained constant at each updating. Then, after the $\hat{m}$-th updating, $\alpha_{\mathrm{L}+m}$ becomes:

$$\begin{aligned} \alpha_{\mathrm{L}+m}^{(\mathrm{L}+m+\hat{m})} &= \eta_m y_{\mathrm{L}+m} \prod_{l=1}^{\hat{m}} \frac{\mathrm{NSV}^{(m)}}{1+\mathrm{NSV}^{(m)}} \\ &= \eta_m y_{\mathrm{L}+m} \left(\frac{\mathrm{NSV}^{(m)}}{1+\mathrm{NSV}^{(m)}}\right)^{\hat{m}} \end{aligned} \tag{5.88}$$

In both cases, $\alpha_{\mathrm{L}+m}^{(\mathrm{L}+m+\hat{m})}$ tends to zero as $\hat{m}$ grows to infinity. However, (5.88) converges exponentially to zero while (5.87) converges linearly. Thus, we can state that the forgetting speed of an expansion coefficient is approximately determined by the number of correct membership predictions in which it participated. Using this fact, we can approximate the number of correct decisions $\hat{M}$ that an expansion coefficient, associated with the training element $x_{\mathrm{L}+m}$, "survives to". Let ε be the machine precision, then (5.88) yields:

$$\varepsilon > \left(\frac{\mathrm{NSV}^{(m)}}{1+\mathrm{NSV}^{(m)}}\right)^{\hat{M}} \tag{5.89}$$

$$\hat{M} = \left\lceil \frac{\log \varepsilon}{\log\left(\frac{\mathrm{NSV}^{(m)}}{1+\mathrm{NSV}^{(m)}}\right)} \right\rceil \tag{5.90}$$

where $\lceil \cdot \rceil$ is the ceiling function, i.e. this function gives the smallest integer that is larger than its argument.

5.6.2 Dynamic updating of the membership threshold and offset

The dynamic updating relations for the membership threshold and offset are obtained by replacing θ in (5.75) by ρ and b respectively.

$$\rho^{(\mathrm{L}+m)} = \begin{cases} \rho^{(\mathrm{L}+m-1)} + \eta_m \nu & \text{if } g^{(\mathrm{L}+m-1)}(x_{\mathrm{L}+m}) \geqslant 0 \\ \rho^{(\mathrm{L}+m-1)} - \eta_m \left(1-\nu\right) & \text{otherwise} \end{cases} \tag{5.91}$$

$$b^{(\mathrm{L}+m)} = \begin{cases} b^{(\mathrm{L}+m-1)} & \text{if } g^{(\mathrm{L}+m-1)}(x_{\mathrm{L}+m}) \geqslant 0 \\ b^{(\mathrm{L}+m-1)} + \eta_m y_{\mathrm{L}+m} & \text{otherwise} \end{cases} \tag{5.92}$$

If the membership of $x_{\mathrm{L}+m}$ is correctly determined by $f^{(\mathrm{L}+m-1)}$ then, the membership threshold is increased by $\eta_m \nu$ and the membership offset remains constant. On the other hand, if the membership of $x_{\mathrm{L}+m}$ is wrongly determined, the decision membership is decreased by $\eta_m \left(1-\nu\right)$ and the decision offset increases by $\eta_m y_{\mathrm{L}+m}$.

5.7 Summary

The recognition of the mental activities that are used to control the BCI is carried out in feature vector spaces that are subject and mental activity dependent. Thus, each mental activity has an associated feature vector space in which we define a target set, composed of the feature vectors produced during the performance of the targeted mental activity.

The recognition goal is to determine the membership of a feature vector with respect to the target set. This is done by means of the membership parameters, namely the membership function, threshold and offset. These parameters are estimated in a supervised way, i.e. using a set of (training) feature vectors whose membership is known.

Since the shape of the target set can change because of different environmental and user related conditions including the adaptation of the subject to the BCI, the membership parameters need to be updated as new training data become available while forgetting the contribution of old training feature vectors.

In this chapter we have presented an efficient method to estimate the membership parameters and dynamically update them. The method is based on the minimization of a regularized version of the risk functional which is a measure of the inadequacy of a given estimation.

The regularization of the risk is made possible by the introduction of a RKHS to which the membership function belongs. Such an RKHS has the property of having a kernel function that generates it. By means of the kernel function each feature vector can be transformed in a function in RKHS.

In addition to make the regularization possible, the RKHS provides the membership function with a particularly flexible structure as a linear combination of the kernel functions associated with the training elements. Thanks to this structure a geometrical interpretation for the membership parameters in terms of a separating hyperplane can be derived and their updating equations are particularly straightforward.

The RKHS properties and ability to classify the feature vectors into classes defined by their membership depend on the choice of the Kernel function. A suitable kernel function is the Gaussian kernel which permits to easily control the over-fitting and generalization through its parameter σ. Such parameter is selected using a theoretical bound on the generalization error.

6 Protocols and evaluation

"It is a capital mistake to theorize before one has data" *Sir Arthur Conan Doyle*

6.1 Introduction

In previous chapters we presented the process through which, a vector of memberships is obtained from an EEG-trial free of artifacts. This vector is sent to the action generation module (see Fig. 6.1) which, in accordance with a set of rules (action rules), produces commands that act on a computer-rendered environment (CRE). These rules are set experimentally and depend on the MAs used to operate the BCI and the subject performance. Henceforth, unless otherwise specified, the term MA refers to a mental activity used to operate the BCI.

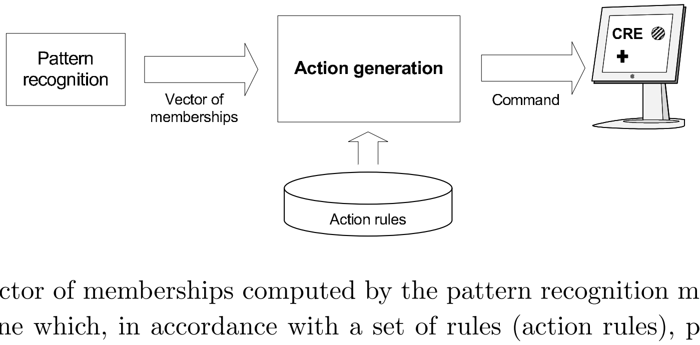

Figure 6.1. The vector of memberships computed by the pattern recognition module is sent to the action generation one which, in accordance with a set of rules (action rules), produces commands that act on a computer rendered environment. The action rules are set experimentally depending on the MAs and on the subject control skills.

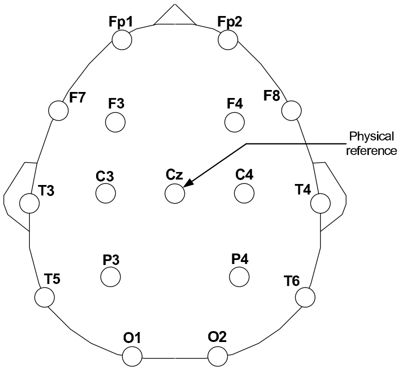

Figure 6.2. Electrodes of the ten-twenty international system at which EEG was measured. Electrode Cz was taken as physical reference. As mentioned in (Chapter 3, Section 3.2.3), the signals were re-referenced with respect to their average.

In this chapter we apply the preprocessing, feature extraction, and recognition algorithms presented in previous chapters, for the training of six subjects who participated in nine training sessions, in the framework of an asynchronous 2D object positioning application.

The first three sessions served to set up initial recognition models for each MA, determine the optimal feature extraction methods or mappings (see Chapter 4), and establish the action rules.

In the next training sessions, feedback was provided, indicating the subjects how well the BCI recognized the MAs they were asked to perform. At the end of each session the recognition models were updated. The controlling skills acquired by the subjects in these sessions were assessed through positioning tests in which the subjects had to move an object on the screen (by performing the trained MAs) to reach a goal. The training schedule was adjusted in function of the recognition error associated with each MA. Thus, those MAs associated with high recognition errors were trained more often.

6.2 Experiment description

Six male subjects (denoted as S1 to S6 respectively) aged 22, 24, 25, 25, 27, and 32 years respectively participated in nine training sessions distributed over two months. These sessions were carried out in a quiet and slightly illuminated room in which the subjects were comfortably sitting in an armchair and placed in front of an LCD computer monitor that was placed at a distance of about 1.5 meters from them.

Signals at electrodes Fp1, Fp2, F7, F3, F4, F8, T3, C3, C4, T4, T5, P3, P4, P4, T6, O1, and O2 were recorded. The electrode Cz was taken as physical reference (see Fig. 6.2). As mentioned in (Chapter 3, Section 3.2.3) the signals were re-referenced with respect to their average. The MAs used to operate the BCI are described in Table 6.1. In addition, we denote as MA0 any mental activity but the controlling ones.

Mental activity	**Description**
MA1: Left index finger imagination	Left index finger movement imagination
MA2: Right index finger imagination	Right index finger movement imagination
MA3: Geometric rotation	Imagined rotation of an object on the screen. Actually, subjects were asked to imagine the spaceship, shown in Fig. 6.3 rotating with respect to the x-axis.
MA4: Visual reverse counting	From a three-digit starting number, subjects imagined the numbers, obtained by successively subtracting 3, being displayed on the screen. The starting number was not provided in the positioning tests.

Table 6.1. Mental activities used to operate the BCI

6.2.1 Object positioning application

We considered an asynchronous 2D object positioning application with four possible movements, namely up, down, left, and right. In Fig. 6.3 we depict the CRE in which the application took place and the associations between the possible movements of the spaceship and the MAs. The controlling skills of a subject were tested in terms of his ability to reach the target (the sun) with the spaceship (Section 6.5).

The CRE was a one-hundred step square where a step corresponded to the smallest movement that the spaceship could execute. The number of steps that the spaceship moved during a single action depended on the action rules (Section 6.4).

6.2.2 Training program

Subjects were trained to gain control of the spaceship according to the following program. In the first three sessions (training-without-feedback sessions), the subjects were asked to perform the MAs according to a defined protocol (see Section 6.3). On completion of these sessions the optimal feature extraction mappings were chosen, the recognition models were built, and the action rules were set.

The next six sessions (training-with-feedback sessions), were divided into two parts. In the first part, the subjects were asked to perform the MAs and a feedback was provided telling them how well the BCI recognized the MA they were asked to perform. In the second part, we carried out two positioning tests in which we asked the subjects to reach the target with the spaceship. In each run, the positions of the spaceship and the target were randomized so that it took fifty steps to reach the target. In the following we detail the two types of training sessions and report the corresponding results.

Three training-without-feedback sessions were carried out before the training-with-feedback sessions, in order to collect enough data to estimate the recognition models.

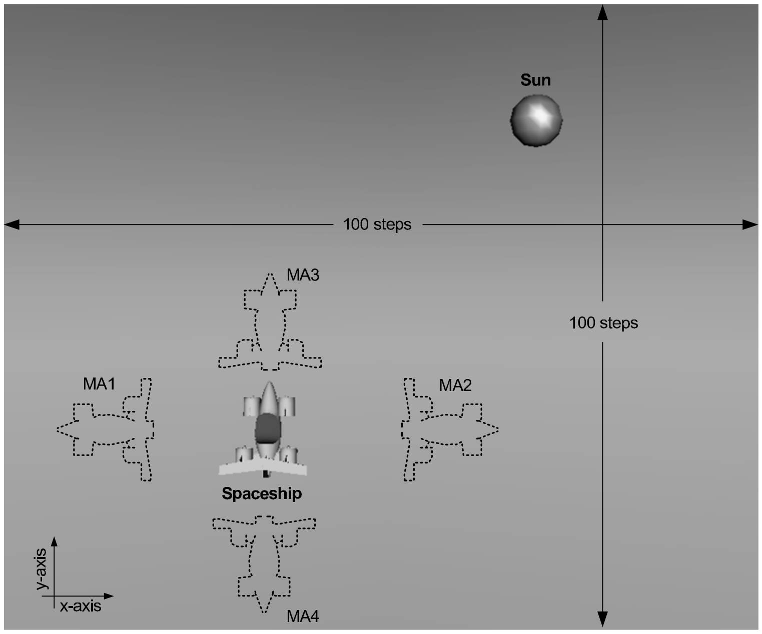

Figure 6.3. CRE in which the object positioning application takes place. The spaceship moves to the left, right, up, and down when the BCI recognizes MA1, MA2, MA3, and MA4 respectively. The CRE is a one-hundred step square where a step corresponds to the smallest movement that the spaceship can execute.

6.3 Training without feedback

Training-without-feedback sessions were structured according to the following protocol. Once the setup (i.e. the placement of the EEG cap, filling of the cap electrodes with the electro-gel, etc.) was completed, the first five to ten minutes were spent at calibrating the noise removal, and artifact detection algorithms according to the procedure explained in Chapter 3, Section 3.4.3. The rest of the sessions were divided into three five-minute slices with subject defined breaks between them (see Fig. 6.4). Each five-minute slice consisted of thirty six second long active-times (i.e. the time segments in which a MA was performed) followed by break-times whose duration was randomized between three to four seconds. During break-times, the subjects were asked to perform any mental activity but the controlling ones. In consequence, we consider the EEG-trials recorded during break-times as generated during the performance of MA0.

The outset of an active-time was signaled by a visual cue which also indicated the MA that had to be performed. Subjects were asked to perform the corresponding MA during the entire active-time. In Fig. 6.5 we portray the visual cues associated with each controlling

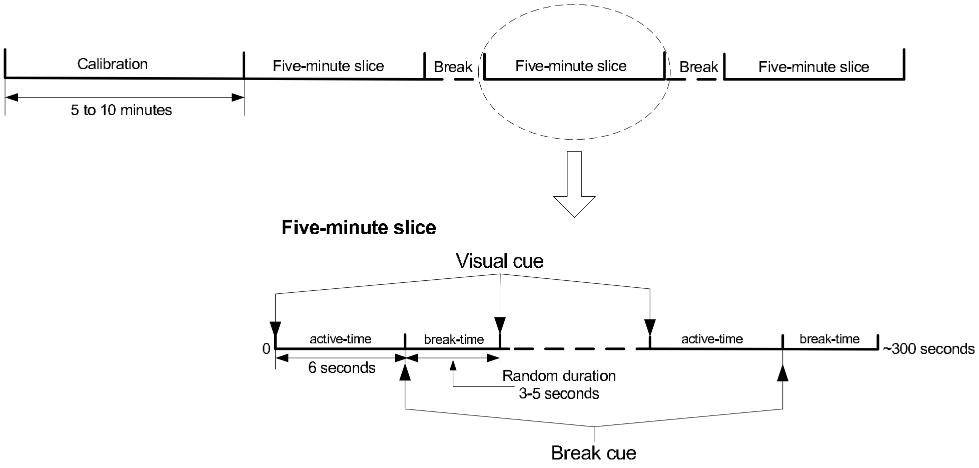

Figure 6.4. Protocol of a training-without-feedback session.

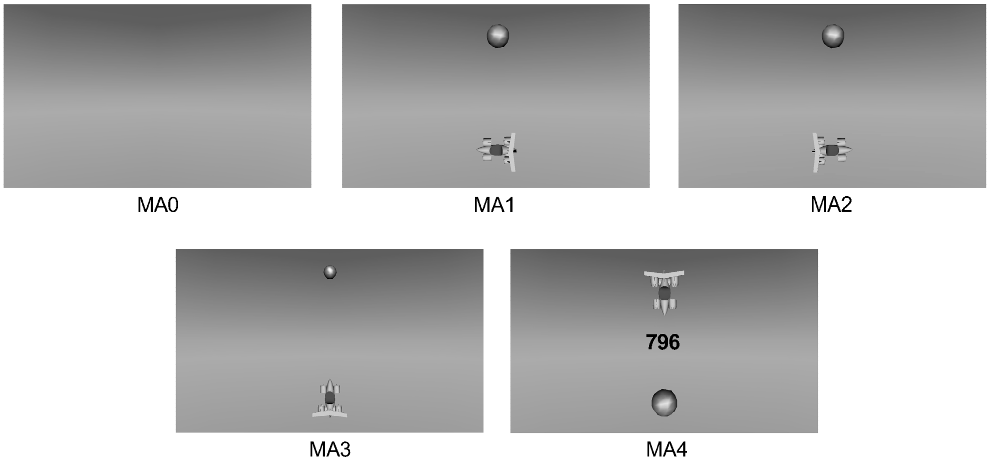

Figure 6.5. Visual cues used to indicate the MA that has to be performed during an active-time. The absence of the sun and spaceship signaled the outset of a break-time whose EEG-trials were considered as produced during the performance of MA0.

MA. The end of an active-time and consequently the outset of a break-time was signaled by the absence of the spaceship and the target in the CRE. The MA corresponding to each active-time was randomly chosen among the four MAs. Yet, we ensured that each MA was requested at least 22 times in each training-without-feedback session.

The EEG-trial duration and action period (see Chapter 2, Section 2.4 for the definition of these parameters) were set to 2000 and 500 milliseconds respectively. Using these values, nine EEG-trials per active-time and an average of four EEG-trials per break-time were potentially available. Indeed, an active or break-time was discarded if an artifact was detected in it. The presence of an ocular or muscular artifact was signaled to the subject by a vertical respectively horizontal oscillation of the sun. The number of EEG-trials per MA that were available at the end of the third training-without-feedback session is reported in Table 6.2 for each subject.

Table 6.2. Number of EEG-trials per MA and subject

MA	S1	S2	S3	S4	S5	S6
MA0	850	809	751	758	799	741
MA1	567	531	540	495	522	513
MA2	576	522	504	504	486	495
MA3	540	540	486	495	540	459
MA4	558	540	459	504	558	495

As mentioned in Chapter 4, different feature extraction methods (mappings) were considered for the recognition of each MA. Table 6.3 summarizes the parameters of those mappings that were common to all the subjects, namely the stationary PSD (denoted as ψ_{P}; see Chapter 4, Section 4.3), coherence (denoted as ψ_{C}; see Chapter 4, Section 4.4), and synchronization (denoted as ψ_{Y}; see Chapter 4, Section 4.8) mappings. The dimensions of the corresponding feature vectors were obtained from the formulae shown in Table 4.1 for N_e(number of electrodes)= 16 and N_{B}(number of frequency bands)= 7.

Table 6.3. Common mapping parameters

Mapping	Frequency bands [Hz]	Feature vector dimension
ψ_{P}	2-6, 6-10, 10-14, 14-18, 18-22, 22-26, 26-30	112
ψ_{C}	2-6, 6-10, 10-14, 14-18, 18-22, 22-26, 26-30	840
ψ_{Y}	2-6, 6-10, 10-14, 14-18, 18-22, 22-26, 26-30	840

The parameters and the dimension of the corresponding feature vectors (denoted as D) of the mappings based on linear prediction parametric models, i.e. the autoregressive (denoted as ψ_{AR}; see Chapter 4, Section 4.5), non-stationary AR (denoted as ψ_{NAR}; see Chapter 4, Section 4.6), and multivariate AR (denoted as ψ_{MVAR}; see Chapter 4, Section 4.7) are reported in Table 6.4. Notice that we chose the same orders Q_m and U_m for each channel. In fact, such values were chosen as the maximum of the values given by the MDL (see Section 4.5) criterion for each channel.

The optimal mapping for the recognition of a given MA was chosen by considering the recognition error associated with each possible mapping. We recall that an EEG-trial is considered as wrongly recognized if its membership is wrongly decided by the membership function (see Chapter 5, Section 5.5). Thus, among the recognition errors associated with

Table 6.4. Parameters of the mappings based on linear prediction models for each subject

Mapping	S1	S2	S3	S4	S5	S6
ψ_{AR} (Q_m/D)	3/48	2/32	2/32	2/32	4/64	4/64
ψ_{NAR} $(Q_m, U_m/\mathrm{D})$	3,2/240	2,1/96	2,2/160	2,3/224	4,2/320	4,1/192
ψ_{MVAR} (Q/D)	3/768	2/512	2/512	2/512	3/768	3/768

each mapping, the optimal mapping was the one associated with the lowest recognition error.

To obtain the recognition error associated with a given mapping and MA (which we call targeted MA), we proceeded as follows. Sixty percent of the EEG-trials (positive trials) available for the targeted MA and the same number of EEG-trials (negative trials) randomly chosen among the other MAs and MA0 were used to build the recognition models[1] using the algorithm explained in Chapter 5, Section 5.3. The rest of the positive EEG-trials and an equal number of EEG-trials (which were not used to build the recognition models) randomly chosen among the other MAs were used to determine the recognition error. To improve the quality of the recognition error estimate, the cross-validation procedure described in Chapter 5, Section 5.5.1 was used. Figure 6.6 shows the recognition errors for each mapping, MA, and subject. The recognition errors are reported in fractions, i.e. a recognition error equal to 0.5 implies that 50% of the tested EEG-trials were wrongly recognized.

Table 6.5 shows the optimal mapping and the associated recognition error for each MA, and subject. From the results reported in this Table, it appears that the optimal mapping choice for each MA is subject dependent. It is worth noticing that for each subject a dominant mapping can be distinguished. In particular, for subjects S1 and S3 the ψ_{NAR} and ψ_{Y} mappings are the optimal ones for each MA.

Figures 6.7 to 6.12 depict the experimental distribution (computed using the optimal mappings reported in Table 6.5) of the normalized membership (see Chapter 5, Section 5.2), associated with each MA for subjects S1 to S6 respectively. We recall that a positive EEG-trial is correctly recognized when its normalized membership is larger than one. On the other hand, a negative EEG-trial is correctly recognized when its normalized membership is smaller than minus one.

It is worth mentioning that positive and negative EEG-trials are relative to the targeted MA. Thus, an EEG-trial generated during the performance of MA1 is a positive trial with respect to MA1 and a negative one with respect to any other MA.

[1] As a matter of fact, the recognition models are composed of the membership parameters, namely the offset, threshold, and membership function (see Chapter 5 for details)

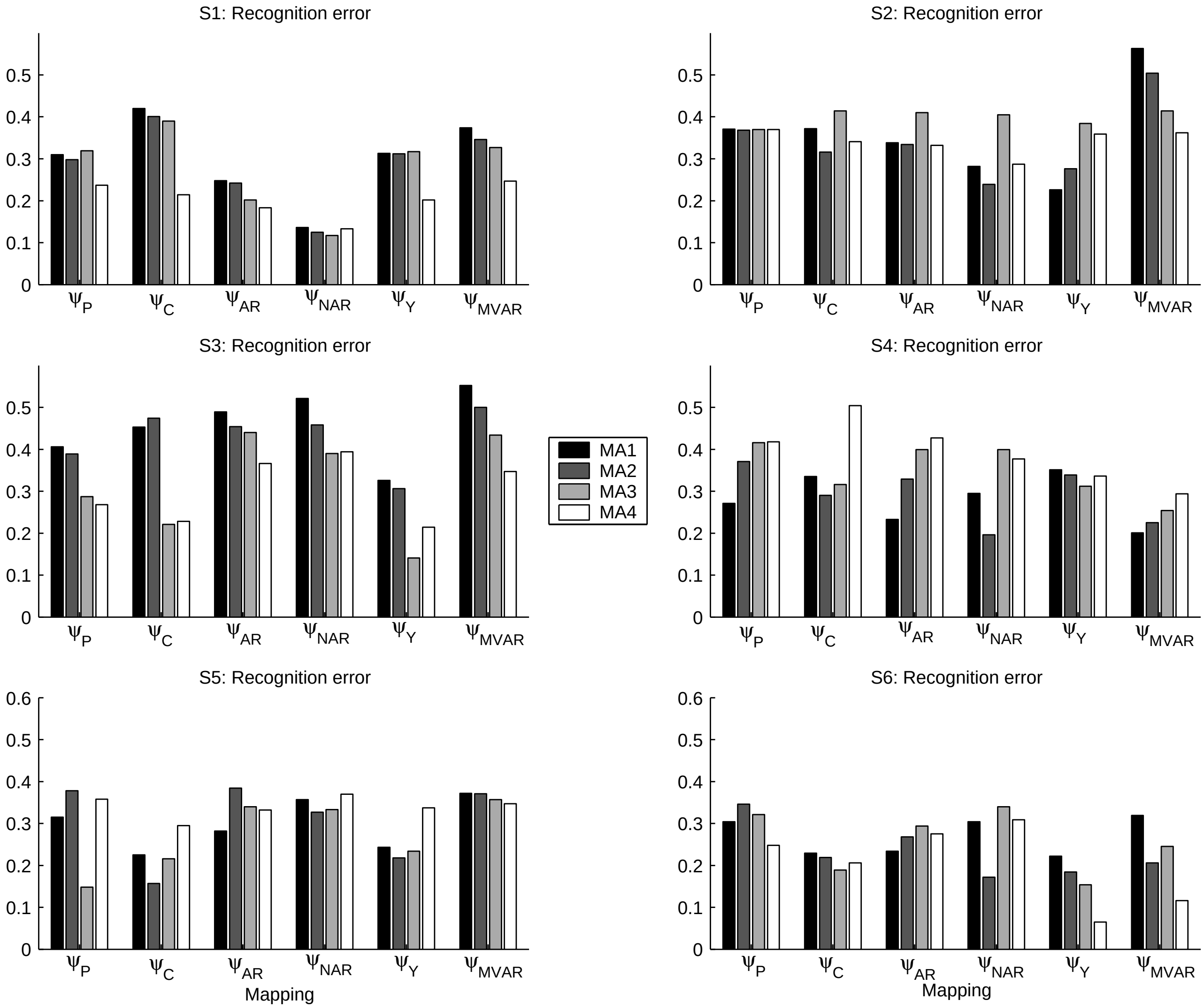

Figure 6.6. Recognition errors for each mapping, MA, and subject. The values are reported in fractions, i.e. a recognition error equal to 0.5 implies that 50% of the tested EEG-trials were wrongly recognized. The numerical values presented in this figure are reported in Appendix B, Section B.1. For a given MA, the mapping providing the smallest recognition error is chosen as the optimal mapping for this MA. For instance, in the case of subject S1, the non-stationary AR mapping constitutes the optimal one for each MA.
The optimal mappings and the corresponding recognition errors for each MA, and subject are reported in Table 6.5.

Table 6.5. Choice of the optimal mapping for each MA and subject[a]

MA	S1	S2	S3	S4	S5	S6
MA1	ψ_{NAR} (0.136)	ψ_{Y} (0.226)	ψ_{Y} (0.326)	ψ_{MVAR} (0.201)	ψ_{C} (0.225)	ψ_{Y} (0.222)
MA2	ψ_{NAR} (0.125)	ψ_{NAR} (0.239)	ψ_{Y} (0.306)	ψ_{NAR} (0.196)	ψ_{C} (0.157)	ψ_{NAR} (0.172)
MA3	ψ_{NAR} (0.117)	ψ_{P} (0.370)	ψ_{Y} (0.141)	ψ_{MVAR} (0.254)	ψ_{P} (0.148)	ψ_{Y} (0.154)
MA4	ψ_{NAR} (0.133)	ψ_{NAR} (0.287)	ψ_{Y} (0.214)	ψ_{MVAR} (0.294)	ψ_{C} (0.295)	ψ_{Y} (0.065)

[a] The numbers in parenthesis correspond to the associated recognition errors

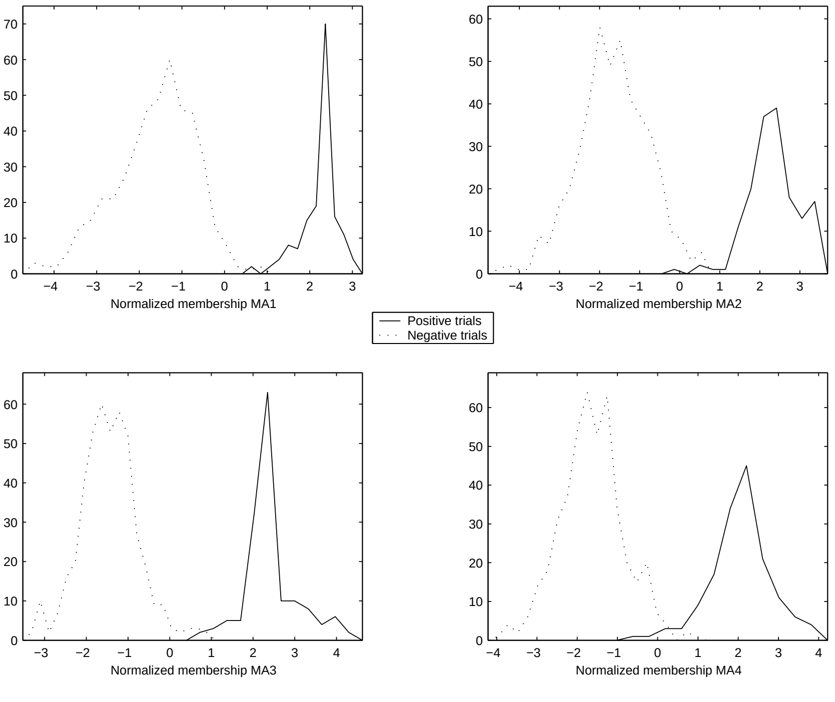

Figure 6.7. Subject S1: Distribution of the normalized memberships corresponding to positive and negative EEG-trials for each MA.

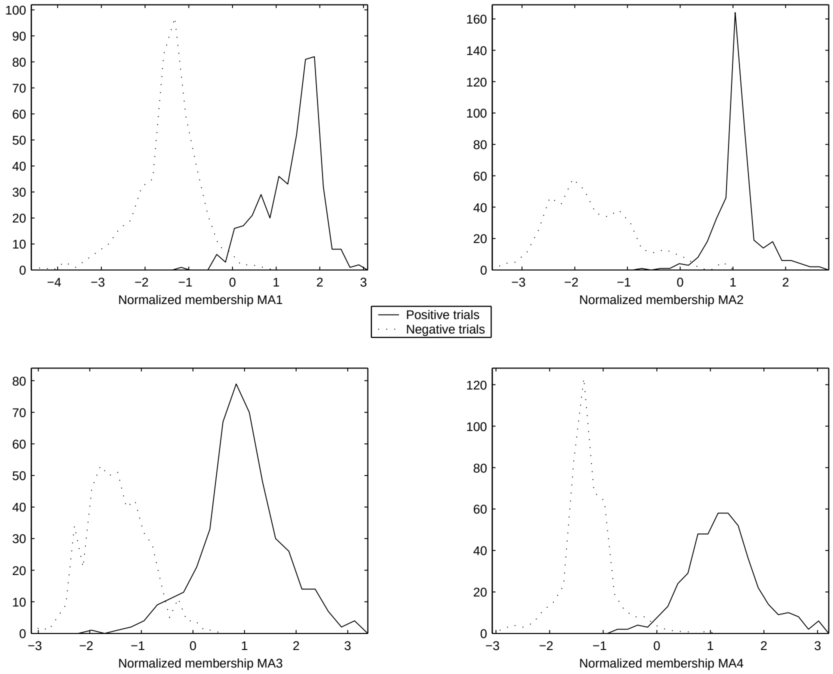

Figure 6.8. Subject S2: Distribution of the normalized memberships corresponding to positive and negative EEG-trials for each MA.

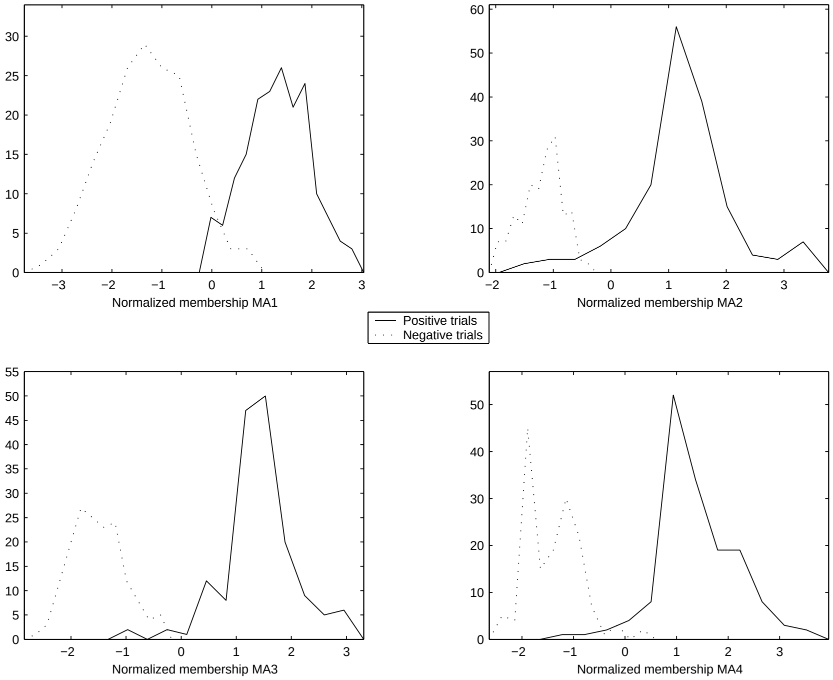

Figure 6.9. Subject S3: Distribution of the normalized memberships corresponding to positive and negative EEG-trials for each MA.

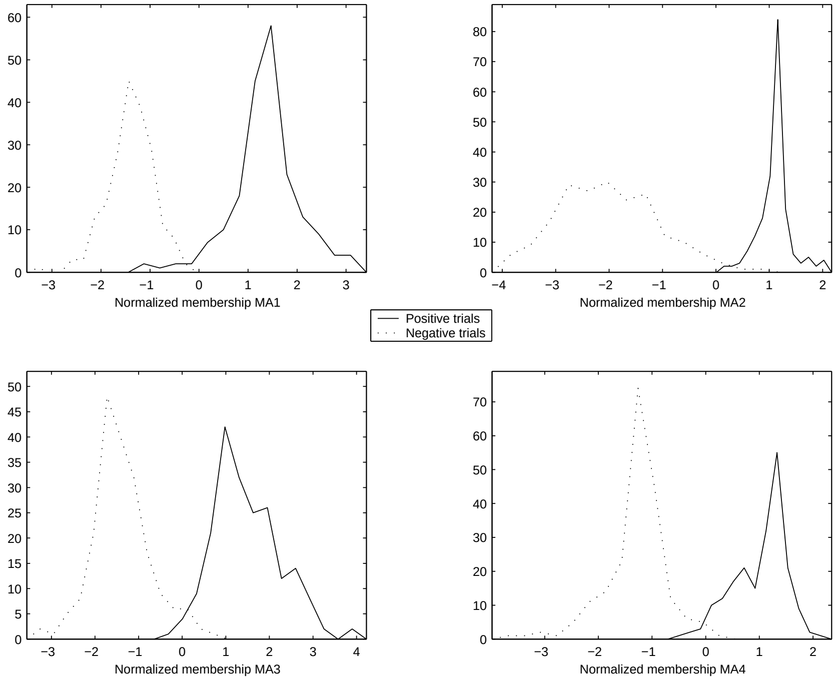

Figure 6.10. Subject S4: Distribution of the normalized memberships corresponding to positive and negative EEG-trials for each MA.

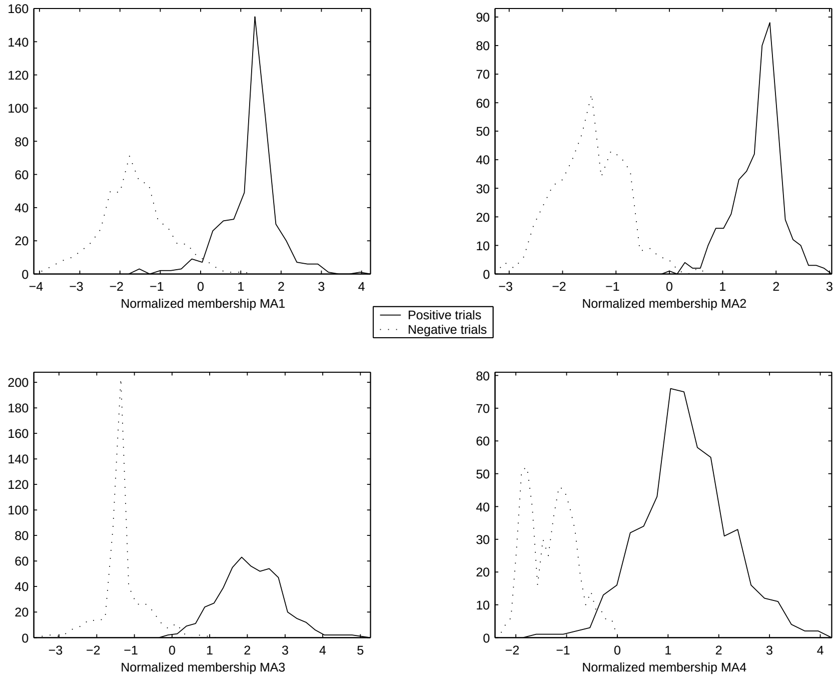

Figure 6.11. Subject S5: Distribution of the normalized memberships corresponding to positive and negative EEG-trials for each MA.

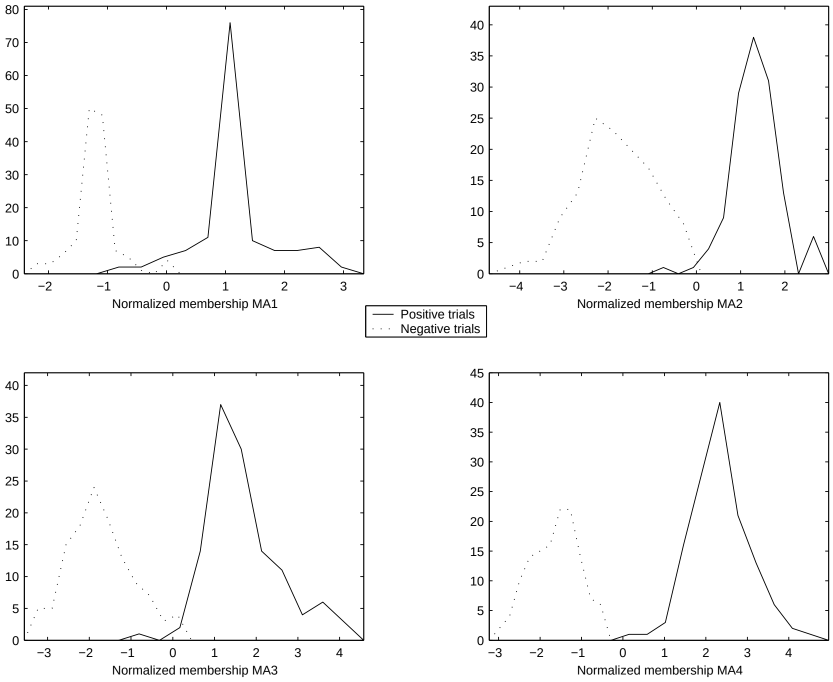

Figure 6.12. Subject S6: Distribution of the normalized memberships corresponding to positive and negative EEG-trials for each MA.

6.4 Action rules

The action rules governing the movement of the spaceship were set by considering the experimental distributions of the normalized membership of positive and negative trials as depicted in Figs. 6.7 to 6.12.

We first define the action strength (denoted as $\daleth_k$) associated with MAk as a piecewise linear function (see Fig. 6.13) of the normalized membership ζ_k:

$$\daleth_k = \begin{cases} 0 & \zeta_k \leqslant u_k \\ \frac{\zeta_k - u_k}{v_k - u_k} & u_k < \zeta_k \leqslant (v_k - u_k)\,\mathrm{M}_{\mathrm{stp}} + u_k \\ \mathrm{M}_{\mathrm{stp}} & u_k > (v_k - u_k)\,\mathrm{M}_{\mathrm{stp}} + u_k \end{cases} \tag{6.1}$$

where u_k is the smallest value of ζ_k for positive trials, v_k corresponds to the intersection between the positive and negative trials distribution of ζ_k (see Fig. 6.13), and $\mathrm{M}_{\mathrm{stp}}$ is the maximum number of steps that the spaceship is allowed to move in a single action. In our experiments $\mathrm{M}_{\mathrm{stp}}$ was equal to 8.

The slope of $\daleth_k$ was set so as to have $\daleth_k = 1$ for $\zeta_k = v_k$. The number of steps that the spaceship moved in the direction corresponding to MAk was equal to the nearest integer function of $\daleth_k$, namely nint($\daleth_k$).

For each EEG-trial, four strengths: $\daleth_1, \ldots \daleth_4$ were computed. The spaceship moved by the corresponding number of steps in each direction. Notice that in a single action the spaceship could simultaneously move in several directions. This possibility was allowed during the positioning tests (Section 6.5). However, in training-with-feedback sessions only the strength associated with the trained MA was considered.

6.5 Training with feedback

Training-with-feedback sessions were composed of two parts. The first part was organized similarly to the training-without feedback sessions, namely the first five to ten minutes were spent in the calibration procedure and the rest was divided into three five-minute slices separated by breaks of variable duration determined by the subject (see Fig. 6.14). Throughout this section the term session refers to a training-with-feedback one.

During active-times, in accordance with the values of the EEG-trial duration and the action period, the feedback was provided at a rate of two per second starting from second two after the presentation of the visual cue which indicated the MA that had to be performed (see Fig. 6.14).

The feedback consisted in moving the spaceship in the direction corresponding to the trained MA (see Table 6.1) by a number of steps determined by the action strength.

The number of times that a given MA was trained was proportional to the recognition error associated with such MA in the previous session. Thus, those MAs with relatively high recognition errors were trained more frequently.

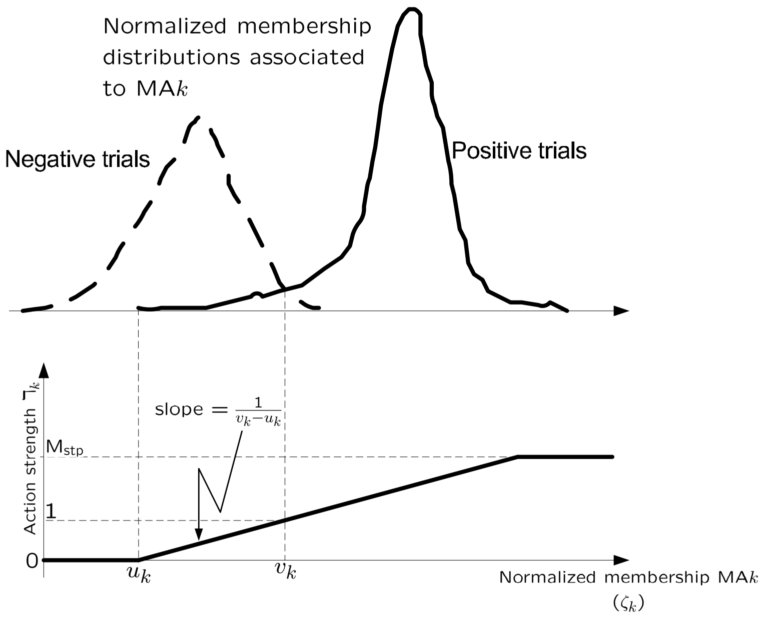

Figure 6.13. Action strength ($\daleth_k$) associated with MAk. The number of steps that the spaceship moved in the direction corresponding to MAk was equal to the nearest integer function of $\daleth_k$. The action strength depended on the experimental distribution of ζ_k. In particular, the value of the smallest ζ_k for positive trials (denoted as u_k) and the value of ζ_k corresponding to the intersection of the positive and negative trials distribution of ζ_k (denoted as v_k) determined the shape of $\daleth_k$. The action strength was limited by the maximum number of steps ($\mathrm{M_{stp}}$) that the spaceship was allowed to move in a single action.

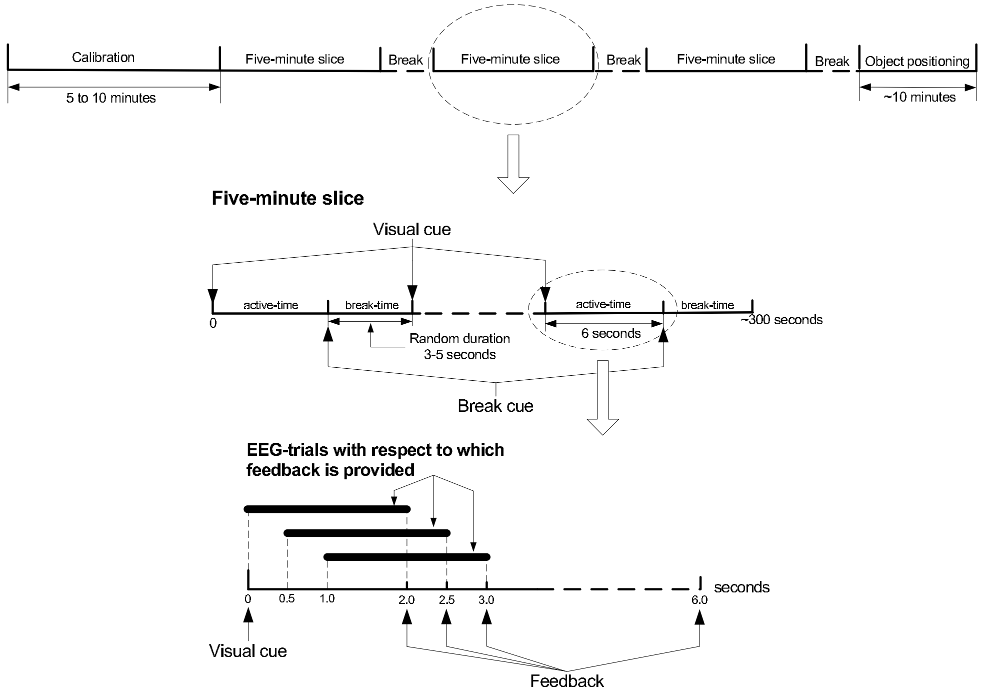

Figure 6.14. Protocol of a training-with-feedback session

The second part consisted in assessing the controlling skills acquired during the session by running two positioning tests. In each run, the positions of the target and the spaceship were randomized in such a way that at least fifty steps were necessary for the spaceship to reach the target. It is important to mention that the horizontal and vertical distances between the target and spaceship were both equal to twenty-five steps. As two bits are necessary to encode a single step in a given direction (up, down, left, or right), the number of bits that are needed to encode any optimal trajectory between the initial position of the spaceship and the target is equal to one-hundred bits. By measuring the average time a subject took to reach the target we can experimentally measure the bit-rate reached in the session.

As in training-without-feedback sessions, the EEG-trials with artifacts were discarded. In Appendix B, Section B.2 we report the number of EEG-trials that were available after artifact detection for each MA and subject.

The recognition errors (for each MA) corresponding to each session were determined using the recognition models updated at the end of the previous session. This makes sense since feedback was provided using such models. The recognition error of the first training-with-feedback session was determined using the recognition models built at the end of the training-without-feedback sessions.

The evolution through the sessions of the recognition errors associated with each MA, are reported in the four upper graphs in Figs. 6.15 to 6.20 for subjects S1 to S6 respectively. The values represented in these curves are reported in the Appendix B, Section B.2.

The recognition errors are represented with their two components, namely the false negative (FN) and false positive (FP) fractions. The FN is the fraction of positive EEG-trials whose normalized membership were smaller than one, and the FP is the fraction of negative EEG-trials whose normalized membership were larger than minus one. The recognition error is related to FN and FP by means of the following equation.

$$\text{Recognition error} = \frac{\text{FN}}{1+\frac{N_n}{N_p}} + \frac{\text{FP}}{1+\frac{N_p}{N_n}} \tag{6.2}$$

where N_n and N_p are the number of negative, respectively positive EEG-trials that were used to compute the recognition error.

To globally evaluate a session, we computed the theoretical bit rate (in bits per minute) by adapting the formula given in Chapter 2, Section 2.10 as follows.

$$\text{Bit rate} = \frac{60}{\text{T}_{\text{act}}}\left(\log_2 \text{N}_{\text{MA}} + (1-p_e)\log_2(1-p_e) + p_e \log_2 \frac{p_e}{\text{N}_{\text{MA}}-1}\right) \tag{6.3}$$

where p_e is the mean recognition error over the MAs, N_{MA} is the number of MAs (equal to four), and T_{act} is the action period in seconds (equal to 0.5 seconds). The theoretical bit rates for each session are represented on the lower left graph in Figs. 6.15 to 6.20 for subjects S1 to S6 respectively..

Since the bit rate computed by means of (6.3) considers EEG-trials that are free of artifacts, it constitutes an over optimistic estimation of the bit rate that can be achieved during actual BCI operation. Therefore, two positioning tests were carried out at the end of each session. As mentioned before, an experimental estimate of the bit rate can be obtained by dividing one hundred (i.e. the number of bits needed to encode an optimal trajectory) by the average time spent in reaching the target. The lower right graph in Figs. 6.15 to 6.20 depicts the bit rates experimentally estimated for subjects S1 to S6 respectively. Missing values correspond to those positioning tests in which the target could not be reached. These situations were due to the subjects who were free to interrupt the experiment at any time. The time spent in reaching the target for each positioning test, session, and subject are reported in Appendix B, Section B.2.

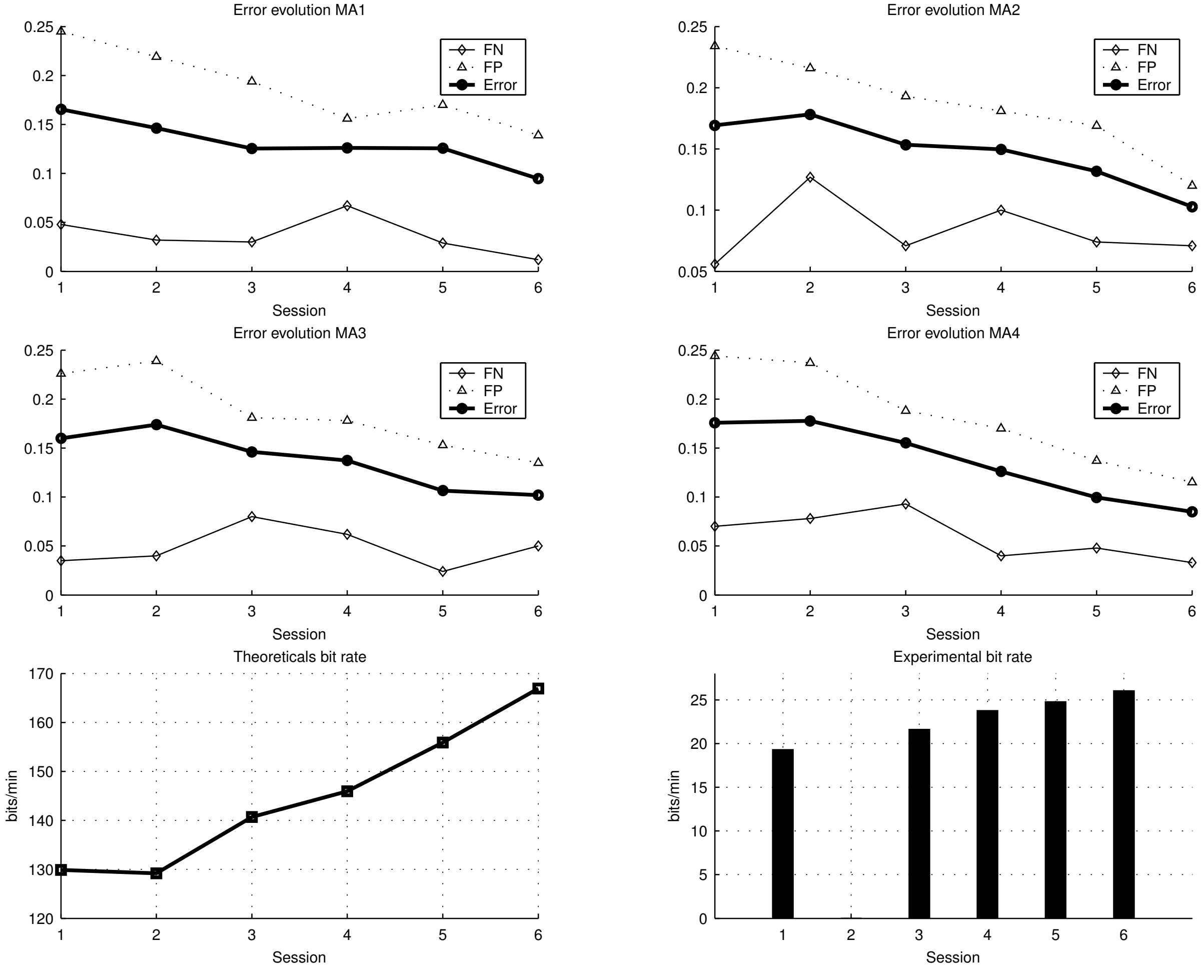

Figure 6.15. Subject S1.*Top four graphs*: Evolution of the recognition errors, associated with MA1 to MA4, throughout training-with-feedback sessions. The recognition errors are reported along with their two components, namely the false negative (FP) and false positive (FP) fractions. *Bottom left*: Theoretical bit rate. *Bottom right*: Experimental bit rate. Missing values correspond to positioning tests in which the target was not reached.
Numerical values presented in these graphs are reported in Appendix B, Section B.2

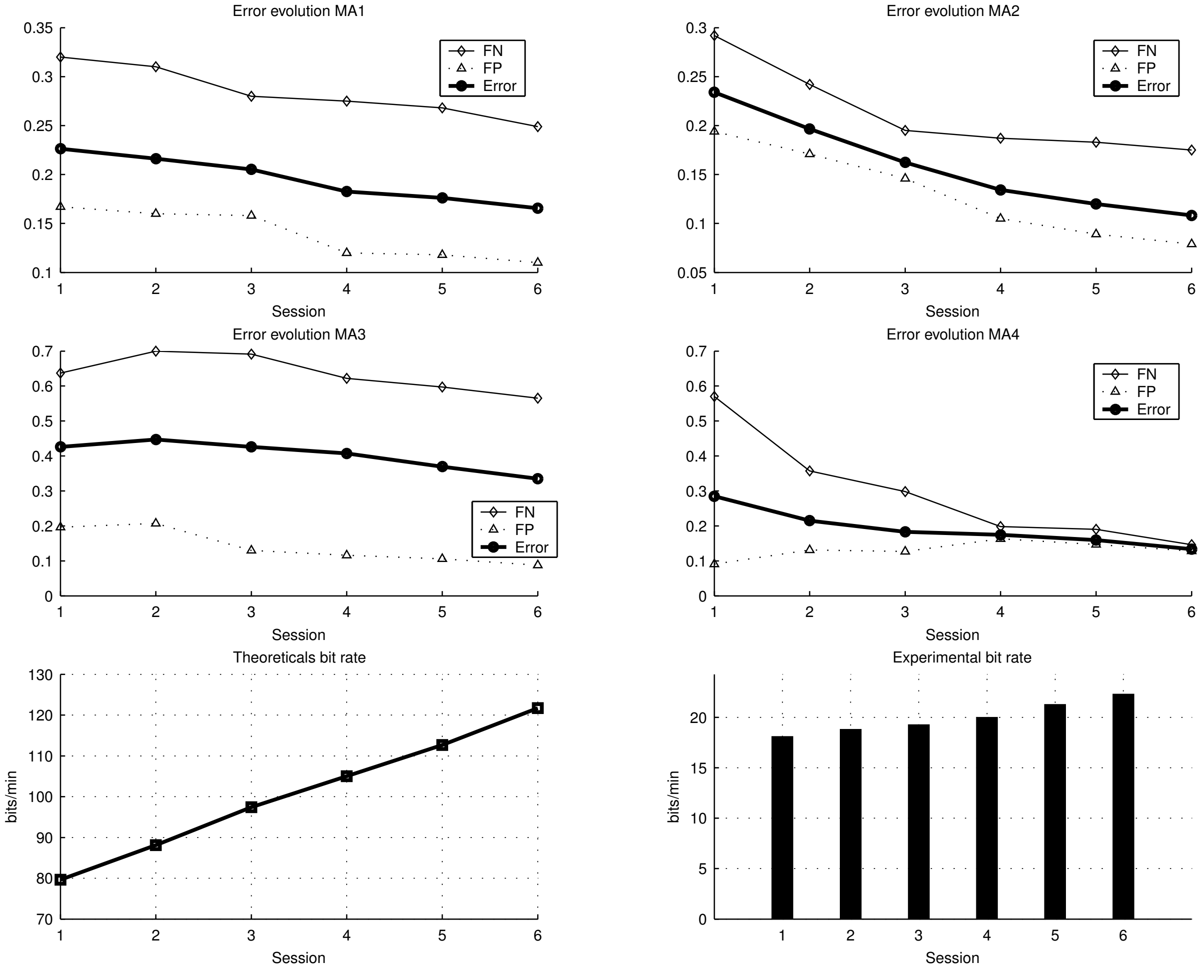

Figure 6.16. Subject S2. *Top four graphs*: Evolution of the recognition errors, associated with MA1 to MA4, throughout training-with-feedback sessions. The recognition errors are reported along with their two components, namely the false negative (FP) and false positive (FP) fractions. *Bottom left*: Theoretical bit rate. *Bottom right*: Experimental bit rate.
Numerical values presented in these graphs are reported in Appendix B, Section B.2

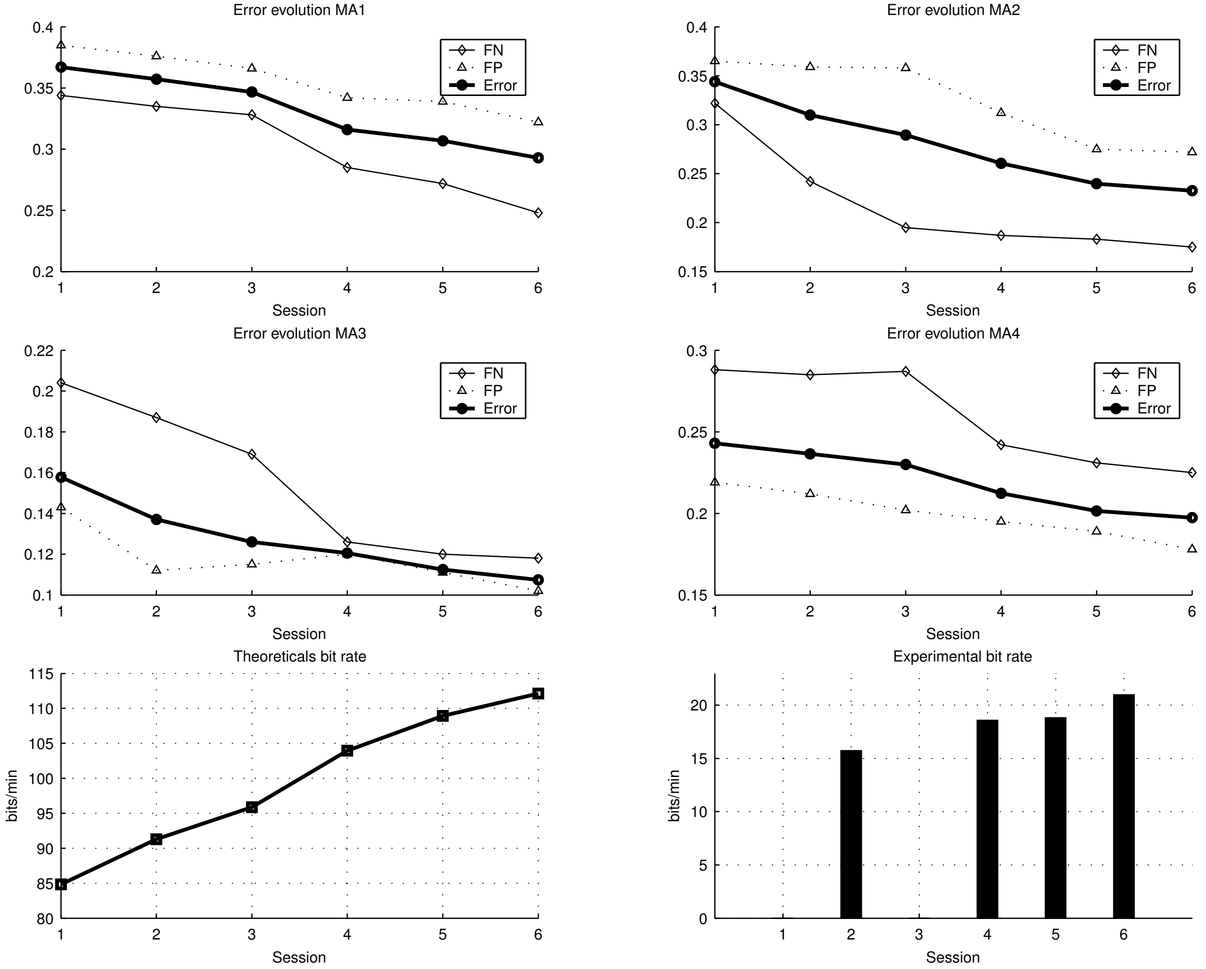

Figure 6.17. Subject S3.*Top four graphs*: Evolution of the recognition errors, associated with MA1 to MA4, throughout training-with-feedback sessions. The recognition errors are reported along with their two components, namely the false negative (FP) and false positive (FP) fractions. *Bottom left*: Theoretical bit rate. *Bottom right*: Experimental bit rate. Missing values correspond to positioning tests in which the target was not reached.

Numerical values presented in these graphs are reported in Appendix B, Section B.2

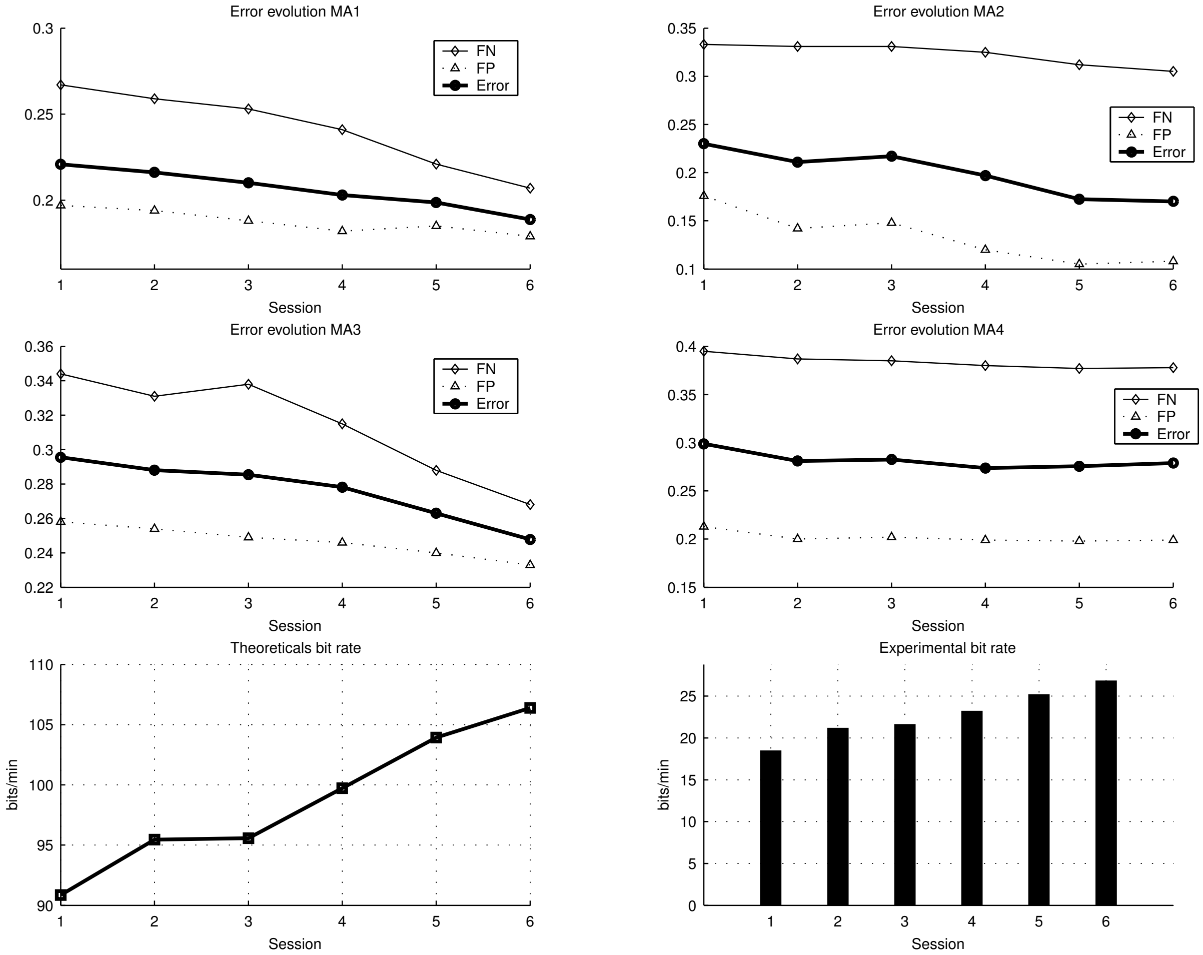

Figure 6.18. Subject S4. *Top four graphs*: Evolution of the recognition errors, associated with MA1 to MA4, throughout training-with-feedback sessions. The recognition errors are reported along with their two components, namely the false negative (FP) and false positive (FP) fractions. *Bottom left*: Theoretical bit rate. *Bottom right*: Experimental bit rate.
Numerical values presented in these graphs are reported in Appendix B, Section B.2

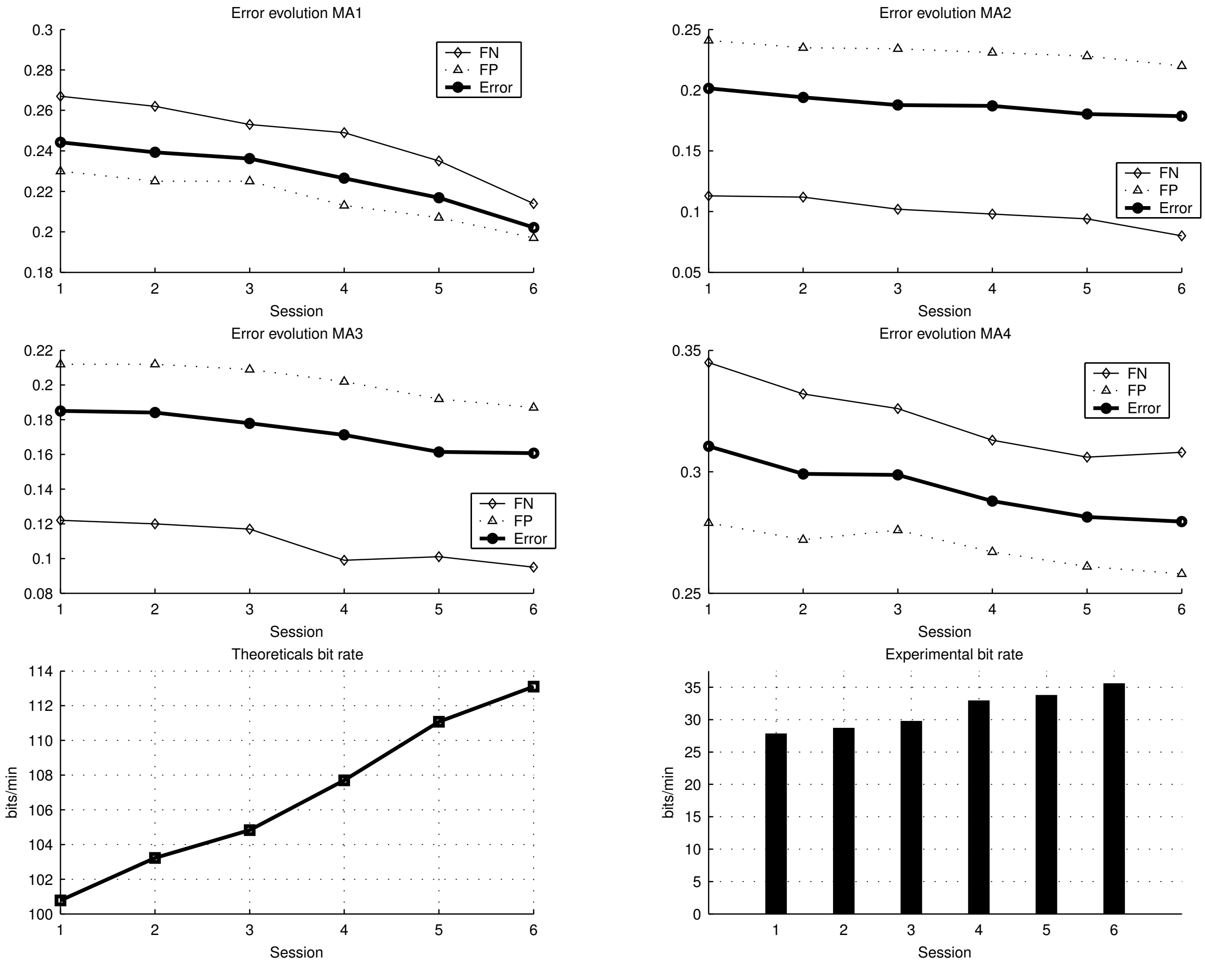

Figure 6.19. Subject S5.*Top four graphs*: Evolution of the recognition errors, associated with MA1 to MA4, throughout training-with-feedback sessions. The recognition errors are reported along with their two components, namely the false negative (FP) and false positive (FP) fractions. *Bottom left*: Theoretical bit rate. *Bottom right*: Experimental bit rate.
Numerical values presented in these graphs are reported in Appendix B, Section B.2

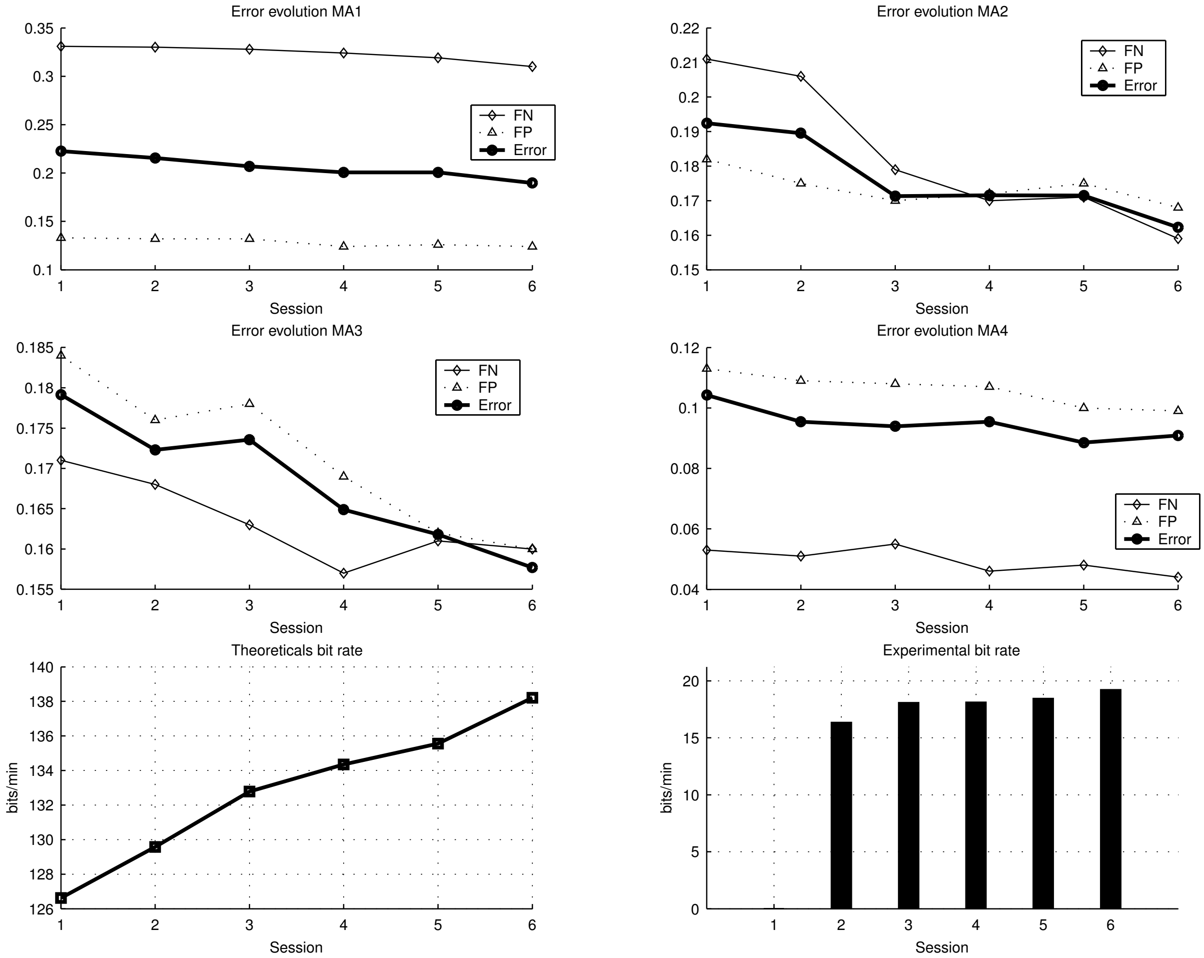

Figure 6.20. Subject S6.*Top four graphs*: Evolution of the recognition errors, associated with MA1 to MA4, throughout training-with-feedback sessions. The recognition errors are reported along with their two components, namely the false negative (FP) and false positive (FP) fractions. *Bottom left*: Theoretical bit rate. *Bottom right*: Experimental bit rate. Missing values correspond to positioning tests in which the target was not reached.
Numerical values presented in these graphs are reported in Appendix B, Section B.2

6.5.1 Discussion

The evolution of the recognition errors throughout the training-with-feedback sessions for each MA and subject (see Figs. 6.15 to 6.20) shows a clear downwards trend. Figure 6.21 depicts the relative decrease of the recognition error between two consecutive sessions for each MA, and subject. The relative decrease of the recognition error associated with MAk, between sessions $i + 1$ and $i \in \{1, \dots, 5\}$ was obtained by subtracting the recognition error associated with MAk corresponding to session $i + 1$ from that of session i. Negative values of the relative decrease indicate that the recognition error increased with respect to that of the previous session. Only a few negative values appear in Fig. 6.21. Notice that increases in the recognition error never affected the whole set of MAs. This explains why the theoretical bit rate, which can be thought of as an aggregate measure of the recognition errors associated with each MA, exhibits an upwards trend.

It is worth mentioning that whereas the false negative and false positive fractions do not necessarily exhibit a strict downwards trend (this is specially true for subject S1, see Fig. 6.15), the recognition models updating ensures that the recognition errors do not increase or at least not to the same extent. This sort of automatic control clearly appears in Fig. 6.15 in which, the increases in the false negative fraction curves are countered by corresponding decreases in the false positive fraction curves.

Figure 6.22 shows the relative increase of the theoretical (top) and experimental (bottom) bit rate over sessions for each subject. Relative increase of theoretical (experimental) bit rates between sessions $i + 1$ and i were obtained by subtracting the theoretical (experimental) bit rates corresponding to session i from that of session $i + 1$. The missing values in the experimental bit rates were replaced by the values corresponding to previous sessions, i.e. we assume that the controlling skills were maintained since subjects themselves decided to interrupt positioning testings. The hypothesis of controlling skills remanence seems reasonable, as the subsequent experimental bit rates always increase.

The theoretical bit rates for each subject almost always increase; the exception corresponds to the relative theoretical bit rate between sessions two and one for subject S1 which exhibit a slight decrease. This is confirmed by Fig. 6.21 in which the relative error decrease between sessions two and one for subject S1, shows that three out of the four MAs take negative values. Thus, the global theoretical evaluation of sessions indicates that the information transfer (measured by the bit rate) and consequently BCI operation improved over sessions.

A similar upwards trend is observed for the experimental bit rate. Since missing values were replaced by those corresponding to previous sessions, null increases correspond to those sessions. Yet, the relative experimental bit rate is always positive, meaning that when subjects carried out the positioning tests they always improved their previous performance.

At the end of the sixth session subjects reached experimental bit rates of 26, 22, 21, 27, 35, and 19 bits per minute respectively. Thus, an average of 22 bits per minute was achieved. This result situates our work among the most outstanding in the BCI research community (see Table 2.1). It is worth mentioning that while the theoretical bit rate gives much higher values, the experimental bit rate presents the advantage of being measured in the framework of a real application, and thus constitutes a closer approximation to the real information transfer rate.

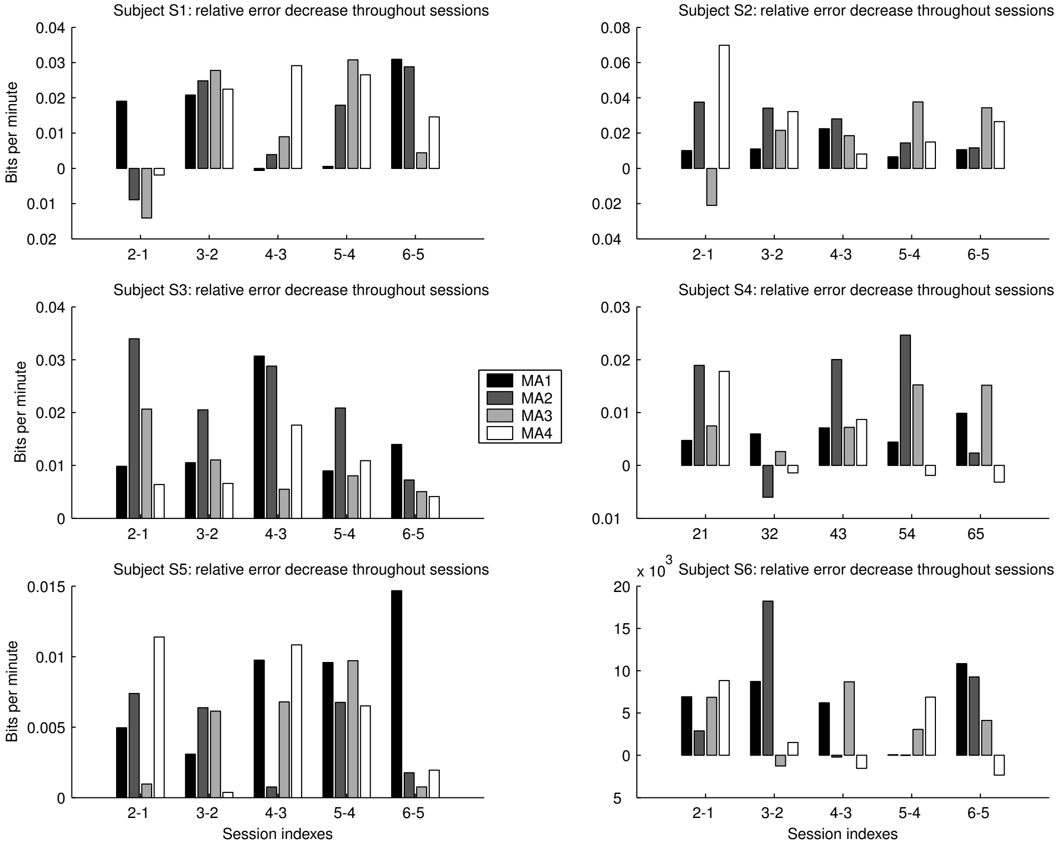

Figure 6.21. Relative error decrease over training-with-feedback sessions for each MA and subject. The relative decrease of the recognition error associated with MAk, between sessions $i+1$ and $i \in \{1, \ldots, 5\}$ was obtained by subtracting the recognition error of MAk corresponding to session $i+1$ from that of session i. Negative values of the relative decrease indicate that the recognition error increased with respect to that of the previous session.

6.6 Summary

In this chapter we applied the artifact detection, feature extraction, and pattern recognition algorithms developed in previous chapters, to the training of six subjects throughout nine training sessions in the framework of an asynchronous 2D positioning application. Four mental activities were used to move an object in the screen in four possible directions, namely left, right, up, and down.

The first three sessions served to build the initial recognition models through the choice of the optimal mappings for each MA and subject. Moreover, the action rules which determined the BCI operation were set with respect to the distribution of the normalized memberships associated with each MA.

The next six sessions were used to adjust the BCI recognition models and improve user controlling skills by means of feedback. Mental activities were trained on in such a way that those MAs that had large recognition errors were trained on more often.

In addition, in each of the last six sessions positioning tests were carried out to evaluate the subject controlling skills in a real application. The bit rate was estimated both theoretically and experimentally. Both estimates exhibited clear upwards trends throughout the sessions.

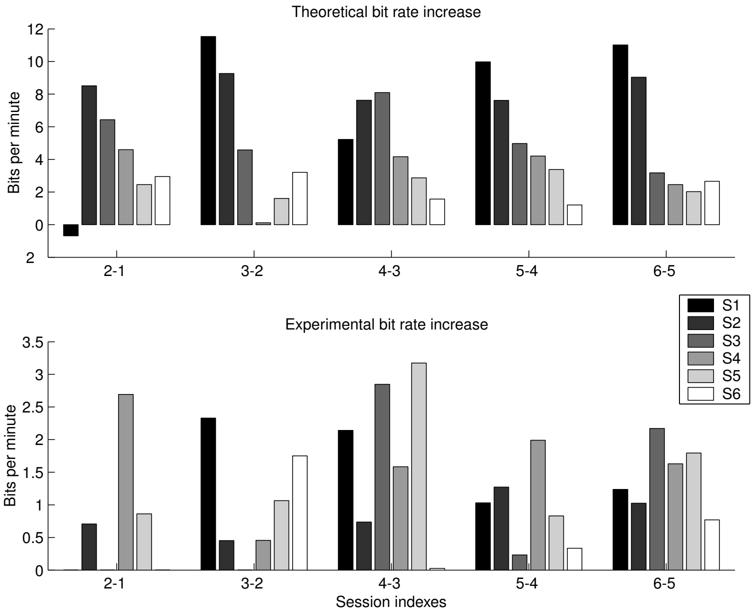

Figure 6.22. Theoretical (left) and experimental (right) bit rate increase over training-with-feedback sessions for each subject. Relative increase of theoretical (experimental) bit rates between sessions $i+1$ and i were obtained by subtracting the theoretical (experimental) bit rates corresponding to session i from that of session $i+1$. The missing values in the experimental bit rates were replaced by the values corresponding to previous sessions, i.e. we assume that the controlling skills were maintained.

7 Conclusions

In this chapter we review the most important issues and contributions presented in this thesis. Then we discuss possible extensions and continuations to the presented work.

The objectives of this thesis were to:

- Design and develop an asynchronous operant conditioning based BCI system which implements three adaptation levels, namely initial adaptation to the subject's signal characteristics, continuous adjustment of the BCI to maintain subject's controlling skills and reduce the impact of possible EEG changes, and subject adaptation through feedback.

- Ensure that the BCI is not controlled by other type of signals such as ocular and muscular artifacts.

- Design of efficient evaluation schemes and training protocols

7.1 Summary of achievements

The major achievements can be summarized as follows.

- An asynchronous operant conditioning BCI that operates with four mental activities in the framework of a 2D object positioning application was developed. Such BCI produces actions each half second based on the analysis of the last two second long EEG segment (EEG trial).

- Development of an efficient algorithm to detect ocular and muscular artifacts based on the kernel novelty detection framework. The parameters of this algorithm were set during a calibration procedure that took place before each experimental session. The BCI did not attempted to generate an action from an EEG trial with an artifact in it. Instead, it generated especial actions to notify the subject which type of artifact had been detected.

- Several types of feature extraction methods were considered to characterize the EEG trials produced during each controlling MA. From a general framework that considered the generalized interaction between the univariate signals composing EEG, different feature extraction methods (or mappings) were derived by assuming certain hypotheses on the nature of EEG. For a given MA, the mapping associated with the lowest recognition error was chosen. Thus, the BCI presented in this thesis used multiple types of feature vectors to operate.
- Recognition of MAs from feature vectors was done through the use of kernel based learning methods. We have developed an efficient theoretical framework which permits the dynamic updating of recognition models parameters as new training data become available.
- Definition of action rules that adapt the BCI operation mode to the subject performance. As the recognition models are dynamic these rules change accordingly.
- The algorithms and methods mentioned above were applied to the training of six subjects who participated in nine training sessions. The controlling skills acquired by subjects were measured using a theoretical and a experimental measure of the bit rate. Through the sessions, both measures increased for each subject. At the end of the ninth session, an average, over subjects of 126 and 25 bits per minute respectively was achieved. This result situates our research among the most outstanding ones in the BCI community.

7.2 Future directions

BCI research is still in its infancy, its continued success depends on further exploration of neuroscience results, psychological methods, signal processing and machine learning algorithms, evaluation criteria, operation modes, and applications. The following list contains a non-exhaustive number of proposals for further extensions to the work presented in this thesis:

- In this thesis we have considered a hierarchical model for the recognition of MAs, i.e. feature vector extraction and classification were done independently. A possible way to simultaneously consider the feature extraction and recognition problems would consist in applying the Bayesian framework which is able to select those features that make the recognition error lower.
- Generative models such as hidden Markov models (HMMs) were considered in the framework of synchronous BCI operation only. A generalization to asynchronous

operation can be made through ergodic HMMs. The training of such models could be done in two phases. In the first stage general states can be identified, i.e. characterizing all controlling MAs. In a second stage, the transitions that specifically characterize each MA can serve as recognition model for such an MA. The advantage of this approach resides in the fact that through the study of statistical properties of each state, valuable physiological insights into the nature of the MAs can be obtained.

- In this thesis we have considered four MAs that were chosen in accordance with current BCI studies which in turn made their choice based on hemispheric brain specialization studies. In addition to the selection of the optimal feature spaces in which these MAs can be recognized, it can be important to select the MAs as well. In this way, subjects could select the MAs through which they are best able to operate the BCI.

- The improvements obtained by using more adequate signal processing and machine learning algorithms aim at achieving large communication bandwidth as measured by the information transfer rate. To some extent, the most promising avenues for improvement will be determined by the particular application. Indeed, while higher information transfer rate is clearly desirable, the design of intelligent applications can handle much of the communication details. In this way, the subject can focus on communicating goals rather than on the details of control.

- BCI research should adhere to standards for designing studies and for assessing and comparing their results, both in the laboratory and in actual applications. In this way, direct comparisons among different BCI designs will be facilitated.

- The degrees of freedom required for adaptive automation of cognitive tasks, prosthetics, and complex robotics may lie beyond the range of current BCI signals and methods. However, BCIs can be used in combination with other human-computer interface devices. In particular, through the study of the mutual influence (through statistical measurements such as mutual information) between signals coming from other input devices and EEG, one can determine the extent to which EEG can enrich human-computer interaction.

- The position and number of electrodes can be optimized in accordance with physiological considerations and evaluation criteria. Feature selection algorithms can be used to rank the electrodes following their discriminative power among the MAs used to operate the BCI. Yet, a minimum number of electrodes should be maintained in order to cope with possible EEG changes.

Appendix A

A.1 Membership boundary induced by the Gaussian kernel

In Chapter 5 we described the membership parameters in the functional space H. In particular, we saw that the separating boundary between the images, under the map ϕ of the feature vectors belonging and not belonging to the target set, is a hyperplane. In this section we study the shape of the separation boundary in the feature vector space $\mathcal{X}$.

If the map ϕ uses a Gaussian kernel function, we know that a small value of the Gaussian kernel parameter lead to a small fractions of training error (FTE) (see Proposition 5.4). On the other hand, in real applications the joint distribution of feature vectors and labels associated to the training set does not necessarily reflect the real joint distribution. In addition, training errors are possible (this is particularly true in the BCI framework). Consequently, having a too small FTE is not required and even not suitable since the membership parameters can over-fit the training data and exhibit poor performance in determining the membership of unseen feature vectors.

In order to have a small FTE, one can intuitively understand that intricate separating boundaries in $\mathcal{X}$ are needed. As a matter of fact, it can be shown that there is connection between the minimization of the regularized risk and the "complexity" of the decision boundary [147]. Complexity in this context means that the separating boundary is highly irregular and intricate.

To illustrate the connection between the complexity of the separating boundary in $\mathcal{X}$ and the Gaussian kernel parameter we report the solutions obtained in the framework of a 2D toy problem.

In Fig. A.1 we depict the separating boundaries (between the dots and crosses) associated to different values of the Gaussian kernel parameter normalized to the minimum Euclidean distance in the training set (see Eq. 5.61) as it can be seen, the smaller σ_r the more complex the decision boundary and the smaller the FTE.

As discussed in Proposition 5.4, the fraction of support vectors (FSV) decreases as σ increases. In Fig. A.2 we report the distribution of the absolute value of the expansion coefficients in function of σ_r. For σ_r sufficiently small, the expansion coefficients are all equal to $\frac{\nu}{\mathrm{L}}$, where L is the number of training elements, and $\nu = \frac{1}{2}$ (see Proof of proposition 5.4) and consequently the FSV is equal to one.

In Section 5.3.5 we mentioned that the membership of a new feature vector is completely determined by the support vectors because their associated expansion coefficients are non-zero. In fact, the Gaussian kernel parameter is associated with the area of influence associated with a support vector. For a small σ, the area of influence is small and a large number of support vectors is required to define the separating boundary which is highly irregular. On the other hand, a large σ allow a support vector to have a strong influence over a larger area reducing thus the number of support vectors needed to define the separating boundary.

In Fig. A.3 we report the evolution of the FSV, the FTE and the estimation of the generalization error via cross-validation (GE) (see Section 5.5.1) for growing σ. As σ increases, the FTE increases and the FSV decreases. However, the GE exhibit a minimum for σ_r near eight. This value constitutes a good compromise between generalization and training errors. The optimal choice of σ is determined according to the procedure detailed in Section 5.5.2.

A.2 Computing the radius of the smallest sphere containing the training data in H

Let $\mathcal{S}_{\mathrm{tr}} = \{(x_l, y_l) | \, x_l \in \mathcal{X}, \, y_l \in \{-1, +1\}, \text{ and } l = 1, 2, \ldots, \mathrm{L}\}$ be the training set and ϕ the map from $\mathcal{X}$ into H.

We denote as R and $C \in \mathrm{H}$ the radius and the center, respectively of the smallest sphere containing the training data in H. The sphere parameters are found by solving:

$$\min R^2 \tag{A.1}$$

constrained to:

$$\|C - \phi(x_l)\|_{\mathrm{H}}^2 \leqslant R^2 \quad \text{for } l = 1, \ldots, \mathrm{L} \tag{A.2}$$

where $C \in$H is the center of the sphere. Introducing positive Lagrange multipliers $\lambda_1, \ldots, \lambda_{\mathrm{L}}$, we obtain the primal Lagrangian:

$$\Lambda_P = R^2 - \sum_{l=1}^{\mathrm{L}} \lambda_l \left(R^2 - \|C - \phi(x_l)\|_{\mathrm{H}}^2 \right) \tag{A.3}$$

Computing the derivatives of Λ_P with respect to R and C and setting them to zero

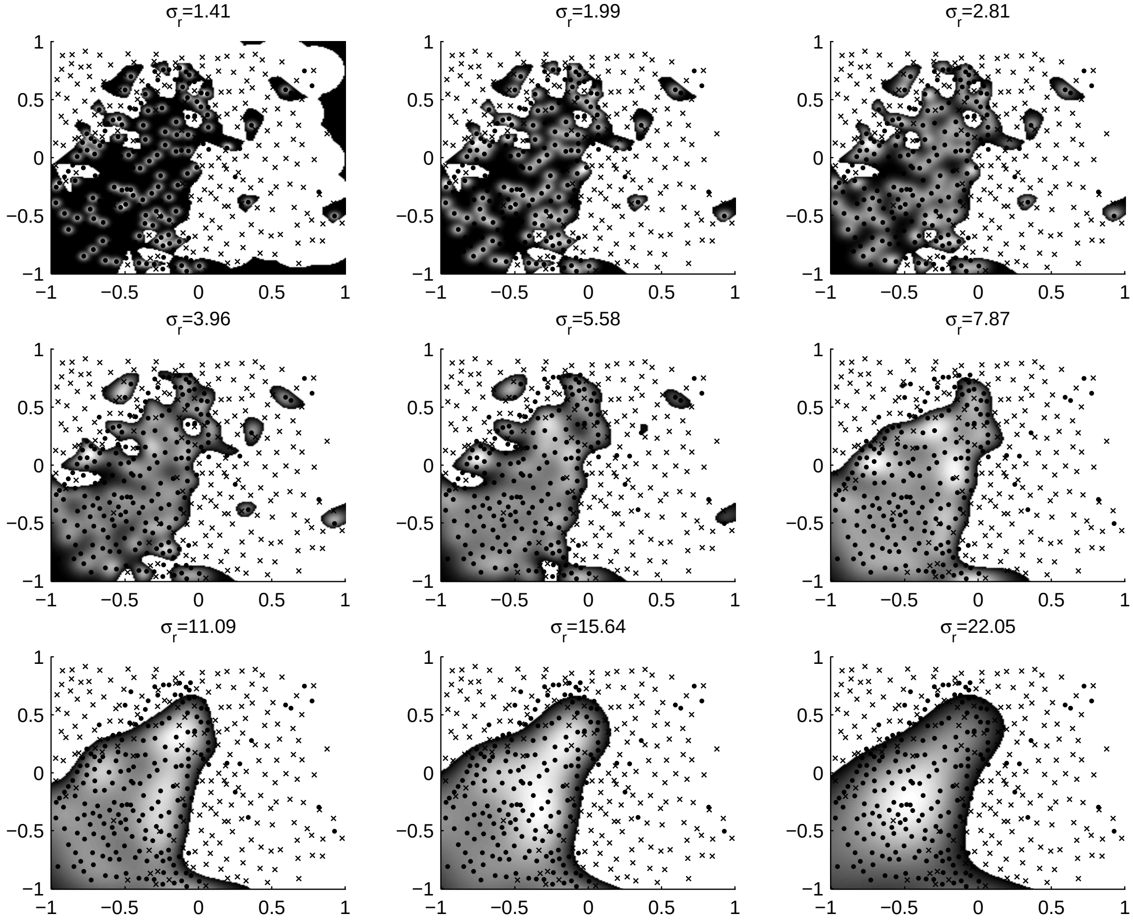

Figure A.1. Separating boundary (to discriminate between dots and crosses) for growing values of the Gaussian kernel parameter normalized to the minimum Euclidean distance in the training set (i.e. $\sigma_r = \frac{\sigma}{\Delta_min}$). The dark region encloses the dots. As σ_r increases more training errors are allowed and the separating boundary becomes more regular.

leads to

$$\partial_R \Lambda_P = 0 \quad \Rightarrow \quad \sum_{l=1}^{\mathrm{L}} \lambda_l = 1 \tag{A.4}$$

$$\partial_C \Lambda_P = 0 \quad \Rightarrow \quad C = \sum_{l=1}^{\mathrm{L}} \lambda_l \phi\left(x_l\right) \tag{A.5}$$

By replacing (A.4) and (A.5) in (A.3) we obtain the dual lagrangian that should be maximized with respect to $\lambda_1, \ldots, \lambda_{\mathrm{L}}$. Then,

$$R^2 = \max_{\lambda_1, \ldots, \lambda_{\mathrm{L}}} \left(\sum_{l=1}^{\mathrm{L}} \lambda_l - \sum_{l1,l2} \lambda_{l1} \lambda_{l2} K_\sigma\left(x_{l1}, x_{l2}\right) \right) \tag{A.6}$$

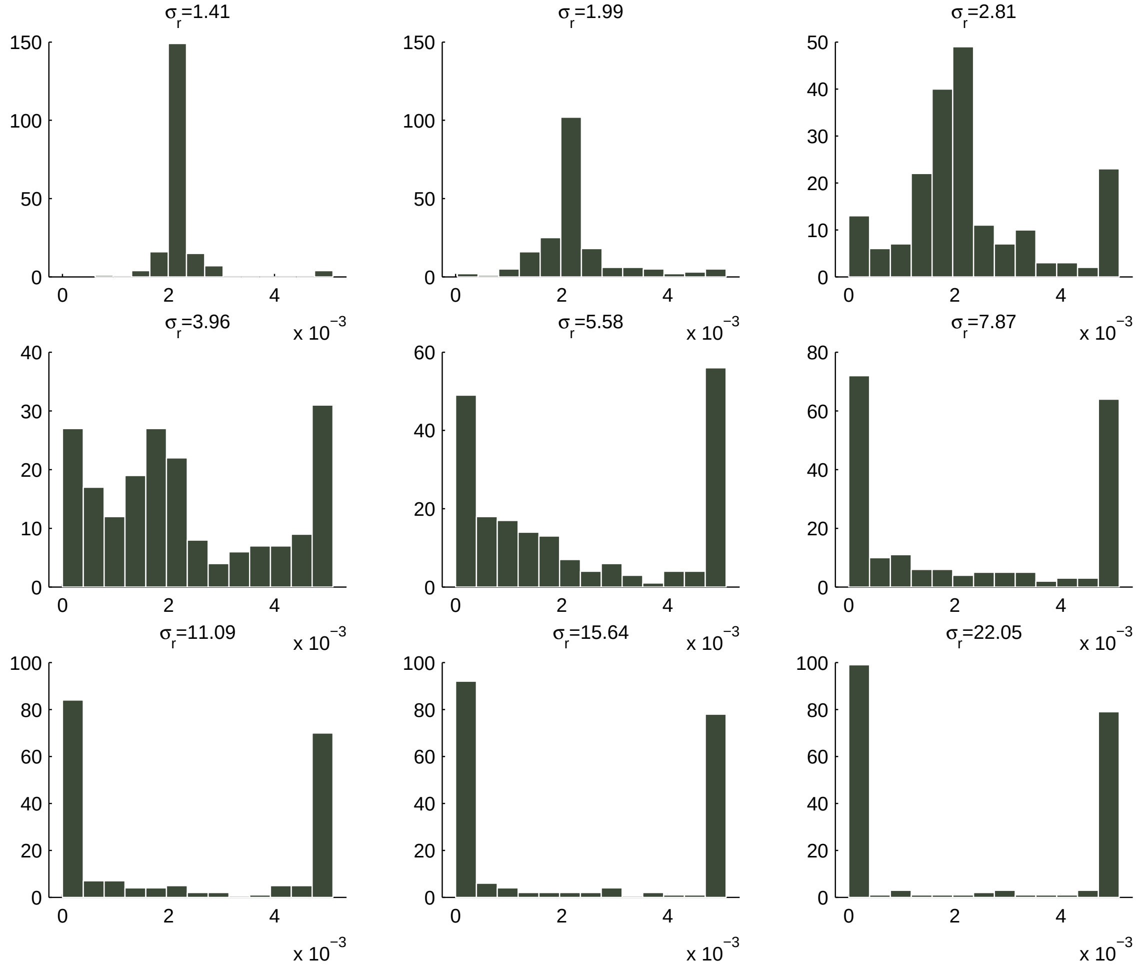

Figure A.2. Distribution of the absolute value of the expansion coefficients for growing values of σ_r. As σ_r increases more expansion coefficients become equal to zero. For large values of σ_r the distribution of the expansion coefficients becomes bimodal, i.e. concentrated in zero and $\frac{1}{\mathrm{L}}$ (the maximum allowed value).

subject to

$$\sum_{l=1}^{\mathrm{L}} \lambda_l = 1 \tag{A.7}$$

By replacing (A.7) in (A.6), the $\lambda_1, \ldots, \lambda_{\mathrm{L}}$ are found by solving:

$$(\lambda_1, \ldots, \lambda_{\mathrm{L}}) = \min_{\lambda_1,\ldots,\lambda_{\mathrm{L}}} \left(\sum_{l1,l2} \lambda_{l1} \lambda_{l2} K_\sigma \left(x_{l1}, x_{l2} \right) \right) \tag{A.8}$$

constrained to (A.7).

This convex quadratic problem [23] can be readily solved using standard quadratic programming techniques [168].

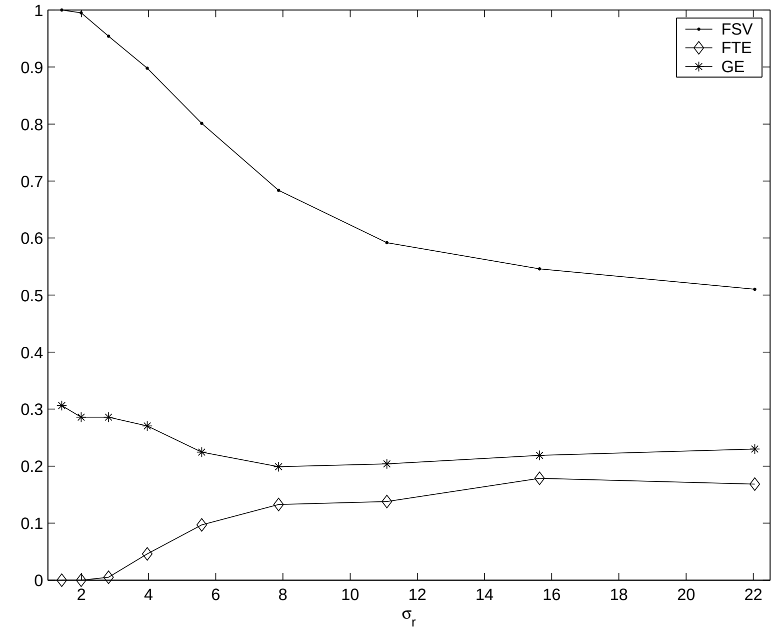

Figure A.3. Evolution of the fraction of support vectors (FSV), the fraction of training errors (FTE) and the cross-validation estimate of the generalization error (GE) in function of σ_r. While the FSV and the FTE exhibit a monotonically behavior in function of σ_r, the GE has a minimum which corresponds to the tradeoff between generalization and the FTE.

A.3 Computing the derivative of the theoretical bound $\mathcal{B}$ with respect to σ

To compute the derivative $\partial_\sigma \mathcal{B}$ we apply the chain rule and obtain:

$$\frac{\partial \mathcal{B}}{\partial \sigma} = \left.\frac{\partial \mathcal{B}}{\partial \sigma}\right|_{\vec{\alpha}\text{ fixed}} + (\nabla_{\vec{\alpha}} \mathcal{B})^{\mathrm{t}} \frac{\partial \vec{\alpha}}{\partial \sigma} \tag{A.9}$$

where $\vec{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{\mathrm{L}}, b)^{\mathrm{t}}$ and $^{\mathrm{t}}$ stands for the transpose operator.

Using (5.65), we have

$$\begin{aligned}\partial_\sigma \mathcal{B} = {} & \frac{1}{\mathrm{L}} \frac{\|w\|_{\mathrm{H}}^2}{\rho^2} \left.\frac{\partial R^2}{\partial \sigma}\right|_{\vec{\alpha}\text{ fixed}} + \frac{1}{\mathrm{L}} \frac{R^2}{\rho^2} \left.\frac{\partial \|w\|_{\mathrm{H}}^2}{\partial \sigma}\right|_{\vec{\alpha}\text{ fixed}} \\ & - \frac{2}{\mathrm{L}} \frac{\|w\|_{\mathrm{H}}^2 R^2}{\rho^3} \left.\frac{\partial \rho}{\partial \sigma}\right|_{\vec{\alpha}\text{ fixed}} + \frac{R^2}{\mathrm{L}} \left(\nabla_{\vec{\alpha}} \left(\frac{\|w\|_{\mathrm{H}}^2}{(\rho)^2}\right)\right)^{\mathrm{t}} \frac{\partial \vec{\alpha}}{\partial \sigma}\end{aligned} \tag{A.10}$$

A.3.1 Computing $\left.\frac{\partial R^2}{\partial \sigma}\right|_{\vec{\alpha}\text{ fixed}}$

This derivative can be computed directly from (A.6) using the following lemma [26]:

Lemma A.1. *Suppose we are given a vector $v_\zeta \in \mathbb{R}^d$ and a $d \times d$ matrix $\mathcal{A}_\zeta$ smoothly depending on a parameter ζ. Consider the function:*

$$\begin{aligned} \mathcal{G}(\zeta) &= \max_{u \in \Upsilon} \left(u^t v_\zeta - \frac{1}{2} u^t \mathcal{A}_\zeta u \right) \\ \Upsilon &= \left\{ u \middle| B^t u = a, u \geqslant 0, B \in \mathbb{R}^d, a \in \mathbb{R} \right\} \end{aligned}$$

Let $u^* \in \Upsilon$ be the vector where the maximum in $\mathcal{G}(\zeta)$ is attained. If this maximum is unique then

$$\frac{\partial \mathcal{G}(\zeta)}{\partial \zeta} = (u^*)^t \frac{\partial v_\zeta}{\partial \zeta} - \frac{1}{2} (u^*)^t \frac{\partial \mathcal{A}_\zeta}{\partial \zeta} u^*$$

In other words, it is possible to differentiate $\mathcal{G}$ with respect to ζ just as if u^* did not depend on ζ.

Thus, from (A.6) we have:

$$\left. \frac{\partial R^2}{\partial \sigma} \right|_{\vec{\alpha} \text{ fixed}} = -\sum_{l1,l2} \lambda_{l1} \lambda_{l2} \partial_\sigma K_\sigma (x_{l1}, x_{l2}) \tag{A.11}$$

A.3.2 Computing $\left. \frac{\partial \|w\|^2_{\mathbf{H}}}{\partial \sigma} \right|_{\vec{\alpha} \text{ fixed}}$

Using (5.32) we have:

$$\left. \frac{\partial \|w\|^2_{\mathrm{H}}}{\partial \sigma} \right|_{\vec{\alpha} \text{ fixed}} = \sum_{l1,l2} \tilde{\alpha}_{l1} \tilde{\alpha}_{l2} y_{l1} y_{l2} \frac{\partial K_\sigma (x_{l1}, x_{l2})}{\partial \sigma} \tag{A.12}$$

A.3.3 Computing $\left. \frac{\partial \rho}{\partial \sigma} \right|_{\vec{\alpha} \text{ fixed}}$

Using (5.50) we have:

$$\left. \frac{\partial \rho}{\partial \sigma} \right|_{\vec{\alpha} \text{ fixed}} = \frac{1}{2} \sum_{l=1}^{\mathrm{L}} y_l \tilde{\alpha}_l \left(\frac{\partial K_\sigma (x_l, x_{l1})}{\partial \sigma} - \frac{K_\sigma (x_l, x_{l2})}{\partial \sigma} \right) \tag{A.13}$$

A.3.4 Computing $\frac{\partial \vec{\alpha}}{\partial \sigma}$

We define the set $\mathcal{L} = \left\{ l \middle| 0 < \tilde{\alpha}_l < \frac{1}{\mathrm{L}} \right\}$ and the vector $\vec{\alpha}_{\mathcal{L}} = [\tilde{\alpha}_{l \in \mathcal{L}}, b]^{\mathrm{t}}$ (i.e. the vector composed of those $\tilde{\alpha}$'s corresponding to the $\phi(x_l)$ that are on the margins and the membership threshold b). Since $\tilde{\alpha}_{l \notin \mathcal{L}}$ is either 0 or $\frac{1}{\mathrm{L}}$, it is clear that:

$$\frac{\partial \tilde{\alpha}_{l \notin \mathcal{L}}}{\partial \sigma} = 0$$

.

Using the results shown in Fig. 5.7 and (5.40) we have:

$$\underbrace{\begin{pmatrix} K_Y & Y_{\mathcal{L}} \\ Y^{\mathrm{t}}_{\mathcal{L}} & 0 \end{pmatrix}}_{\mathcal{K}} \vec{\alpha}_{\mathcal{L}} = \rho \underbrace{\begin{pmatrix} \mathbf{1}_{|\mathcal{L}|} \\ 0 \end{pmatrix}}_{\mathcal{U}} \tag{A.14}$$

where $\mathbf{1}_{|\mathcal{L}|}$ is the $|\mathcal{L}|\times 1$ matrix with unitary elements, K_Y is a $|\mathcal{L}|\times|\mathcal{L}|$ matrix with elements $\mathcal{K}_{l,m\in\ \mathcal{L}} = y_l y_m K_\sigma\left(x_l, x_m\right)$ and $Y_\mathcal{L} = \left(y_{l\in\mathcal{L}}\right)$ (i.e. the $|\mathcal{L}|\times 1$ matrix of labels corresponding to the $\phi(x_l)$ that are on the margins).

The matrix $\mathcal{K}$ is always invertible, then:

$$\begin{aligned}\frac{\partial\vec{\alpha}_\mathcal{L}}{\partial\sigma} &= \frac{\partial\left(\mathcal{K}^{-1}\rho\mathcal{U}\right)}{\partial\sigma} & \text{(A.15)}\\ &= -\rho\mathcal{K}^{-1}\left(\frac{\partial\mathcal{K}}{\partial\sigma}\right)\mathcal{K}^{-1}\mathcal{U}+\mathcal{K}^{-1}\mathcal{U}\frac{\partial\rho}{\partial\sigma} & \text{(A.16)}\end{aligned}$$

The derivative $\frac{\partial\mathcal{K}^{-1}}{\partial\sigma}$ is computed using:

$$\begin{aligned}\mathcal{K}^{-1}\mathcal{K} &= \mathcal{I}\\ \Rightarrow\left(\partial_\sigma\mathcal{K}^{-1}\right)\mathcal{K}+\mathcal{K}^{-1}\partial_\sigma\mathcal{K} &= 0\\ \Rightarrow\frac{\partial\mathcal{K}^{-1}}{\partial\sigma} &= -\mathcal{K}^{-1}\left(\frac{\partial\mathcal{K}}{\partial\sigma}\right)\mathcal{K}^{-1}\end{aligned}$$

A.3.5 Computing $\nabla_{\tilde{\alpha}_\mathcal{L}}\|w\|^2_\mathbf{H}$

Using (5.32) we have

$$\begin{aligned}\frac{\partial\|w\|^2_\mathrm{H}}{\partial\tilde{\alpha}_{l\in\mathcal{L}}} &= \sum_{m\neq l|m\in\mathcal{L}} y_l y_m\tilde{\alpha}_m K_\sigma\left(x_l, x_m\right)+2\tilde{\alpha}_l & \text{(A.17)}\\ \frac{\partial\|w\|^2_\mathrm{H}}{\partial b_k} &= 0 & \text{(A.18)}\end{aligned}$$

A.3.6 Computing $\frac{\partial\rho}{\partial\vec{\alpha}_\mathcal{L}}$

Since $\phi\left(x_{l\in\mathcal{L}}\right)$ is in the margin

$$\begin{aligned}\frac{\partial\rho}{\partial\tilde{\alpha}_{l\in\mathcal{L}}} &= \frac{\partial}{\partial\tilde{\alpha}_{l\in\mathcal{L}}}\left(y_l\sum_{m\in\mathcal{L}} y_m\tilde{\alpha}_m K_\sigma\left(x_l, x_m\right)\right)=1 & \text{(A.19)}\\ \frac{\partial\rho}{\partial b} &= 0 & \text{(A.20)}\end{aligned}$$

Appendix B

B.1 Training without feedback sessions

Subject S1: Recognition error associated with each MA and mapping

Mapping		ψ_{P}	ψ_{C}	ψ_{AR}	ψ_{NAR}	ψ_{Y}	ψ_{MVAR}
MAs	**MA1**	0.310	0.420	0.248	**0.136**	0.313	0.374
	MA2	0.298	0.401	0.242	**0.125**	0.312	0.346
	MA3	0.319	0.390	0.202	**0.117**	0.317	0.327
	MA4	0.237	0.214	0.183	**0.133**	0.202	0.247

Table B.1.

Subject S2: Recognition error associated with each MA and mapping

Mapping		ψ_{P}	ψ_{C}	ψ_{AR}	ψ_{NAR}	ψ_{Y}	ψ_{MVAR}
MAs	**MA1**	0.371	0.372	0.338	0.282	**0.226**	0.563
	MA2	0.368	0.316	0.334	**0.239**	0.276	0.504
	MA3	**0.370**	0.414	0.410	0.405	0.384	0.414
	MA4	0.370	0.341	0.332	**0.287**	0.359	0.362

Table B.2.

Subject S3: Recognition error associated with each MA and mapping

Mapping		ψ_{P}	ψ_{C}	ψ_{AR}	ψ_{NAR}	ψ_{Y}	ψ_{MVAR}
MAs	**MA1**	0.406	0.453	0.489	0.521	**0.326**	0.552
	MA2	0.389	0.474	0.454	0.458	**0.306**	0.500
	MA3	0.287	0.221	0.440	0.390	**0.141**	0.434
	MA4	0.268	0.228	0.366	0.394	**0.214**	0.347

Table B.3.

Subject S4: Recognition error associated with each MA and mapping

Mapping		ψ_{P}	ψ_{C}	ψ_{AR}	ψ_{NAR}	ψ_{Y}	ψ_{MVAR}
MAs	**MA1**	0.271	0.335	0.233	0.295	0.351	**0.201**
	MA2	0.371	0.290	0.329	**0.196**	0.339	0.225
	MA3	0.416	0.316	0.399	0.399	0.312	**0.254**
	MA4	0.418	0.504	0.427	0.377	0.336	**0.294**

Table B.4.

Subject S5: Recognition error associated with each MA and mapping

Mapping		ψ_{P}	ψ_{C}	ψ_{AR}	ψ_{NAR}	ψ_{Y}	ψ_{MVAR}
MAs	**MA1**	0.315	**0.225**	0.282	0.357	0.243	0.372
	MA2	0.378	**0.157**	0.384	0.327	0.218	0.371
	MA3	**0.148**	0.216	0.340	0.333	0.234	0.357
	MA4	0.358	**0.295**	0.332	0.370	0.337	0.347

Table B.5.

Subject S6: Recognition error associated with each MA and mapping

Mapping		ψ_{P}	ψ_{C}	ψ_{AR}	ψ_{NAR}	ψ_{Y}	ψ_{MVAR}
MAs	**MA1**	0.304	0.229	0.234	0.304	**0.222**	0.319
	MA2	0.346	0.219	0.268	**0.172**	0.184	0.206
	MA3	0.321	0.189	0.294	0.340	**0.154**	0.245
	MA4	0.248	0.206	0.275	0.309	**0.065**	0.116

Table B.6.

B.2 Training with feedback sessions

Subject S1: number of EEG-trials after artifact detection

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	105	95	100	75	70	85
	MA2	90	110	70	90	95	85
	MA3	85	75	75	80	85	100
	MA4	100	90	75	75	105	90

Table B.7.

Subject S1: Recognition error evolution over training-with-feedback sessions FN/FP/Error[a]

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	0.048/0.245 0.165	0.032/0.219 0.146	0.030/0.194 0.126	0.067/0.156 0.126	0.029/0.170 0.126	0.012/0.139 0.095
	MA2	0.056/0.234 0.169	0.127/0.216 0.178	0.071/0.193 0.153	0.100/0.181 0.150	0.074/0.169 0.132	0.071/0.120 0.103
	MA3	0.035/0.226 0.160	0.040/0.239 0.174	0.080/0.181 0.146	0.062/0.178 0.137	0.024/0.153 0.106	0.050/0.135 0.102
	MA4	0.070/0.244 0.176	0.078/0.237 0.178	0.093/0.188 0.155	0.040/0.170 0.126	0.048/0.137 0.100	0.033/0.115 0.085

Table B.8.

[a] FN and FP stand for false negative and false positive fractions respectively. See Chapter 6, Section 6.5 for details

Subject S1: Positioning tests results

Session		1	2	3	4	5	6
Time [s]	**Test 1**	315	-	278	249	238	227
	Test 2	307	-	277	256	246	234
Exp. bit rate [bits/min]		19.29	-	21.62	23.76	24.79	26.03

Table B.9. Time spent in reaching the target and experimental bit rate.

Subject S2: number of EEG-trials after artifact detection

Session		1	2	3	4	5	6
MAs	**MA1**	323	342	361	342	361	399
	MA2	342	323	285	285	285	266
	MA3	513	513	589	608	608	589
	MA4	342	342	285	247	247	247

Table B.10.

Subject S2: Recognition error evolution over training-with-feedback sessions FN/FP/Error

Session		1	2	3	4	5	6
MAs	**MA1**	0.320/0.167 0.226	0.310/0.160 0.216	0.280/0.158 0.205	0.275/0.120 0.183	0.268/0.118 0.176	0.249/0.110 0.165
	MA2	0.292/0.194 0.234	0.242/0.171 0.196	0.195/0.146 0.162	0.187/0.105 0.134	0.183/0.089 0.120	0.175/0.079 0.108
	MA3	0.637/0.196 0.426	0.700/0.207 0.447	0.691/0.130 0.426	0.622/0.116 0.407	0.597/0.106 0.369	0.565/0.088 0.335
	MA4	0.570/0.091 0.285	0.357/0.131 0.215	0.298/0.127 0.183	0.198/0.163 0.175	0.190/0.147 0.160	0.146/0.128 0.133

Table B.11.

Subject S2: Positioning tests results

Session		**1**	**2**	**3**	**4**	**5**	**6**
Time [s]	**Test 1**	327	313	318	308	279	273
	Test 2	337	326	306	293	286	266
Exp. bit rate [bits/min]		18.07	18.78	19.23	19.97	21.24	22.26

Table B.12. Time spent in reaching the target and experimental bit rate.

Subject S3: number of EEG-trials after artifact detection

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	150	144	150	150	144	132
	MA2	138	132	132	126	114	132
	MA3	66	78	60	54	60	78
	MA4	96	90	102	108	96	114

Table B.13.

Subject S3: Recognition error evolution over training-with-feedback sessions FN/FP/Error

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	0.344/0.385 0.367	0.335/0.376 0.357	0.328/0.366 0.347	0.285/0.342 0.316	0.272/0.339 0.307	0.248/0.322 0.293
	MA2	0.322/0.365 0.344	0.242/0.359 0.310	0.195/0.358 0.289	0.187/0.312 0.261	0.183/0.275 0.240	0.175/0.272 0.232
	MA3	0.204/0.143 0.158	0.187/0.112 0.137	0.169/0.115 0.126	0.126/0.120 0.120	0.120/0.111 0.112	0.118/0.102 0.107
	MA4	0.288/0.219 0.243	0.285/0.212 0.237	0.287/0.202 0.230	0.242/0.195 0.212	0.231/0.189 0.201	0.225/0.178 0.197

Table B.14.

Subject S3: Positioning tests results

Session		**1**	**2**	**3**	**4**	**5**	**6**
Time [s]	**Test 1**	-	378	-	319	321	284
	Test 2	-	385	-	327	317	288
Exp. bit rate [bits/min]		-	15.73	-	18.58	18.81	20.98

Table B.15. Time spent in reaching the target and experimental bit rate.

Subject S4: number of EEG-trials after artifact detection

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	144	144	135	144	162	144
	MA2	135	144	153	153	135	126
	MA3	180	189	171	189	207	189
	MA4	207	189	198	171	198	207

Table B.16.

Subject S4: Recognition error evolution over training-with-feedback sessions FN/FP/Error

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	0.267/0.197 0.221	0.259/0.194 0.216	0.253/0.188 0.210	0.241/0.182 0.203	0.221/0.185 0.199	0.207/0.179 0.189
	MA2	0.333/0.176 0.230	0.331/0.142 0.211	0.331/0.148 0.217	0.325/0.120 0.197	0.312/0.105 0.172	0.305/0.108 0.170
	MA3	0.344/0.258 0.296	0.331/0.254 0.288	0.338/0.249 0.285	0.315/0.246 0.278	0.288/0.240 0.263	0.268/0.233 0.248
	MA4	0.395/0.213 0.299	0.387/0.200 0.281	0.385/0.202 0.283	0.380/0.199 0.274	0.377/0.198 0.276	0.378/0.199 0.279

Table B.17.

Subject S4: Positioning tests results

Session		**1**	**2**	**3**	**4**	**5**	**6**
Time [s]	**Test 1**	319	289	274	265	244	229
	Test 2	332	279	282	253	233	219
Exp. bit rate [bits/min]		18.43	21.13	21.58	23.17	25.16	26.79

Table B.18. Time spent in reaching the target and experimental bit rate.

Subject S5: number of EEG-trials after artifact detection

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	441	420	420	378	399	357
	MA2	315	357	357	315	378	294
	MA3	294	315	336	273	357	273
	MA4	588	546	546	504	525	483

Table B.19.

Subject S5: Recognition error evolution over training-with-feedback sessions FN/FP/Error

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	0.267/0.230 0.244	0.262/0.225 0.239	0.253/0.225 0.236	0.249/0.213 0.226	0.235/0.207 0.217	0.214/0.197 0.202
	MA2	0.113/0.241 0.202	0.112/0.235 0.194	0.102/0.234 0.188	0.098/0.231 0.187	0.094/0.228 0.180	0.080/0.220 0.179
	MA3	0.122/0.212 0.185	0.120/0.212 0.184	0.117/0.209 0.178	0.099/0.202 0.171	0.101/0.192 0.161	0.095/0.187 0.161
	MA4	0.345/0.279 0.310	0.332/0.272 0.299	0.326/0.276 0.299	0.313/0.267 0.288	0.306/0.261 0.281	0.308/0.258 0.279

Table B.20.

Subject S5: Positioning tests results

Session		**1**	**2**	**3**	**4**	**5**	**6**
Time [s]	**Test 1**	218	213	192	186	176	172
	Test 2	214	206	212	179	180	166
Exp. bit rate [bits/min]		27.78	28.64	29.70	32.88	33.71	35.50

Table B.21. Time spent in reaching the target and experimental bit rate.

Subject S6: number of EEG-trials after artifact detection

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	156	132	102	114	126	96
	MA2	120	114	90	90	108	84
	MA3	108	102	78	96	108	78
	MA4	48	60	42	48	60	42

Table B.22.

Subject S6: Recognition error evolution over training-with-feedback sessions FN/FP/Error

Session		**1**	**2**	**3**	**4**	**5**	**6**
MAs	**MA1**	0.331/0.133 0.223	0.330/0.132 0.216	0.328/0.132 0.207	0.324/0.124 0.201	0.319/0.126 0.201	0.310/0.124 0.190
	MA2	0.211/0.182 0.192	0.206/0.175 0.190	0.179/0.170 0.171	0.170/0.172 0.172	0.171/0.175 0.172	0.159/0.168 0.162
	MA3	0.171/0.184 0.179	0.168/0.176 0.172	0.163/0.178 0.174	0.157/0.169 0.165	0.161/0.162 0.162	0.160/0.160 0.158
	MA4	0.053/0.113 0.104	0.051/0.109 0.095	0.055/0.108 0.094	0.046/0.107 0.095	0.048/0.100 0.089	0.044/0.099 0.091

Table B.23.

Subject S6: Positioning tests results							
Session		**1**	**2**	**3**	**4**	**5**	**6**
Time [s]	**Test 1**	-	363	329	334	328	309
	Test 2	-	371	334	328	322	315
Exp. bit rate [bits/min]		-	16.35	18.10	18.13	18.46	19.23

Table B.24. Time spent in reaching the target and experimental bit rate.

Bibliography

[1] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–834, 1964.

[2] H. Akaike. Fitting autoregressions for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] M. Akay. *Time Frequency and Wavelets in Biomedical Signal Processing.* IEEE Press Series on Biomedical Engineering, 1998.

[5] J. Allanson. *upporting the Development of Electrophysiologically Interactive Computer Systems.* PhD thesis, Lancaster University, 2000.

[6] C.W. Anderson, E.A. Stolz, and S. Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45:277–286, 1998.

[7] F. Babiloni, F. Cincotti, L. Lazzarini, J. Millan, J. Mourino, M. Varsta, J. Heikkonen, L. Bianchi, and M. G. Marciani. Linear Classification of Low-Resolution EEG Patterns Produced by Imagined Hand Movements. *IEEE Transactions Rehabilitation Engineering*, 8(2):186–188, 2000.

[8] A.B. Barreto, S.D. Scargle, and M. Adjouadi. A practical emg-based human-computer interface for users with motor disabilities. *Journal of Rehabilitation Research and Development*, 37(1):53–63, 2000.

[9] A.R. Barron, J. Rissanen, and B. Yu. The mdl principle in modeling and coding. *IEEE Transactions on Information Theory - Special issue commemorating 50 years of information theory*, 44:2743–2760, 1998.

[10] P.J. Bartlett, B. Schölkopf, D. Schuurmans, and A.J. Smola, editors. *Advances in Large-Margin Classifiers (Neural Information Processing).* MIT Press, 2000.

[11] J.D. Bayliss. *A Flexible Brain-Computer Interface.* PhD thesis, Department of Computer Science University of Rochester, 2001.

[12] J.D. Bayliss. Use of the Evoked Potential P3 Component for Control in a Virtual Apartment. *IEEE Transactions Rehabilitation Engineering*, 11(2):113–116, June 2003.

[13] R. Beisteiner, P. Höllinger, G. Lindinger, W. Lang, and A. Berthoz. Mental representations of movements. Brain potentials associated with imagination of hand movements. *Electroencephalography and Clinical Neurophysiology*, 96:183–192, 1995.

[14] J.S. Bendat and G.G. Piersol. *Random Data Analysis and Measurement Procedures.* Wiley-Interscience, 2000.

[15] H. Berger. Ueber das elektrenkephalogramm des menschen. *Arch. Psichiatr. Nervenkr.*, 87:527–570, 1929.

[16] J. Bhattacharya. Reduced degree of long-range phase synchrony in pathological human brain. *Acta neurobiologiae experimentalis*, 61(4):309–318, 2001.

[17] N. Birbaumer, T. Hinterberger, A. Kübler, and N. Neumann. The Thought-Translation Device (TTD): Neurobehavioral Mechanisms and Clinical Outcome. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):120–123, 2003.

[18] N. Birbaumer, A. Kübler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor. The Thought Translation Device (TTD) for Completely Paralyzed Patients. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 8:190–193, 2000.

[19] G.E. Birch, S.G. Mason, and J.F. Borisoff. Current trends in brain-computer interface research at the Neil Squire foundation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):123–126, 2003.

[20] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[21] B. Blankertz, G. Curio, and K.-R. Müller. *Advances in Neural Information Processing Systems (NIPS 01)*, volume 14, chapter Classifying single trial EEG: Towards brain-computer interfacing. MIT Press, 2002.

[22] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer-Verlag, second edition, 1996.

[23] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[24] C.J.C. Burges. Uniqueness of the SVM Solution. In *Proceedings of the Twelfth Conference on Neural Information Processing Systems*. MIT Press, 1999.

[25] G.L. Calhoun, G.R. McMillan, D.F. Ingle, and M.S. Middendorf. Eeg-based control: Neurologic mechanisms of steady-state self-regulation. Technical Report AL/CF-TR-1997-0047, Wright-Patterson Air Force Base, 1997.

[26] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 46(1-3):131–159, 2002.

[27] Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Computation*, 14(12):2791–2846, 2002.

[28] M. Cheng, X. Gao, S. Gao, and D. Xu. Design and Implementation of a Brain-Computer Interface With High Transfer Rates. *IEEE Transactions on Biomedical Engineering*, 49(10):1181–1186, 2002.

[29] A. Choppin. EEG-Based Human Interface for Disabled Individuals: Emotion Expression with Neural Networks. Master's thesis, Tokyo Institute of Technology - Department of Information Processing, 2000.

[30] L. Cohen. *Time-Frequency Analysis*. Prentice Hall, 1995.

[31] R.J. Croft and R.J. Barry. Removal of ocular artifact from the eeg: a review. *Clinical Neurophysiology*, 30:5–19, 2000.

[32] N.E. Crone, D.L. Miglioretti, B. Gordon, J.M. Sieracki, M.T. Wilson, S. Uematsu, and R.P. Lesser. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. I. Alpha and beta event-related desynchronization. *Brain*, 121:2271–2299, 1998.

[33] F.H. Lopes da Silva. Neural mechanisms underlying brain waves: from neural membranes to networks. *Electroencephalography and Clinical Neurophysiology*, 79:81–93, 1991.

[34] L. Debnath and P. Mikusinski. *Introduction to Hilbert Spaces : With Applications*. Harcourt / Academic Press, 1998.

[35] J. Decety, D. Perani, M. Jeannerod, V. Bettinardi, B. Tadardy, R. Woods, J. Mazziotta, and F. Fazio. Mapping motor representations with positron emission tomography. *Nature*, 371:600–602, 1994.

[36] A. Delorme and S. Makeig. EEG Changes Accompanying Learned Regulation of 12-hz EEG Activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):133–137, June 2003.

[37] L Devroye, L. Gyröfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, 1996.

[38] S. Devulapalli. Nonlinear principal component analysis and classification of EEG during mental tasks. Master's thesis, Department of Computer Science, Colorado State Univ., 1996.

[39] E. Donchin and D.B. Smith. The contingent negative variation and the late positive wave of the average evoked potential. *Electroencephalography and Clinical Neurophysiology*, 29:201–203, 1970.

[40] E. Donchin, K.M. Spencer, and R.S. Wijesinghe. The mental prosthesis: Assessing the speed of a p300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering*, 8:174–179, 2000.

[41] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. *Advances in Neural Inf. Proc. Systems (NIPS 02)*, volume 15, chapter Combining features for BCI. MIT Press, 2003.

[42] J. Durbin. The Fitting of Time-series Models. *Review of the International Institute of Statistics*, 28:233–243, 1960.

[43] P.J. Durka. *Time-ferquency analyses of EEG*. PhD thesis, Institute of Experimental Physics Departement of Physics Warsaw University, August 1996.

[44] J.R. Evans and A. Abarbanel. *Introduction to Quantitative EEG and Neurofeedback*. Academic Press, 1999.

[45] E.V. Evarts. Pyramidal tract activity associated with a conditioned hand movement in the monkey. *Journal of Neurophysiology*, (29):293–301, 1966.

[46] L.A. Farwell and E. Donchin. Talking off the top of your head: A mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, (70):510–523, 1998.

[47] E.E. Fetz and D.V. Finocchio. Correlations between activity of motor cortex cells and arm muscles during operantly conditioned response patterns. *Experimental Brain Research*, 3(23):217–240, 1975.

[48] F. Findji, P. Catani, and C. Liard. Topographical distribution of delta rhythms during sleep: Evolution with age. *Electroencephalography and Clinical Neurophysiology*, 51(6):659–665, 1981.

[49] P. Flandrin. *Temps-fréquence*. Hermès Paris, 1993.

[50] W.J. Freeman. *Dynamics of sensory and cognitive processing by the brain*, chapter Nonlinear neural dynamics in olfaction as a model for cognition, pages 19–28. Springer Verlag, 1988.

[51] W.J. Freeman. *Induced rhythms in the brain*, chapter Predictions on neocortical dynamics derived from studies in paleocortex, pages 183–199. Springer Verlag, 1992.

[52] J. Freudiger, G.N. Garcia, T. Koenig, and T.Ebrahimi. Brain states analysis for direct brain-computer communication. Technical report, Swiss Federal Institute of Technology EPFL, July 2003.

[53] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[54] D. Galin and R.F. Ornstein. Lateral specialization of cognitive mode: An EEG study. *Psycophysiology*, 9:412–418, 1972.

[55] D. Galin and R.F. Ornstein. *Human Behavior and Brain Function*, chapter Hemispheric Specialization and the Duality of Consciousness, pages 3–23. Thomas Books, 1975.

[56] X. Gao, D. Xu, M. Cheng, and S. Gao. A BCI-based Environmental Controller for the Motion Disabled. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):137–140, June 2003.

[57] G.N. Garcia and T. Ebrahimi. Time-Frequency-Space Kernel for Single EEG-Trial Classification. In *Proceedings of the NORSIG conference*, 2002.

[58] G.N. Garcia, T. Ebrahimi, and J.-M. Vesin. Classification of EEG signals in the ambiguity domain for brain computer interface applications. In *Proceedings of the IEEE International Conference on Digital Signal Processing (DSP)*, volume 1, pages 301–305, July 2002.

[59] G.N. Garcia, T. Ebrahimi, and J.-M. Vesin. Correlative exploration of EEG signals for direct brain-computer communication. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 5, pages 816–819, 2003.

[60] G.N. Garcia, T. Ebrahimi, and J.-M. Vesin. Joint Time-Frequency-Space Classification of EEG in a Brain-Computer Interface Application. *EURASIP Journal on Applied Signal Processing*, 2003(7):713–729, 2003.

[61] G.N. Garcia, T. Ebrahimi, and J.-M. Vesin. Support vector EEG classification in the fourier and time-frequency correlation domains. In *Proceedings of the First IEEE EMBS Conference on Neural Engineering*, pages 591–594, March 2003.

[62] G.N. Garcia, T. Ebrahimi, J.-M. Vesin, and A. Villca. Direct Brain-Computer Communication with User Rewarding Mechanism. In *Proceedings of the IEEE International Symposium in Information Theory (ISIT)*, pages 221–221, July 2003.

[63] G.N. Garcia, U. Hoffmann, T. Ebrahimi, and J.-M. Vesin. Direct Brain-Computer Communication through EEG Signals. To appear in IEEE EMBS Book Series on Neural Engineering, 2004.

[64] D. Garret, D.A. Peterson, C.W. Anderson, and M.H. Thaut. Comparison of Linear, Nonlinear, and Feature Selection Methods for EEG Signal Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):141–144, June 2003.

[65] A Gevins, M. Smith, L. McEvoy, H. Leong, and J. Le. Electroencephalographic imaging of higher brain function. *Philisophical Transactions of the Royal Society*, (354):1125–1134, 1999.

[66] A.A. Glover, M.C. Onofrj, M.F. Ghilardi, and I. Bodis-Wollner. P300-like potentials in the normal monkey using classical conditioning and the auditory oddball paradigm. *Electroencephalography and Clinical Neurophysiology*, 65:231–235, 1986.

[67] R.M. Golden. Digital filter synthesis by sampled-data transformation. *IEEE Transactions Audio and Electroacustics*, AU-16:321–329, 1968.

[68] I.I. Goncharova, D.J. McFarland, T.M. Vaughan, and J.R. Wolpaw. Eeg-based brain-computer interface (bci) communication: scalp topography of emg contamination. *Soc Neurosci Abstr*, 26:1229, 2000.

[69] C. Guger, G. Edlinger, W. Harkam, I. Niedermayer, and G. Pfurtscheller. How Many People are Able to Operate an EEG-Based Brain-Computer Interface (BCI). *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):145–147, 2003.

[70] C. Guger, A. Schlgl, C. Neuper, D. Walterspacher, T. Strein, and G. Pfurtscheller. Rapid Prototyping of an EEG-Based BrainComputer Interface (BCI). *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 9(1):49–58, 2001.

[71] S. Gupta and H. Singh. Preprocessing EEG signals for direct human-system interface. In *IEEE International Joint Symposia on Intelligence and Systems (IJSIS)*, pages 32–37, November 1996.

[72] C. Harris, X. Hong, and Q. Gan. *Adaptive Modelling Estimation and Fusion from Data.* Springer-Verlag, 2002.

[73] T. Hinterberger, A. Kübler, J. Kaiser, N. Neumann, and N. Birbaumer. A braincomputer interface (BCI) for the locked-in: comparison of different EEG classifications for the thought translation device. *Clinical Neurophysiology*, 114:416–425, 2003.

[74] K. Hirano, S. Nishimura, and S.K. Mitra. Design of Digital Notch Filters. *IEEE Transactions on Circuits and Systems*, 22(7):964–970, 1974.

[75] J.E. Huggins, S.P. Levine, S.L. BeMent, R.K. Kushwaha, L.A. Schuh, M.M. Rohde, and D.A. Ross. Detection of event related potentials for development of a direct brain interface. *Journal of Clinical Neurophysiology*, 16:448–455, 1999.

[76] D.R. Humprey. Representation of movements and muscles within the primate precentral motor cortex: historical and current perspectives. *Federation Proceedings*, 12(45):2687–2699, 1986.

[77] M. Jachan, G. Matz, and F. Hlawatsch. Time-Frequency-Autoregressive Random Processes: Modeling and Fast Parameter Estimation. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 6, pages 125–128, 2003.

[78] H. Jasper. Report on committee on methods of clinical exam in EEG. *Electroencephalogry and Clinical Neurophysiology*, 7:370–375, 1958.

[79] H.H. Jasper. The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10(1):371–375, 1958.

[80] H.H. Jasper and W. Penfield. Electrocorticograms in man: effect of the voluntary movement upon the electrical activity of the precentral gyrus. *Arch. Psychiat. Z. Neurol.*, 183:163174, 1949.

[81] M.J. Jeannerod. Mental imagery in the motor context. *Neuropsychologia*, 33(11):1419–1432, 1995.

[82] M. Johansson. The Hilbert Transform. Master's thesis, Växjö, 1999.

[83] J. Kalcher, D. Flotzinger, and G. Pfurtscheller. A New Approach to a Brain-Computer-Interface (BCI) based on Learning Vector Quantization (LVQ3). In *Proceedings of the IEEE EMBS International Conference*, volume 14, pages 1658–1659, 1992.

[84] J. Kamiya. Conditioned discrimination of the EEG alpha rhythm in humans. Paper presented at the Western Psychological Association, 1962.

[85] S.M. Kay. *Modern Spectral Estimation: Theory and Application.* Prentice-Hall, 1988.

[86] Z.A. Keirn and J.I. Aunon. Man-Machine Communications Through Brain-Wave Processing. *IEEE Engineering in Medicine and Biology Magazine*, 9(1):55–57, 1990.

[87] S Kelly, D. Burke, P. de Chazal, and R. Reilly. Parametric Models and Spectral analysis for Classification in Brain-Computer Interfaces. In *Proceedings of the IEEE International Conference on Digital Signal Proceesing*, 2002.

[88] P.R. Kennedy. The cone electrode: a long-term electrode that records from neurites grown onto its recording surface. *Journal of Neuroscience Methods*, (29):181–193, 1989.

[89] P.R Kennedy and K.D. Adams. A decision tree for brain-computer interface devices. *IEEE Transactions on Neural Systems and Rehabilitation Enginnering*, 11(2):148–150, 2003.

[90] P.R. Kennedy and R.A. Bakay. Restoration of neural output from a paralyzed patient by a direct brain connection. *NeuroReport*, (9):1707–1711, 1998.

[91] P.R. Kennedy, R.A.E. Bakay, M.M. Moore, K. Adams, and J. Goldwaithe. Direct control of a computer from the human central nervous system. *IEEE Transactions on Rehabilitation Engineering*, (8):198–202, 2000.

[92] G.S. Kimeldorf and G.Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, (33):82–95, 1971.

[93] J. Kivinen, A.J. Smola, and R.C. Williamson. Online Learning with Kernels. Available at http://citeseer.nj.nec.com/kivinen02online.html, 2002.

[94] T. Koenig, K. Kochi, and D. Lehmann. Event-related electric microstates of the brain differ between words with visual and abstract meaning. *Electroencephalography and Clinical Neurophysiology*, 106(6):535–546, 1998.

[95] H.W. Kuhn and A.W. Tucker. Nonlinear programming. In *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, pages 481–492, 1951.

[96] J.P. Lachaux, E. Rodriguez, J. Martinerie, and F.J. Varela. Measuring Phase Synchrony in Brain Signals. *Human Brain Mapping*, 8:194–208, 1999.

[97] N. Levinson. The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction. *Journal of Mathematical Physics*, 25:261–278, 1947.

[98] V. Makarenko and R. Llinas. Experimentally Determined Chaotic Phase Synchronization in a Neural System. *Proceedings of the National Academy of Sciences of the United States of America*, 95:15747–15752, 1998.

[99] S. Makeig, T.-P. Jung, A.J. Bell, D. Ghahremani, and T.J. Sejnowski. Blind Separation of Auditory Event-related Brain Responses into Independent Components. *Proceedings of the National Academy of Sciences of the United States of America*, 94:10979–10984, 1997.

[100] J. Makhoul. Linear Prediction: A tutorial review. *Proceedings of the IEEE*, 63:561–580, 1975.

[101] S.G. Mason and G.E. Birch. A Brain-Controlled Switch for Asynchronous Control Applications. *IEEE Transactions on Biomedical Engineering*, 47(10):1297–1307, 2000.

[102] D.J. McFarland, L.M. McCane, S.V. David, and J.R. Wolpaw. Spatial filter selection for eeg-based communication. *Electroencephalography and Clinical Neurophysiology*, 103:386–394, 1997.

[103] D.J. McFarland, L.M. McCane, and J.R. Wolpaw. EEG-Based Communication and Control:Short-Term Role of Feedback. *IEEE Transactions on Rehabilitation Engineering*, 6(1):7–11, 1998.

[104] D.J. McFarland, W.A. Sarnacki, and J.R. Wolpaw. Brain-computer interface (BCI) operation: optimizing information transfer rates. *Biological Psychology*, 63:237–251, 2003.

[105] M.S. Middendorf, G.R. McMillan, G.L. Calhoun, and K.S. Jones. Brain-Computer Interfaces Based on the Steady-State Visual-Evoked Response. *IEEE Transactions on Rehabilitation Engineering*, 8:211–214, 2000.

[106] J.d.R. Millan. A Local Neural Classifier for the Recognition of EEG Patterns Associated to Mental Tasks. *IEEE Transactions on Neural Networks*, 13:678–686, 2002.

[107] J.d.R. Millan and J. Mourino. Asynchronous BCI and local neural classifiers: an overview of the adaptive brain interface project. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):159–161, 2003.

[108] R. Miller. *Cortico-hippocampal interplay and the representation of contexts of the brain*. Springer Verlag, 1991.

[109] W. Miltner, W. Larbig, and C. Braun. Biofeedback of somatosensory event related potentials: can individual pain sensations be modified by biofeedback-induced self-control of event-related potentials. *Pain*, 35:205213, 1988.

[110] J. Mller-Gerking, G. Pfurtscheller, and H. Flyvbjerg. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Electroencephalography and Clinical Neurophysiology*, 110:787–798, 1999.

[111] F. Mormann, K. Lehnertz, P. David, and C.E. Elger. Mean Phase Coherence as a Measure for Phase Synchronization and its Application to the EEG of Epilepsy Patients. *Physica D*, 144:358–369, 2000.

[112] K.-R. Müller, C.W. Anderson, and G.E. Birch. Linear and Nonlinear Methods for Brain-Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):165–169, 2003.

[113] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–202, 2001.

[114] J. Muthuswamy and N.V. Thakor. Spectral analysis methods for neurological signals. *Journal of Neuroscience Methods*, 83:1–14, 1998.

[115] E. Niedermeyer and F.H. Lopes da Silva. *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*. Williams and Wilkins, 4 edition, 1999.

[116] A.R. Nikolaev and A.P. Anokhin. Eeg frequency ranges during perception and mental rotation of two- and three-dimentional objects. *Neuroscience and Behavioral Physiology*, 6(28):670–677, 1998.

[117] P.L. Nunez, R.B. Silbersteina, Z. Shia, M.R. Carpentera, R. Srinivasana, D.M. Tuckerb, S.M. Doranc, P.J. Caduschd, and R.S. Wijesinghea. EEG coherency II: experimental comparisons of multiple measures. *Electroencephalography and Clinical Neurophysiology*, 110:469–486, 1999.

[118] P.L. Nunez, R. Srinivasan, A.F. Westdorpa, R.S. Wijesinghea, D.M. Tuckerb, R.B. Silbersteine, and P.J. Cadusche. EEG coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and Clinical Neurophysiology*, 103:499–515, 1997.

[119] H. Nyquist. Certain topics in telegraph transmission theory. *Transactions AIEE*, 47:617–644, 1928.

[120] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller. Information Transfer Rate in a Five-Classes BrainComputer Interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 9(3):283–288, 2001.

[121] D.A. Overton and C. Shagass. Distribution of eye movement and eye blink potentials over the scalp. *Electroencephalography and Clinical Neurophysiology*, 27:546, 1969.

[122] R.D. Pascual-Marqui, C.M. Michel, and D. Lehmann. Segmentation of brain electrical activity into microstates: model estimation and validation. *IEEE Transactions on Biomedical Engineering*, 42(7):658–665, 1995.

[123] W.D. Penny, S.J. Roberts, E.A. Curran, and M.J. Stokes. EEG-Based Communication: A Pattern Recognition Approach. *IEEE Transactions on Rehabilitation Engineering*, 8(2):214–215, 2000.

[124] J. Perelmouter and N. Birbaumer. A Binary Spelling Interface with Random Errors. *IEEE Transactions on Rehabilitation Engineering*, 8(2):227–232, 2000.

[125] G. Pfurtscheller and A. Aranibar. Event-related cortical desynchronization detected by power measurements of scalp EEG. *Electroencephalography and Clinical Neurophysiology*, 42:817–826, 1977.

[126] G. Pfurtscheller and C. Neuper. Motor Imagery and Direct Brain-Computer Communication. *Proceedings IEEE*, 89:1123–1134, 2001.

[127] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlgl, B. Obermaier, and M. Pregenzer. Current trends in graz brain-computer interface (BCI) research. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 8:216–219, 2000.

[128] G. Pfurtscheller, C. Neuper, G.R. Müller, B. Obermaier, G. Krausz, A. Schlögl, R. Scherer, B. Graimann, C. Keinrath, D. Skliris, M. Wörtz, G. Supp, and C. Schrank. Graz-bci: State of the Art and Clinical Applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):177–180, 2003.

[129] G. Pfurtscheller, C. Neuper, A. Schloegl, and K. Lugger. Separability of EEG Signals Recorded During Right and Left Motor Imagery Using Adaptive Autoregressive Parameters. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 6:316–325, 1998.

[130] D.A. Pierre. *Optimization Theory with Applications.* Dover Pubns, 1987.

[131] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Science.* Cambridge University Press, 2002.

[132] J.A. Pineda, B.Z. Allison, and A. Vankov. The Effects of Self-Movement, Observation, and Imagination on Rhythms and Readiness Potentials (RPs): Toward a BrainComputer Interface (BCI). *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 8(2):219–222, 2000.

[133] M. Pregenzer and G. Pfurtscheller. Frequency Component Selection for an EEG-Based Brain to Computer Interface. *IEEE Transactions on Rehabilitation Engineering*, 7(4):413–419, 1999.

[134] M.B. Priestley. *Spectral Analysis and Time Series.* Academic Press, 1981.

[135] J.G. Proakis and D. Manolakis. *Digital Signal Processing: Principles, Algorithms and Applications.* Prentice Hall, 1995.

[136] P.Stoica and R. Moses. *Introduction to Spectral Analysis.* Prentice Hall, 1988.

[137] P.Sykacek, S.Roberts, M.Stokes, E.Curran, M.Gibbs, and L.Pickup. Probabilistic Methods in BCI Research. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):192–195, June 2003.

[138] D. Ravden and J. Polich. On p300 measurement stability, habituation, intra-trial block variation, and ultadian rhythms. *Biological Psychology*, 51:5976, 1999.

[139] S. Risau-Gusman and M. B. Gordon. Generalization properties of finite-size polynomial support vector machines. *Physical Review*, 62:7092–7099, 2000.

[140] J. Rissanen. Modeling by Shortest Data Description. *Automatica*, 14:465–471, 1978.

[141] J. Robbins. *A Symphony in the Brain.* Atlantic Monthly Press, 2000.

[142] B. Rockstroh, N. Birbaumer T. Elbert, and W. Lutzenberger. Operant control of EEG and event-related and slow brain potentials. *Biofeedback and Self Regulation*, 9:139160, 1984.

[143] B. Roder, F. Rosler, E. Hennighausen, and F. Nacker. Event-related potentials during auditory and somatosensory discrimination in sighted and blind human subjects. *Cognitive Brain Research*, 4:77–93, 1996.

[144] V.S. Rotenberg and V.V. Arshavsky. Right and left brain hemispheres activation in the representatives of two different cultures. *Homeostasis in Health & Disease*, 38(2):49–57, 1997.

[145] N. Saiwaki, N. Kimura, and S. Nishida. An analysis of EEGS based on Information Flow with SD method. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 4115–4119, 1998.

[146] A. Schlögl and G. Pfurtscheller. Considerations on Adaptive Autoregressive Modelling in EEG Analysis. In *Proceedings of the First International Sysmposium on Communication Systems and Digital Signal Processing CSDSP*, volume 98, pages 367–370, 1998.

[147] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[148] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. *IEEE Transactions on Signal Processing*, 45:2758–2765, 1997.

[149] M.D. Serruya, N.G. Hatsopoulos, L. Panininski, M.R. Fellows, and J.P. Donoghue. Instant neural control of a movement signal. *Nature*, 416:141–142, 2002.

[150] C.E. Shannon. Communication in the presence of noise. *Proceedings Institute of Radio Engineers*, 37(1):10–21, 1949.

[151] C.E. Shannon and W. Weaver. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

[152] W. Singer. Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, 55:349–374, 1993.

[153] B.F. Skinner. *The Behavior of Organizms*. New York:Appleton, 1938.

[154] W. Sommer and S. Schweinberger. Operant conditioning of P300. *Biological Psychology*, 33:37–49, 1992.

[155] S. Sutton, M. Braren, J. Zubin, and E.R. John. Evoked correlates of stimulus uncertainty. *Science*, 150:1187–1188, 1965.

[156] C. Tallon-Boudry, O. Bertrand, F. Peronnet, and J. Pernier. Induced gamma-band activity during the delay of a visual short-term memory task in humans. *Journal of Neuroscience*, (11):4244–4254, 1998.

[157] D.M.J. Tax. *One-class classification*. PhD thesis, Technische Universiteit Delft, 2001.

[158] J.J. Tecce, J. Gips, C.P. Olivieri, L.J. Pok, and M.R. Consiglio. Eye movement control of computer functions. *International Journal of Psychophysiology*, 29:319–325, 1998.

[159] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, pages 520–522, 1996.

[160] C. Toro, G. Deuschl, R. Thatcher, S. Sato, C. Kufta, and M. Hallett. Event-related desynchronization and movement-related cortical potentials on the ECoG and EEG. *Electroencephalography and Clinical Neurophysiology*, 93:380–389, 1994.

[161] L.J. Trejo, K.R. Wheeler, C.C. Jorgensen, R. Rosipal, S.T. Clanton, B. Matthews, A.D. Hibbs, R. Matthews, and M. Krupka. Multimodal Neuroelectric Interface Development. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11:199–204, 2003.

[162] D.M. Tumey, P.E. Morton, D.F. Ingle, C.W. Downey, and J.H. Schnurer. Neural Network Classification of EEG using Chaotic Preprocessing and Phase Space Reconstruction. In *Proceedings of the Seventeenth IEEE Annual Northeast Bioengineering Conference*, 1991.

[163] W.R. Utall. *The War Between Mentalism and Behaviorism: On the Accessibility of Mental Processes*. NJ:Erlbaum, 1999.

[164] M. van de Velde, G. van Erp, and P. J. M. Cluitmans. Detection of muscle artefact in the normal human awake EEG. *Electroencephalography and Clinical Neurophysiology*, 107(2):149–158, April 1998.

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

[166] V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[167] T.M. Vaughan, W.J. Heetderks, L.J. Trejo, and W.Z. Rymer. Brain-Computer Interface Technology: A Review of the Second International Meeting. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):94–109, June 2003.

[168] P. Venkataraman. *Applied Optimization with MATLAB Programming*. John Wiley and Sons, 2002.

[169] G. Wahba. *Advances in Kernel Methods-Support Vector Learning*, chapter Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, pages 69–88. MIT Press, 1999.

[170] G. Walker. On periodicity in Series of Related Terms. *Proceedings of the Royal Society (London) A*, 131:518–532, 1931.

[171] W.G. Walter, R. Cooper, V.J. Aldridge, W.C. McCallum, and A.L. Winter. Contingent negative variation: an electric sign of sensorimotor association and expectancy in the human brain. *Nature*, 203:380–384, 1964.

[172] P.D. Welch. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Transactions on Audio Electroacoustics*, AU-15:70–73, 1967.

[173] P. Whittle. On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50:129–134, 1963.

[174] U. Windhorst and H. Johansson. *Modern Techniques in Neuroscience Research*. Springer Verlag, 1999.

[175] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113:767–791, 2002.

[176] J.R. Wolpaw, D.J. McFarland, and T.M. Vaughan. Brain-computer interface research at the Wadsworth Center. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 8:222–226, 2000.

[177] J.R. Wolpaw, H. Ramoser, D.J. McFarland, and G. Pfurtscheller. Eeg-based communication: improved accuracy by response verification. *IEEE Transactions on Rehabilitation Engineering*, 6(3):326–333, 1998.

[178] J. Wright, R. Kydd, and A Sergejev. Autoregressive models of EEG. *Biological Cybernetics*, 62:201–210, 1990.

[179] A.R. Wyler and K.J. Burchiel. Factors influencing accuracy of operant conditioning of tract neurons in monkey. *Brain Research*, (152):418–421, 1978.

[180] A.R. Wyler, K.J. Burchiel, and S.A. Robbins. Operant control of precentral neurons in monkeys: evidence against open loop control. *Brain Research*, (171):29–39, 1979.

[181] G. Xiaorong, X. Dingfeng, M. Cheng, and G. Shangkai. A BCI-based environmental controller for the motion-disabled. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):137–140, 2003.

[182] G.U. Yule. On a Method for Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society of London Series A-Mathematical and Physical Sciences*, 226:267–298, 1927.