Pattern Recognition in Electric Brain Signals

- mind reading in the sleeping brain

Christian Vad Karsten

Kongens Lyngby 2012
IMM-M.Sc.-2012-91

Technical University of Denmark
DTU Informatics
Asmussens Alle, building 305, DK-2800 Lyngby, Denmark
Phone +45 4525 3351, Fax +45 4588 2673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-M.Sc.-2012-91

Abstract

A machine learning framework for analyzing experimental EEG data is presented. The question of whether the human brain is capable of more abstract processing during sleep is partly answered by analyzing data from 18 sleeping subjects tested at a semantic level using two different classes of auditory input. Using a pattern recognition algorithm it is possible to localize significant discriminating activity in 12 subjects during sleep. To validate the method, it is applied to data from the same experiment obtained during wakefulness. Here it produces significant results for 16 subjects.

The purpose of the presented pattern-based analysis is twofold. The first objective is to consider whether classification is possible with the underlying presumption that if a classifier can label new examples with a better accuracy than chance, then the two conditions are indeed differently represented in the brain. The second is to make claims about information representation in the brain.

Both objectives are fulfilled. Regardless of differences in latency and morphology at a single-subject level, patterns similar to results from the relevant literature concerning wakefulness do arise. This can be an indication of cognitive processing during sleep all the way up to motor planning.

The presented results are obtained using a novel method for image based analysis of EEG spectrograms at the sensor level. A non-linear support vector machine is trained directly on spectrograms and combined with an embedded feature selection scheme to overcome the challenges posed by low sample size high dimensional data. Opposite to classical analysis of EEG data, this method allows analysis at an individual subject level, where results are normally obtained at a group level. In addition to answering the question of whether there is information of interest (pattern discrimination), the method also to some degree answer the questions of where and how the information is encoded (pattern localization and pattern characterization).

Resumé

I hvor høj grad hjernen fortsætter med at behandle eksterne input, mens mennesket sover, har været diskuteret gennem årtier. I det følgende vises, at hjernen i en vis grad forsætter med at behandle auditoriske input på et semantisk niveau i de lette søvnfaser.
Der anvendes EEG-optagelser fra et forsøg, hvor testpersoner skal skelne mellem ordklasser i vågen tilstand og i forlængelse deraf i sovende tilstand. Til analysen udvikles en selvlærende algoritme, som finder signifikant diskriminerende hjerneaktivitet i mindst 12 ud af 18 sovende testpersoner. Metoden valideres på data fra den vågne tilstand af forsøget. Her er det muligt at vise diskriminerende aktivitet i 16 ud af de 18 testpersoner.
Den foreslåede metode til analyse af EEG-optagelser fra et eksperimentelt setup, benytter en ikke-lineær support vektor maskine med en inkorporeret automatisk udvælgelse af relevante datapunkter. Det giver mulighed for både at vise diskriminerende aktivitet i hjernen, samt at undersøge hvor og hvordan informationen er indkodet. Resultaterne sammenlignes med litteraturen på området, og der findes generelle ligheder. Mellem forsøgspersonerne ses betydelige forskelle i latens og morfologi ved aktivering af hjernen. Resultaterne tyder på, at hjernen behandler information helt op til forberedelse af bevægelse.
Metoden udmærker sig ved, at det i modsætning til klassisk analyse af EEG-data er muligt at analysere data på individniveau. Ligeledes giver metoden mulighed for at analysere direkte på spektrogrammer fra enkelte elektroder på trods af de højdimensionelle repræsentationer med få forsøgseksempler.
Metoden er generisk i den forstand, at relevante spatiotemporale strukturer i hjernen udvælges automatisk på baggrund af relevans for klassifikationen. Det gør den særligt anvendelig i kendte såvel som ukendte eksperimentelle paradigmer, hvor forhåndsantagelser kan være problematiske.

Preface

This thesis was written at the Section for Cognitive Systems, Department of Informatics and Mathematical Modelling, Technical University of Denmark as a partial fulfillment of the requirements for acquiring the M.Sc. in Mathematical Modelling and Computation.
This thesis was carried out between March 1st and August 15th 2012 with a workload of 30 ECTS credits.

I would like to thank my three supervisors, Professor, Lars Kai Hansen, and Post. Doc., Carsten Stahlhut, both Department of Informatics and Mathematical Modeling, DTU and Dr. Sid Kouider, École Normale Supérieure, Paris, for their genuine interest in my project and for asking all the questions.
Also I would like to acknowledge Leonardo Barbosa, École Normale Supérieure, for having made the pre-processing of data used in this thesis a lot easier.

Kgs. Lyngby, 15-August-2012

Christian Vad Karsten

Contents

Abstract i

Resumé iii

Preface v

List of Abbreviations xi

1 Introduction 1

2 Physiological Background 5
2.1 Brain Activity and EEG . . . 5
2.1.1 Classification Studies . . . 10
2.2 The Sleeping Brain . . . 10
2.2.1 Processing During Sleep . . . 11
2.3 Experimental Setup . . . 15
2.3.1 Procedure . . . 15
2.3.2 Stimuli . . . 15
2.3.3 Sleep Assessment . . . 17
2.3.4 Subjects . . . 17
2.3.5 EEG Equipment . . . 18

3 Support Vector Machines 19
3.1 The Learning Setting . . . 20
3.2 Separating Hyperplanes . . . 20
3.3 Optimal Margin Hyperplanes . . . 22
3.4 Soft Margin Optimal Hyperplanes . . . 24
3.5 The Non-linear Support Vector Machine for Non-separable Data 25
3.5.1 Kernel . . . 28

3.6 Numerical Optimization . . . 28
3.6.1 Subset selection . . . 29
3.6.2 Sequential Minimal Optimization . . . 29
3.6.3 Stopping Criterion . . . 31
3.6.4 Implementation . . . 33

4 Feature Extraction **35**
4.1 Feature Construction . . . 35
4.1.1 Spectral Decomposition . . . 36
4.2 Feature Selection . . . 37
4.2.1 NIPS 2003 Feature Selection Challenge Summary . . . 39
4.2.2 Univariate Methods . . . 40
4.2.3 Multivariate Embedded Methods . . . 41

5 Model Selection **45**
5.1 The Curse of Dimensionality . . . 45
5.2 Cross-validation . . . 46
5.3 Permutation Test . . . 47
5.4 Synthetic Data . . . 48
5.5 Parameter Selection . . . 49
5.5.1 SVM Parameters . . . 49
5.5.2 SMO Parameter . . . 52
5.5.3 RFE Parameters . . . 59
5.6 Evaluation of Model Performance . . . 59

6 Results **61**
6.1 Data Pre-processing . . . 61
6.1.1 Data Set Dimensionality Reduction . . . 63
6.1.2 Artifact Rejection . . . 63
6.2 Feature Selection . . . 64
6.3 T-Test Based Feature Selection . . . 64
6.4 RFE Feature Selection . . . 64
6.4.1 Permutation Test . . . 66
6.4.2 Data From Pre-motor Cortex . . . 71
6.4.3 Data From Visual Cortex . . . 71
6.4.4 Data From the Awake Condition . . . 72
6.4.5 Spectro-histo-grams . . . 73
6.4.6 Inter-Subject Learning & Group Level Results . . . 75
6.4.7 Learning Curves . . . 75
6.4.8 Included Electrodes . . . 75
6.4.9 Kernel Parameter . . . 78
6.4.10 Number of Cross-validation Iterations . . . 78

7 Discussion **79**
7.1 Spatiotemporal Cortical Dynamics 79
7.2 Anti-Learning 81
7.3 Data Representation 82
7.4 Computational Issues 83
7.5 Future Work - Harnessing the Machine Learning Approach . . . 84

8 Conclusion **87**

A Classifier Performance and Spectro-histo-grams **89**

B Learning Curves **105**

C Awake Classifier Performance **113**

Bibliography **117**

List of Abbreviations

BCI	Brain Computer Interface
EEG	Electroencephalography
ERD	Event-Related Desynchronization
ERP	Event-Related Potentials
ERS	Event-Related Synchronization
fMRI	Functional Magnetic Resonance Imaging
ICA	Independent Component Analysis
KKT	Karush-Kuhn-Tucker
LDA	Linear Discriminant Analysis
LRP	Lateralized Readiness Potential
PCA	Principal Component Analysis
QP	Quadratic Programming
RBF	Radial Basis Function
REM	Rapid Eye Movement
RFE	Recursive Feature Elimination
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
XOR	Exclusive OR

Chapter 1

Introduction

Sleep is a recurring and readily reversible state of unconsciousness in every human being. It is revealed by inactivity of most voluntary muscles and apparent unresponsiveness to, and interaction with, external stimuli. To what extend the brain actually shuts down and whether semantic processes take place during sleep however, is relatively unknown. This thesis deals with the question of whether, and to what extent, the brain continues to respond to and process external stimuli during sleep.

Though [Emmons and Simon, 1956] state that material presented a number of times during sleep cannot be subsequently recalled and in contrast [Huxley, 1932]'s conditioning of children in "Brave New World" via hypnopedia seems far fetched, both some older [Oswald et al., 1960, Formby, 1967] and newer [Halperin and Iorio, 1981, Bastuji et al., 2002] studies have shown that auditory stimuli with a relevant meaning are more likely to lead to awakening, indicating some discriminating brain activity. However, it is still unclear if the brain is capable of more abstract processing or even preparation of task-relevant responses.
To get an insight on this issue, data from an experiment, where subjects were presented a task-set while awake and tested later whether this task-set would be maintained while the same subjects were asleep, is used. Subjects were presented with auditory stimuli before sleep onset and had to give a behavioral response, classifying the stimulus as animals vs. objects by pressing a button using the right and left hand respectively. This is known to induce contralateral

activity patterns in the brain, [Pfurtscheller and Lopes da Silva, 1999]. Hence, this task allowed the mapping of two specific categories with a specific manual response. It was reasoned that the induction of a category-response mapping just before sleep onset would promote the maintenance of this task-set after the disappearance of behavioral responsiveness. Following sleep onset, subjects were exposed to new auditory stimuli within the same two categories to ensure the involvement of semantic categorization rather than simple stimulus-response associations.

Sleeping subjects present several significant challenges since classical analysis of motor action is not possible. A lot of cognitive experiments are a combination of a presented stimulus and a simple forced-choice perceptual report. This is the case in the awake case of the present experiment. During sleep there is no direct way to evaluate the response. Both during sleep and wakefulness however, it is possible to measure and evaluate brain activity using Electroencephalography, EEG.
Usually grand averages of EEG-signals within a group of subjects might give some insights using event-related potentials, ERP, and event-related desynchronization, ERD and synchronization, ERS. However, in EEG signals from sleeping subjects the signal-to-noise ratio is reduced and simple statistical analyses is more challenging, also due to a considerable intra- and inter- subject variability. Furthermore, when dealing with high dimensional data representations in neuroscience studies, classical statistics are insufficient for analysis of single subjects. Simplified analysis where the dimension is reduced to a single measurement might be unsatisfactory and hence, the analysis can often only be conducted at a group level.
More advanced mathematical tools may provide additional insights. One approach to high dimensional problems is to use pattern recognition systems. Pattern recognition is about seeing similarities in diverse data and machine learning techniques, especially the Support Vector Machine, SVM [Boser et al., 1992, Cortes and Vapnik, 1995, Vapnik, 1998], has proven very powerful for various classification- and pattern recognition problems, including EEG data interpretation in Brain Computer Interfaces, BCI, [Lotte et al., 2007]. Hence they may help to answer the question of the degree of semantic processing continuing in the brain during sleep. Utilizing the structure of the models might even help to identify specific cortical dynamics associated with the described tasks. Using SVMs it is possible to do the analysis at the level of individual subjects.

The adaption of support vector machines for neuro-scientific image-based statistical analysis is inspired by several functional magnetic resonance imaging, fMRI studies [Pereira et al., 2009]. Two or more sets of images, in the present thesis EEG spectrograms corresponding to two different conditions, are analyzed with the goal of identifying differences. This is done by training an SVM on one part

of the data set and then predict the labels on the rest of the data set. The underlying presumption is that if an SVM can label new examples with a better accuracy than chance, the two conditions are indeed different, and the SVM implicitly find the differences. Furthermore, the benefit of the SVM is that it does not assume independence between the data set features.

The performance of pattern recognition systems obviously depends on the input features and the applied classification algorithm. Furthermore, tuning of parameters in the different algorithms is of great importance. SVMs are usually considered to have two user defined hyperparameters. The kernel parameter, σ, to control the degree of nonlinearity applied to the feature space and a regularization term, C, which determines the trade off between minimizing the training error and minimizing model complexity. However, there is another parameter in many of these systems, which is usually less investigated. This parameter is the stopping criterion, ϵ, employed in the different optimization algorithms. It could potentially impact the results of the SVM in terms of accuracy and speed. This parameter is explored in this thesis.

SVMs normally use regularization to avoid overfitting without requiring space dimensionality reduction. However e.g.[Guyon et al., 2002] show that SVMs can benefit from space dimensionality reduction, not only concerning computational tractability, but also regarding prediction performance.
One solution to space dimensionality reduction is feature selection. In addition to performance improvements, it is an intuitive way to get a better understanding of the influence of the input data, i.e. identifying the characteristics of the cortical dynamics needed for classification.
Feature selection can be done using the weights of a linear SVM itself, by sensitivity analysis, correlation coefficients, ranking criterions etc, [Guyon et al., 2006]. Recursive feature elimination, [Guyon et al., 2002], is a technique that removes features iteratively using an internal measure. It is reported to obtain a better ranking of features than by using the weights of a single classifier. In this thesis the recursive feature elimination is expanded to comply with a non-linear version of the SVM to identify and utilize possible non-linear patterns as well.

Hence to summarize, the goal of the current thesis is to harness machine learning techniques to analyze EEG recordings obtained during sleep at a level not possible using classical statistical analysis. The two main objectives are:
Objective 1: Classify (SVM) and locate (Feature Selection) patterns in sleeping brains which can indicate discriminating activity at a semantic level during sleep.
Objective 2: Improve (Feature Selection and Feature Construction) and investigate computational tractability (Feature Selection and Optimization Tolerance) of the classification model.

The thesis is organized as follows. The physiological background and the experiment used for the data acquisition are presented in Chapter 2. The SVM and the algorithm for solving it will be derived in Chapter 3. The SVM is enhanced for the current setting by feature selection schemes introduced in Chapter 4. Considerations related to parameter tuning and model selection are presented in Chapter 5 along with a resampling framework for evaluation of the methods. Small examples are provided to illustrate the impact of the proposed heuristics. Finally the data obtained from the experiment are analyzed in Chapter 6 using the presented framework. In Chapter 7 the results are discussed and extensions are proposed before a final conclusion is provided in Chapter 8.

Chapter 2

Physiological Background

2.1 Brain Activity and EEG

A fundamental property of the brain is the ability of groups of neurons to work in synchrony and generate oscillatory activity, [Bear et al., 2007]. Brain activity and inactivity is widely studied and large-scale activity can be measured by non-invasive techniques such as EEG, which is the recording of electrical activity along the scalp, see Figure 2.1 and 2.2. In general, EEG signals have a good time resolution and a broad spectral content. The signals are usually described using the oscillatory activity in specific frequency bands, and it appears that the frequency of brain oscillations is negatively correlated with amplitude. The amplitude is furthermore proportional to the number of synchronously active neurons. This indicates that slowly oscillating cell assemblies comprise more neurons than fast oscillating assemblies [Brown and Singer, 1993]. These fluctuations of field potentials from large cell assemblies are measured by electrodes that are fixed to the outside of the scalp. Single electrodes can easily be mounted and removed, but since the EEG signal is usually recorded at many locations simultaneously, the use of EEG caps is an advantage. In such a setup, the distance between neighboring electrodes is usually in the range of one to a few centimeters.
The spatial resolution of EEG is not only limited by the distance between electrodes, [Nunez and Srinivasan, 2006]. EEG recordings suffer from the fact that

the electric signal has to pass through the intra-cerebral liquor, the meninges, the skull bone, and the skin, see Figure 2.2. These layers act as a low-pass filter to the electrical fields and act as a spatial low-pass filter as well. EEG is furthermore very prone to so-called artifacts. Artifacts are signal components picked up by EEG electrodes that are not caused by neural activity and can be so strong in amplitude that interesting signals are not detectable any more. The fact that artifacts are picked up with highest intensity at electrodes closest to their origin can help to identify them. Typical artifacts in EEG comprise: muscle activity, movements of the eyeball, eye blinks and exterior signal sources. Most artifacts however, can be controlled by proper instruction of the subjects by using additional control electrodes close to possible artifact locations and by proper frequency filtering of the recorded signals, [Nunez and Srinivasan, 2006].
[Berger, 1929] discovered the oscillatory behavior and described the commonly occurring alpha activity (8 -13 Hz) after his invention of EEG around 1930. It can be detected from the occipital lobe during relaxed wakefulness and increases when the eyes are closed. Later defined frequency bands are delta (1 - 4 Hz), theta (4 - 8 Hz), beta (13 - 30 Hz) and gamma ($>$ 30 Hz). Generally oscillations in the beta band and above indicate an activated cortex [Bear et al., 2007]. The exact definitions of the frequency ranges are varying slightly in the literature, especially the transition from beta to gamma is defined over a broader range, see e.g. [Nunez and Srinivasan, 2006, Pfurtscheller and Lopes da Silva, 1999].
It has been shown that several kinds of events can induce time-locked changes in the oscillatory activity of ensembles of neurons or neural networks. These changes are commonly referred to as event-related potentials, ERP. Averaging techniques are commonly used to detect ERPs since these will enhance the signal-to-noise ratio. The underlying assumption is that the brain signal has a relatively fixed time-delay to the stimulus and other ongoing activity behaves as additive noise. This is however just a simplification of the real condition. Certain events are indeed time-locked but not phase-locked and can either desynchronize or even block the present alpha activity. These types of changes can not be extracted by simple averaging methods, but may be detected by frequency analysis [Pfurtscheller and Lopes da Silva, 1999]. These kind of phenomena where power in a given frequency band is either increased or decreased may be viewed as an increase or decrease in synchrony of the underlying neural network. It is referred to as event-related synchronization, ERS, [Pfurtscheller, 1992] and event-related desynchronization, ERD, [Pfurtscheller, 1977].
[Pfurtscheller and Lopes da Silva, 1999] propose that ERPs reflect changes in afferent activity in the cortical neurons and ERD/ERS reflect changes in the activity of local interactions between main neurons and interneurons.
[Pfurtscheller and Lopes da Silva, 1999] and [Donner et al., 2009] summarizes from several studies how rhythmic neural activity carries information about sensory stimuli, cognitive processes, or motor tasks. Limb movements especially are reported to be accompanied by suppression of low-frequency activity and enhancement of high frequency activity in motor cortex, which is the part of

the brain responsible for planning and execution of movements. It is noteworthy that both suppression and enhancement of activity is stronger contralateral than ipsilateral to the movement. Furthermore [Donner et al., 2009] finds in accordance with other studies that neural activity exhibit robust lateralized effector selectivity of opposite polarity in beta and gamma bands.
Figure 2.3 shows the areas of the brain responsible for planning and execution of movements. Two main areas, area 4 just anterior to the central sulcus and area 6 which lies anterior to area 4, constitute the areas responsible for planning and execution of movements. All the highlighted areas in Figure 2.3 are however involved to some extend in voluntary limb movement. It has been shown that stimulation of area 4 leads to movements of the muscles on the contralateral side of the body and EEG recordings from this area show clear patterns when limb movements are present. Likewise, as described in the classification studies in the following section, area 6 is active when movements are imagined or planned but not executed whereas area 4 is dominant during the actual movement [Bear et al., 2007].

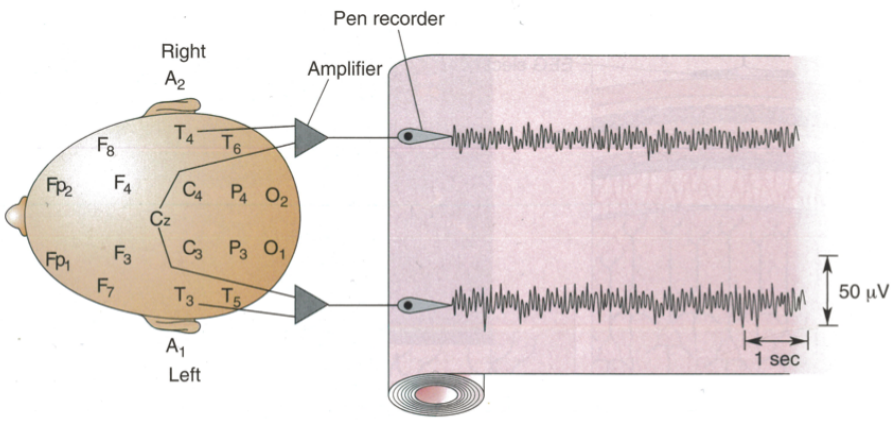


Figure 2.1: Principle of EEG recordings. Small voltage fluctuations are measured between selected pairs of electrodes placed on the scalp. Different areas of the brain, e.g. anterior, posterior, left or right, can be compared by selecting corresponding electrodes. The output from the amplifier either drives a pen recorder or is recorded digitally. (Source: [Bear et al., 2007])

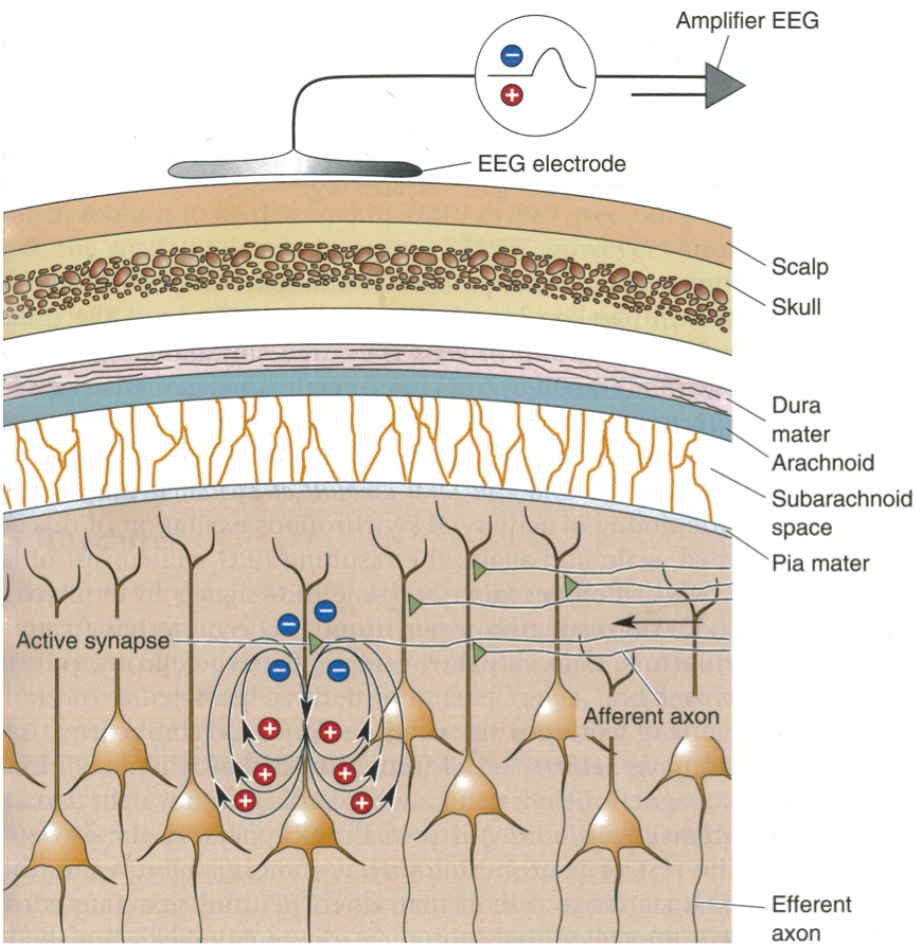


Figure 2.2: A small collection of pyramidal cells is seen. This is the prevalent neuron type of the cerebral cortex. The human nervous system consists of approximately 10^{10} neurons, where most of the neurons are situated in the central nervous system consisting of the brain and the spinal chord.
A neuron receives information either from sensory cells or from other neurons in the form of electrical or chemical stimulations which can be excitatory or inhibitory. If excitatory stimulations prevail, an inflow of Na+ ions through the membrane occurs. This inflow transiently disturbs the resting cell potential, depolarizes the membrane, and leads to a so-called excitatory postsynaptic potential. This depolarization only lasts for 1-2 ms before the influx of K+ ions reestablishes the original polarization.
The electrical contribution from a single neuron is extremely small and it must pass through several layers of non-neural tissue as seen in the figure to reach the electrode. Only if thousands of cells contribute to the signal, it is large enough to be measured by the electrode. The amplitude of the signal depends on how synchronous the underlying activity is. By convention EEG signals are plotted with negativity upward. (Source: [Bear et al., 2007])

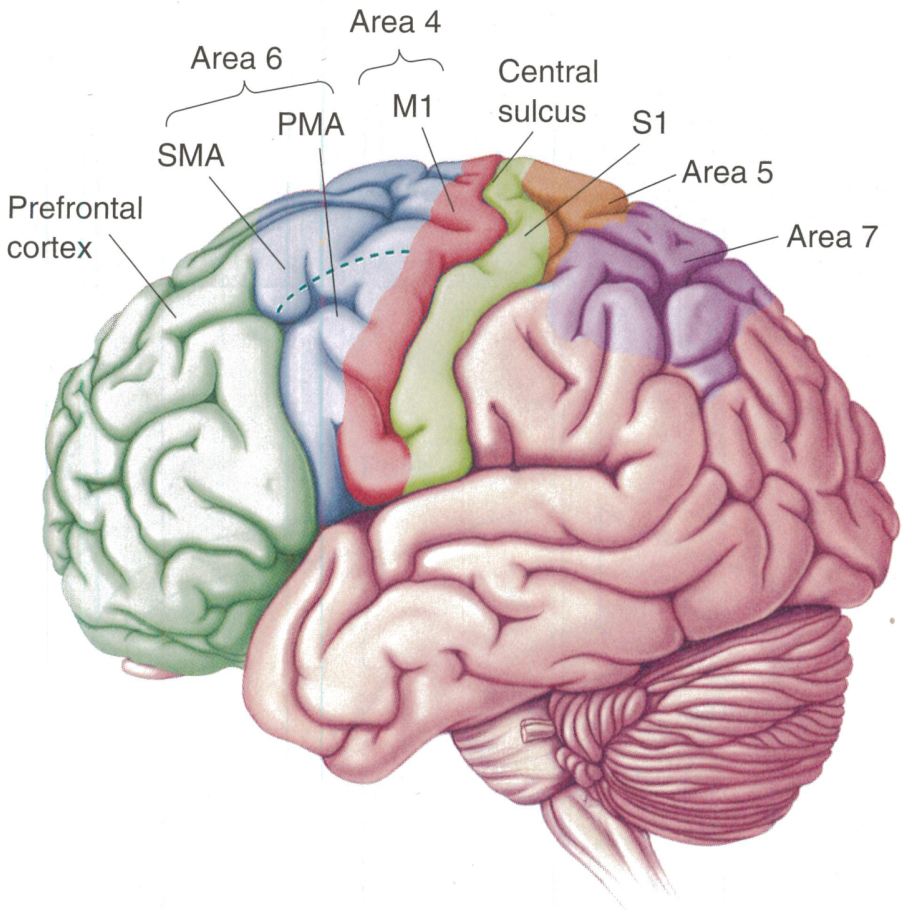


Figure 2.3: The areas of the cerebral cortex related to planning and execution of voluntary movements. Area 4 is the primary motor cortex and area 6 constitutes the premotor cortex involved in planning of movements. (Source: [Bear et al., 2007])

2.1.1 Classification Studies

[Pfurtscheller et al., 2006] present a classification study of four motor imagery task and concludes that the discrimination improved when ERD end ERS patterns were induced in at least one or two tasks. The most important electrode positions for the classification are found to be C3, C4, and Cz of the international 10-20 system, see Figure 2.5. Furthermore an optimal spatial filtering reveals electrodes in the neighborhood of C3 and C4 to be the most important. They finally conclude there is a great inter- and intra-subject variability concerning the reactivity of upper mu rhythm (9-13 Hz), which is the typical rhytmic activity exhibited by motor cotical areas at rest.
[Morash et al., 2008] use neural signals preceding movement and motor imagery to predict which of the four movements/motor imageries is about to occur, and to access this utility for BCI applications [Crone et al., 1998].
Within BCI applications machine learning techniques are widely used in connection with EEG recordings,[Blankertz et al., 2004, Lotte et al., 2007, Lal et al., 2004]. [Lotte et al., 2007] review classification algorithms used in a BCI setting and find that especially the support vector machine performs well. Using similar techniques to analyze experimental data in an EEG framework is less widespread whereas they are commonly used to analyze neuroimaging data in fMRI settings [Pereira et al., 2009, Haynes and Rees, 2006, Norman et al., 2006].
[Cruse et al., 2012] and [Cruse et al., 2011] investigate motor imagery tasks in a group of patients in the minimally conscious state and vegetative state using a linear SVM and find robust responses in some cases. They use artifact rejected, downsampled EEG signals recorded over the motor cortex and calculate log power values at every time step in four frequency bands ranging from 7-30 Hz. 60 to 203 trials contribute to each subjects single-trial analysis and accuracies in the range 38-78 % are obtained for the two-class analysis.

2.2 The Sleeping Brain

EEG signals change dramatically during sleep and show a transition from faster to increasingly slower frequencies. The spectral content is therefore one of the measures used to characterize different sleep stages.
Sleep has several very distinct phases but can overall be characterized by two main stages, the the Rapid Eye Movement, REM, and non-REM sleep discovered in the 50's by [Aserinsky and Kleitman, 1953]. The non-REM is characterized by high voltage and slow synchronized EEG rhythms whereas the REM sleep is characterized by desynchronized, fast and low voltage signals. When falling

asleep the EEG alpha rhythms of relaxed wakefulness become less regular and decline along with the eyes making slow, rolling movements. This is the first of four stages in the non-REM sleep [Bear et al., 2007], see Figure 2.4, and is also referred to as the drowsiness period. The second stage (light sleep) usually enters after a few minutes and lasts 5-15 minutes. This stage is slightly deeper and usually considered to be the actual onset of sleep. It is characterized by occasional sleep spindles and K-complexes. Sleep spindles are longer lasting oscillatory brain activity in the 8-14 Hz domain whereas the K-complex is a brief high-amplitude sharp wave, see Figure 2.4. The K-complex can occur spontaneously but also in response to e.g. auditory stimuli [Roth et al., 1956] and is often followed by spindles. Around actual sleep onset and before K-complexes and spindles occur, vertex sharp waves can be observed. They are small spike-like positive discharges that occur spontaneously or in response to sensory stimuli, [Rodenbeck et al., 2006]. Stage 3 (deep sleep) show large amplitude slow delta waves and sleep spindles gradually disappear as sleep becomes deeper. Stage 4 (very deep sleep) is the deepest of the four stages with large delta waves of 2 Hz or less. After stage 4 sleep lightens again and ascends to stage 2 from where it enters a brief period of REM sleep with fast beta rhythms. Physically, the REM sleep is characterized by rapid eye movements, rapid and irregular heart rate and breathing, increased blood pressure and the muscles of the body are virtually paralyzed. During a night the brain cycles through the different stages and generally moves towards more REM sleep as the night progress, [Bear et al., 2007]. Newer publications, e.g. [Iber, 2007], revise this classical view slightly with new class definitions and introduction of micro awakenings and deepening of sleep within sleep stages.

2.2.1 Processing During Sleep

During sleep external stimuli are processed to some extent. The fact that people are more easily awoken by presentation of their own name and mother's by their baby's cry is a clear indication that relevant external stimuli do get some attention, [Hennevin et al., 2007, Oswald et al., 1960, Formby, 1967, Burton et al., 1988, Bruck et al., 2009]. Another indication of the processing of external stimuli is observed from the phenomenon that external stimuli can be incorporated in dreams [Kramer et al., 1982].
[Edeline et al., 2000] show for the guinea pig that auditory messages sent by thalamic cells to cortical neurons are reduced but preserved both in terms of rate and frequencies, which indicate that the messages sent to cortical cells are not deprived of relevant information and can explain how processing of relevant stimuli is possible during sleep. There is hence also evidence of maintained cortical responsiveness to auditory stimuli during REM and non-REM sleep in

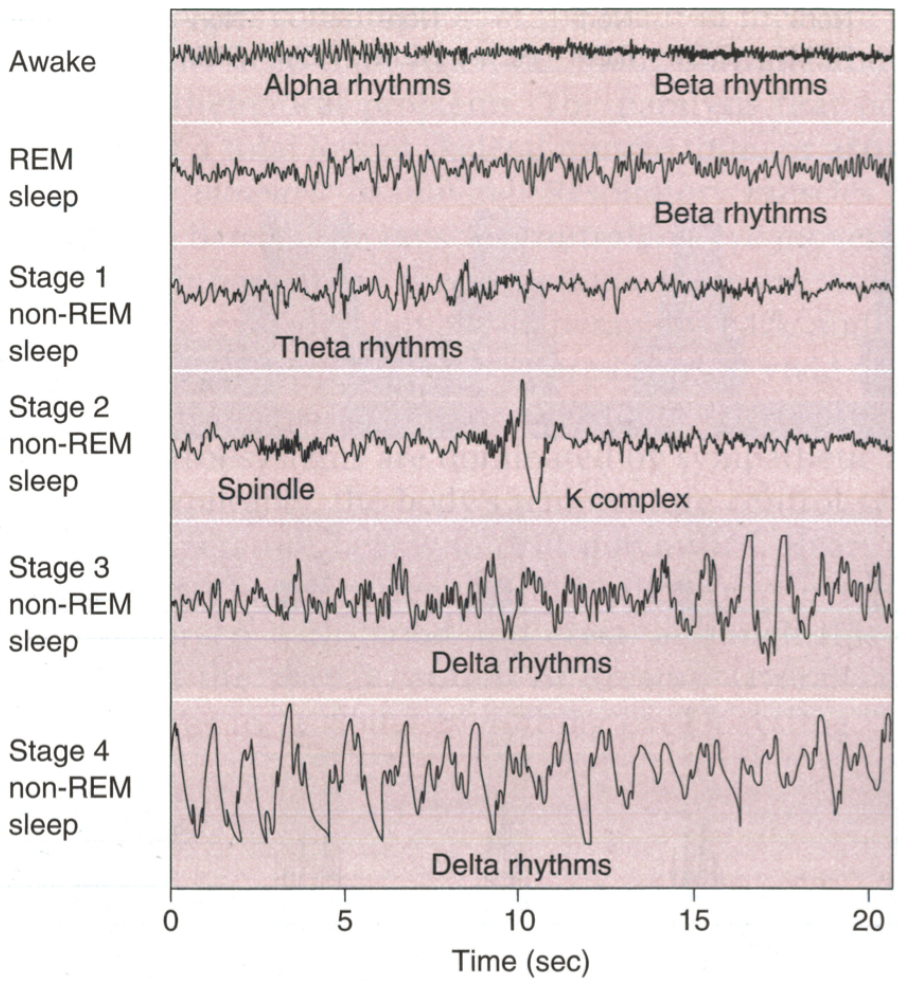


Figure 2.4: Typical EEG rhythms recorded during wakefulness and during sleep. The different signals illustrate the signals that characterize different sleep stages. (Source: [Bear et al., 2007])

both humans and animals, [Hennevin et al., 2007].
Human ERP studies of response to external auditory stimuli, such as one's own name, [Atienza et al., 2001], indicate that auditory information processing is possible though it is affected differently during the different stages of sleep. The P300 effect is an ERP component showing a positive deflection (relative to reference electrode) in voltage with a latency of 250 to 500 ms, which during wakefulness is evoked in the process of decision making. During sleep, studies of the P300 component indicate that discriminating processes occur though shape, latency and amplitude compared to the normal P300 component is different [Hennevin et al., 2007, Atienza et al., 2001, Perrin, 2004, Perrin et al., 1999], at least during early sleep stages and REM sleep. Compelling are also ERP studies which show that the N400 effect appears from word associations during both REM and early stages of non-REM sleep [Brualla et al., 1998, Ibáñez et al., 2008, Bastuji et al., 2002]. The N400 is a negative potential (relative to reference electrode) seen in the EEG which peaks around 400 ms post-stimulus in response to a wide array of meaningful or potentially meaningful stimuli such as auditory words. However, it is a rather automatic mechanism [Federmeier and Kutas, 2009, Kutas and Federmeier, 2011], but manipulations that affect the extent to which attention is allocated to N400-eliciting stimuli do influence the size of N400 effects. Whether processes indexed by the N400 require awareness has been debated for decades, but experiments suggest that N400 effects can be obtained even when manipulations are incidental to the task and when the stimuli themselves elicit little conscious awareness, e.g. during sleep.
At another level, [Antony et al., 2012] recently showed that a partly auditory task learned during wakefulness can be promoted during sleep, which is another indication that more active processing takes place.
Hence there are various indications that during sleep, auditory stimuli are integrated at a semantic level, but clearly further evidence and investigation of the level of cognitive processing is needed. Whether sleep involves deeper processing not only at a semantic level but all the way "up" to the preparation of task-relevant responses remains unclear. According to [Nofzinger et al., 2002, Maquet et al., 2000, Manganotti et al., 2004] the relative cortical activity in sleep and awakening of motor cortex indicates that motor cortex is not fully deactivated during sleep. However, it does remain a question whether cortical motor processes are active during sleep. And it is indeed a question posing certain challenges to investigate since the initialization of a new task-set might prove difficult during sleep stages, because the prefrontal regions dealing with executive functions are particularly suppressed in comparison to other cortical regions [Muzur et al., 2002, Maquet et al., 2000]. This is somewhat addressed in the experiment described in Section 2.3, which is the background of this thesis. Here an induction strategy is used as an approach for the study of non-conscious perception. The results presented in the following are obtained during the early stages of non-REM sleep and hence it remains to investigate whether it generalize to other sleep stages including REM sleep.

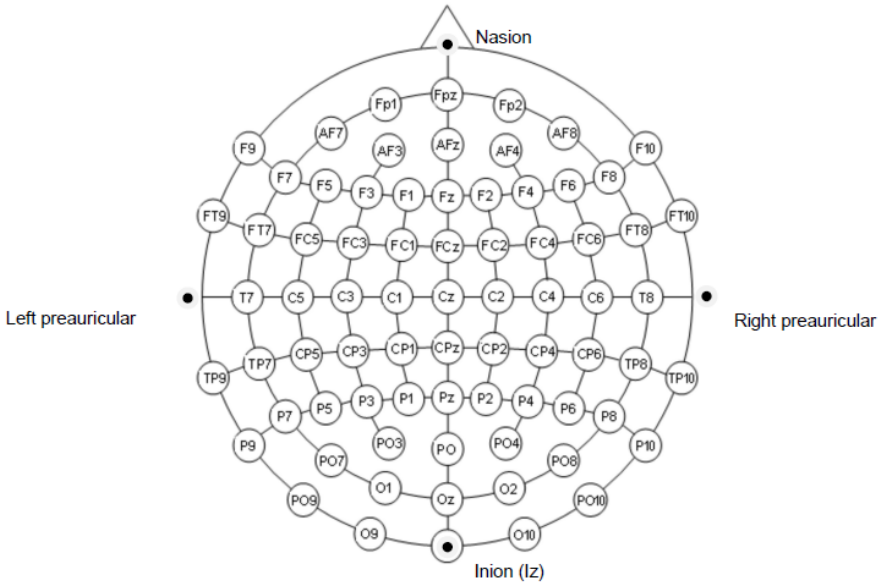


Figure 2.5: Standard placement of the electrodes according to the international 10-20 system. Each location is paired with a letter and a number to identify the lobe and hemisphere location respectively. F, T, P and O corresponds to frontal, temporal, parietal, and occipital lobe, and additionally C (central lobe) is introduced for further identification. Even numbers refer to electrode positions on the right hemisphere, and odd numbers refer to those on the left hemisphere. Z refers to an electrode placed on the midline.
The nasion, which is the point between the forehead and the nose, and the inion, which is the lowest point of the skull from the back of the head, are used as reference for the EEG electrodes. (Source: www.gtec.at)

2.3 Experimental Setup

The depth of unconscious cognitive processing can be investigated at various levels and using various approaches. More specifically the present thesis deals with the question of whether there is maintained some semantic processing in the unconscious state of sleep and if it is possible to show that auditory stimuli presented to the sleeping subject can reach higher levels of processing.
To answer the question, data from an experiment conducted by Dr. Sid Kouider at Laboratoire de Sciences Cognitives et Psycholinguistique, École Normale Supérieure is analyzed. In this study it was tested if an association between a semantic category and a specific motor response learned during wakefulness can be maintained in sleep. More specifically, it was tested whether a learned motor mapping association between a lateralized motor action and specific semantic category is preserved during early sleep stages. The experiment relies on an induction strategy to overcome the problem of learning a completely new task-set during sleep. Subjects were presented a task-set while they were still awake and then tested whether this task-set was maintained after subjects fell asleep, see Figure 2.6. In this section the experimental setup will be described briefly.

2.3.1 Procedure

Subjects were instructed to do a categorization of spoken words by pressing a button with their left or right hand depending on corresponding semantic category, i.e. animals or objects, see Figure 2.6. While doing the classification subjects were lying in a comfortable chair in a dark room with their eyes closed to encourage the transition towards sleep. The auditory stimuli was presented in headphones and subjects were instructed that they could fall asleep anytime during the experiment but were also asked not to stop responding voluntarily to easier fall asleep. When the subjects were assessed to be asleep a new list of words was introduced. This new list had the same properties, see Section 2.3.2, as the list presented during sleep, but was introduced to test for genuine semantic effects rather than simple stimulus-response associations [Kouider and Dehaene, 2007].

2.3.2 Stimuli

The spoken words used as stimuli were selected from the CELEX lexical database (Linguistic Data Consortium, University of Pennsylvania). There were 48 names

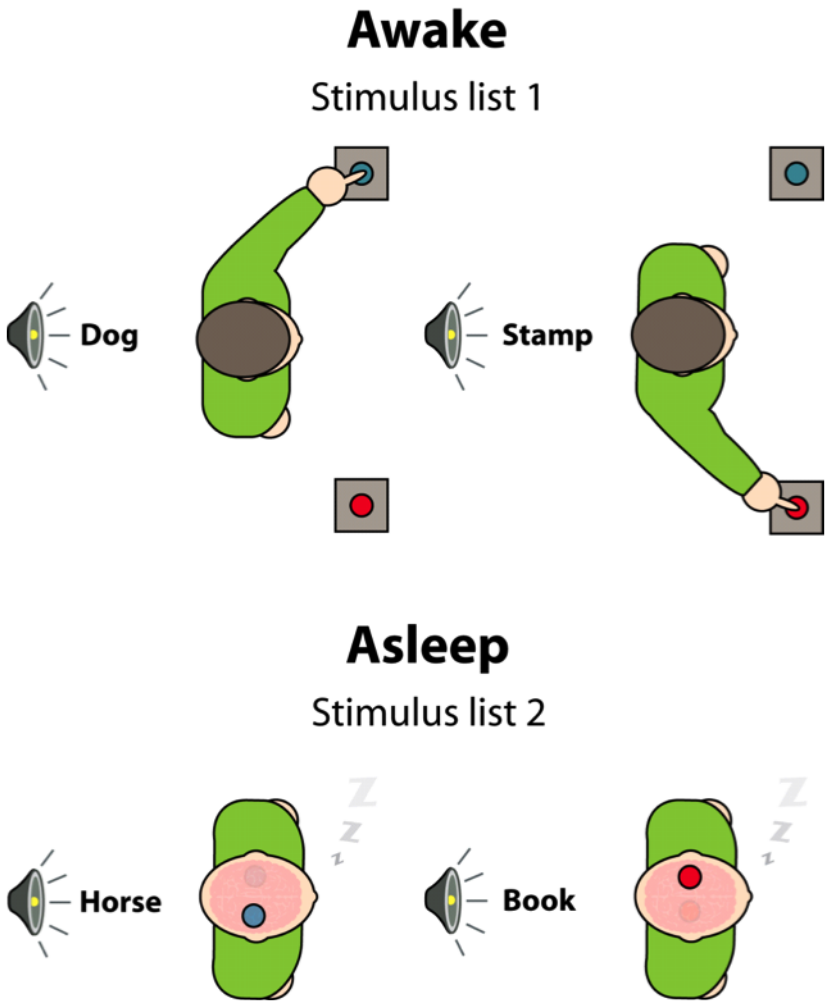


Figure 2.6: Before falling asleep subjects had to classify a word presented to them through headphones every 6 to 9 seconds as either animals or objects. This task allowed the mapping of each specific category with a specific motor response. This induction of a category-response mapping just before the onset of sleep is believed to promote the maintenance the task-set even after sleep onset. Testing conditions encouraged the transition towards sleep while remaining engaged with the same task-set. For each subject one of two lists of words was presented during wakefulness and the other list during sleep ensuring actual abstract categorization rather than simple stimulus-response associations. (Source: Sid Kouider)

of objects and 48 names of animals. Half were monosyllabic and the other half disyllabic, with animal and object names matched as closely as possible in terms of combined (spoken and written) log lemma frequencies, as confirmed by an independent t-test ($p > 0.10$). Additionally, words within the two categories were matched in a pair-wise fashion regarding their phonological properties: each object name was matched with a similar animal name (for example "quilt" was matched with "quail"), ensuring that animal and object names could not be differentiated in terms of phonological properties. The words were tape-recorded by a female voice and digitized. Durations of the resulting stimuli ranged from 357 to 800 ms. Two lists of 48 stimuli each were produced, one for the awakening period and the other for the sleeping period.

2.3.3 Sleep Assessment

Sleep onset was assessed both online and offline. During the experiment subjects were assessed to be asleep when showing no overt response for at least two minutes of stimulation and if the EEG showed sleep markers before and after the presentation of each word, i.e. vertex sharp waves, regular spontaneous and evoked K-complexes, sleep spindles, and an overall reduction of fast, alpha/beta rhythms in favor of slower delta/theta rhythms, cf. Section 2.2. After the experiment was finished, this was verified offline, and ambiguous epochs were excluded from the analysis.

2.3.4 Subjects

18 out of 47 subjects fell asleep for at least 9 consecutive minutes and were included in this study. Of these, 6 were women and 12 were men in the age range 18-30-years-old. They were all healthy native English speakers, right-handed and reported no auditory, neurological or psychiatric alterations. Only self reported easy sleepers [Johns et al., 1991] were chosen for the experiment to increase the probability that subjects would fall asleep. Subjects were also asked to avoid exciting substances as coffee, and to sleep 1-2 hours less than usual the night preceding the experiment. They signed a written consent and were paid for their participation.

2.3.5 EEG Equipment

The electroencephalogram was continuously recorded from 64 Ag/AgCl electrodes mounted on an electrode cap (Easycap, Falk Minow Services, Herrsching-Breitbrunn, Germany) using SynAmps amplifiers (NeuroScan Labs, Sterling, VA), with Cz as a reference. The impedance for electrodes was kept below 6 kΩ. Data were acquired with a sampling rate of 500 Hz, and then down sampled at 250 Hz. An electrooculogram (EOG) was recorded through electrodes placed above and below the right eye (vertical) and at the outer canthi (horizontal). Amplifier band pass was 0.1-100 Hz.

Chapter 3

Support Vector Machines

Statistical learning theory [Vapnik, 2006, Vapnik, 2000, Vapnik, 1998] provides the theoretical basis for machine learning and Support Vector Machines, SVM. It deals with the problem of inferring a predictive function based on empirical data using concepts from the fields of statistics and functional analysis.
SVMs have developed in several directions such that they have applications both within regression estimation as well as single-class and multi-class classification. This thesis deals with the formulation dealing with two-class pattern recognition in the non-linear version [Boser et al., 1992, Cortes and Vapnik, 1995, Vapnik, 1998].
The foundation of the SVM is the separating hyperplane. From this an optimal margin hyperplane can be defined and extended to the case where data are non-separable. Furthermore the optimal margin hyperplane can be generalized to a non-linear version where it is computed in a feature space non-linearly related to the input space. The following review is inspired by [Schölkopf and Smola, 2002, Bishop, 2007].

3.1 The Learning Setting

Suppose m observations, where each observation belongs to only one of two different classes, are given. Each observation consists of a pattern vector $x_i \in \mathcal{X},\ i = 1, ..., m$ and the associated class label y_i, which for mathematical convenience is labelled by either $+1$ or -1 in the simple binary classification problem. In the present thesis this corresponds to the classification of words belonging to one of two classes. In the frame of mathematical learning the goal is to be able to generalize to unseen data, i.e. for a new $\mathbf{x} \in \mathcal{X}$ it is possible to predict the corresponding y $\in \pm 1$. Again in the present thesis, given an EEG epoch recorded during auditory stimulation, it is possible to predict the word class. In the following it is assumed that there exists some unknown but fixed probability distribution P(x,y) from which these data are drawn and the data are assumed i.i.d. For all $\mathbf{x} \in \mathcal{X}$ we want to estimate a function $f : \mathcal{X} \to \{\pm 1\}$.

3.2 Separating Hyperplanes

If the pattern vectors are given in a dot product space $\mathbf{x} \in \mathcal{H}$, then any hyperplane can be written as

$$\{\mathbf{x} \in \mathcal{H} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\},\ \mathbf{w} \in \mathcal{H},\ b \in \Re, \tag{3.1}$$

where $\mathbf{w}$ is an orthogonal vector to the hyperplane, see also Figure 3.1.

The dot product is a simple similarity measure, since it represents geometric constructions that can be formulated in terms of angles, lengths and distances. Later a *kernel* function, k, will be introduced since it turns out that in many problems the dot product is not sufficiently general. However, both the dot product and kernels gives a way to characterize similarity in two patterns geometrically and hence the ability to construct learning algorithms using linear algebra and analytic geometry. In both cases the space $\mathcal{H}$ is called a feature space.
If the hyperplane is scaled such that the point closest to the hyperplane has a distance of $\frac{1}{\|\mathbf{w}\|}$, the hyperplane is said to be canonical and much freedom in choosing a hyperplane is gone. Nevertheless, it is still possible to choose both $(\mathbf{w}, b)$ and $(-\mathbf{w}, -b)$. Without class labels it is not possible to distinguish these hyperplanes. For pattern recognition problems they are different as they make

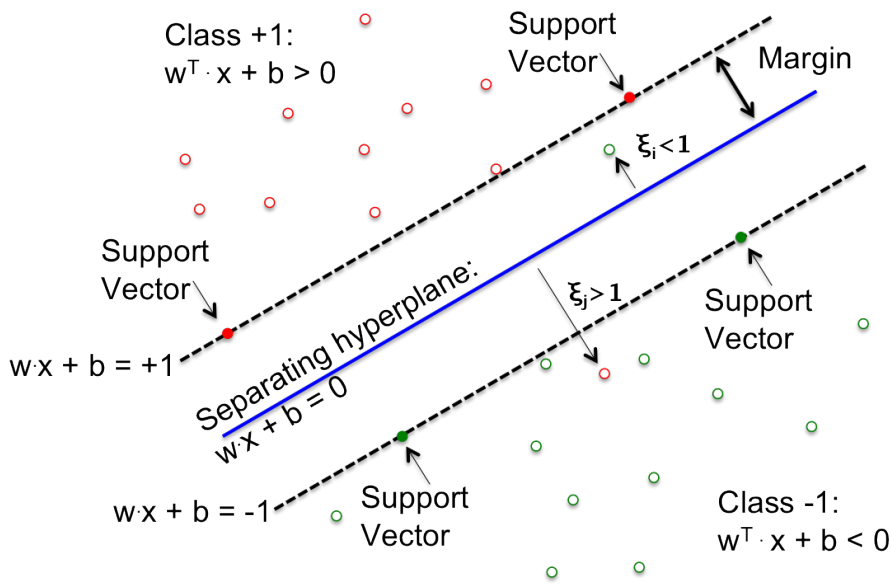


Figure 3.1: The SVM "learns" a hyperplane which gives the best separation of two classes. The figure shows a 2-dimensional classification problem where red dots correspond to the class label $+1$ and green dots corresponds to the class label -1.

opposite class assignments using the two inversely correlated decision functions

$$\begin{aligned} f_{\mathbf{w},b} &: \mathcal{H} \to \{\pm 1\} \\ x &\mapsto f_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b). \end{aligned} \tag{3.2}$$

From this it is clear how a separating hyperplane can be constructed if the data are separable, i.e. two disjoint sets with corresponding different class labels.
In creating learning algorithms it proves useful to define the margin as well. The margin is defined as the perpendicular distance between the decision boundary and the closest data point, hence for a hyperplane, the geometric margin of the point $(\mathbf{x}, y) \in \mathcal{H} \times \{\pm 1\}$ is defined as

$$\rho_{(\mathbf{w},b)}(\mathbf{x}, y) := y(\langle \mathbf{w}, \mathbf{x} \rangle + b)/\|\mathbf{w}\|, \tag{3.3}$$

and the minimum value

$$\rho_{(\mathbf{w},b)} := \min_{i=1,\ldots,m} \rho_{(\mathbf{w},b)}(\mathbf{x}_i, y_i), \tag{3.4}$$

is the geometrical margin, or just the margin, of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$.
For a correctly classified point $(\mathbf{x}, y)$ the margin is the distance from $\mathbf{x}$ to the hyperplane and for a misclassified point the margin gives a negative distance. This

can be seen from the fact that for the canonical hyperplane considered the margin is $1/\|\mathbf{w}\|$ and the length of the weight vector is 1. Hence it is the projection of $\mathbf{x}$ onto the direction orthogonal to the hyperplane. The idea behind SVMs is to choose a decision boundary for which the margin of the separating hyperplane is maximized. The intuition behind this is that a large margin will give the optimal separation of the data. Furthermore, since the observed data is assumed to have been generated by the same underlying process it seems reasonable to assume that new observations will lie close (in $\mathcal{H}$) to one of the training patterns.

3.3 Optimal Margin Hyperplanes

An important property of support vector machines is that the determination of the model parameters eventually corresponds to a convex optimization problem, hence any local solution is also a global optimum. Finding the optimal separating hyperplane (maximum margin) is the heart of the support vector machine and it basically boils down to optimizing the parameters w and b in order to maximize the decision boundary.
For the canonical hyperplane all points will satisfy $y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1$ and the decision boundary is optimized when $\|\mathbf{w}\|^{-1}$ is maximized. Hence, for a linearly separable set of training data the optimal separating hyperplane can be found from the following quadratic optimization problem

$$\min_{\mathbf{w}\in\mathcal{H}, b\in\Re} \quad \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.5}$$

$$\text{subject to: } y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1 \quad \forall \quad i = 1, ..., m. \tag{3.6}$$

Solving this primal problem will result in $(\mathbf{w}, b)$ with the largest possible geometric margin with respect to the training set. The dual problem however, can give some additional insight and it is here the foundation of the SVM is found. The Lagrangian of the inequality constrained primal convex quadratic optimization problem is given by

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - 1), \tag{3.7}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.
The corresponding Karush-Kuhn-Tucker, KKT, optimality conditions, [Nocedal and Wright, 1999] are both necessary and sufficient conditions for optimality since both the objective function and the inequality constraints are continuously differentiable convex functions.
To obtain the same solution as to the primal problem, the Lagrangian must be

minimized with respect to $\mathbf{w}$ and b and maximized with respect to α_i, hence the solution is found at a saddle point. Minimizing the Lagrangian with respect to $\mathbf{w}$ and b yields two conditions

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0 \tag{3.8}$$

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \tag{3.9}$$

which implies that

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.10}$$

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i. \tag{3.11}$$

From 3.11 it can furthermore be seen that the unique solution vector (due to convexity) has an expansion only in terms of the training data. If 3.11 is plugged into the Lagrangian 3.7, we obtain

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - b \sum_{i=1}^{m} \alpha_i y_i, \tag{3.12}$$

which can be reduced using 3.10 such that

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j. \tag{3.13}$$

Combined with the constraints, we have the dual form of the primal optimization problem

$$\max_{\alpha \in \Re^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{3.14}$$

$$\text{subject to:} \quad \alpha_i \geq 0, \quad i = 1, ..., m \tag{3.15}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0. \tag{3.16}$$

This is the foundation of the SVM algorithm, written only in terms of the inner product between points in the input feature space and the parameters (Lagrange multipliers) α_i. For every training point there is a Lagrange multiplier α_i. At the solution, those points, $\mathbf{x}_i$ for which $\alpha_i > 0$ are called support vectors and they lie exactly on the margin. All other training points have $\alpha_i = 0$ and are irrelevant as they do not appear in (3.11).

Now suppose the models α_i's are found using a training set, and we wish to make a prediction at a new input $\mathbf{x}$. We would then calculate $\langle \mathbf{w}, \mathbf{x} \rangle + b$, and predict y = 1 if and only if this quantity is bigger than zero. But using (3.11) , this quantity can also be written

$$\mathbf{w}^T\mathbf{x} + b = (\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i)^T \mathbf{x} + b = \sum_{i=1}^{m} \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b. \tag{3.17}$$

Hence the prediction only depends on the inner product between the new point x and the points in the training set. Moreover, the α_i's will all be zero except for the support vectors. Thus, many of the terms in the sum will be zero, and only the inner products between x and the support vectors (of which there is often only a small number) need to be calculated in order to calculate (3.17) and make a prediction using the *sgn* function.

3.4 Soft Margin Optimal Hyperplanes

The classifier considered so far is ideally suited for linearly separable data without outliers. However, in practice data is rarely in that condition, and the algorithm must be adjusted to work for non-separable data sets and to be less sensitive to outliers. To allow some of the training points to be misclassified, the optimization problem can be reformulated by introducing slack variables for each training point, $\boldsymbol{\xi}_i \geq 0, i = 1, ..., m$ and relax the separation constraints 3.6 such that

$$y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \boldsymbol{\xi}_i \quad i = 1, ..., m. \tag{3.18}$$

Correctly classified training points that are either on the margin or on the correct side of the margin yields $\boldsymbol{\xi}_i = 0$. Points that lie inside the margin, but on the correct side of the decision boundary have $0 < \boldsymbol{\xi}_i \leq 1$, and misclassified data points on the wrong side of the decision boundary yields $\boldsymbol{\xi}_i \geq 1$, see Figure 3.1. This allows the constraints to be satisfied by making $\boldsymbol{\xi}_i$ large enough as it relaxes the hard margin constraint to give a "soft" margin and allows some of the training set data points to be misclassified. However, it is necessary to penalize large values of $\boldsymbol{\xi}_i$ in order not to obtain the trivial solution where all $\boldsymbol{\xi}_i$'s are large. This can be done by including the slack variables in the objective function of (3.5), hence our goal is now to maximize the margin while penalizing points

that lie on the wrong side of the margin. This can be formulated as follows

$$\min_{\mathbf{w}\in\mathcal{H},b\in\Re} \quad \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\boldsymbol{\xi}_i \tag{3.19}$$

$$\text{subject to:} \quad y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1 - \boldsymbol{\xi}_i \quad i = 1, ..., m \tag{3.20}$$

$$\boldsymbol{\xi}_i \geq 0 \quad i = 1, ..., m, \tag{3.21}$$

where the parameter $C > 0$ is similar to a regularization coefficient because it controls the trade-off between minimizing training errors (corresponding to the non-zero slack variables and the corresponding penalty) and controlling model complexity (maximizing margin). The original formulation can for separable data be recovered in the limit where $C \to \infty$. The selection of the C parameter has proven to be rather unintuitive, and there is no obvious a priory way of selecting it, other than searching a wide range of values, [Shawe-Taylor and Cristianini, 2004]. As for the original primal problem (3.5) - (3.6) it is possible to obtain a dual formulation. The Lagrangian is given by

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\boldsymbol{\xi}_i - \sum_{i=1}^{m}\alpha_i[y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - 1 + \boldsymbol{\xi}_i] - \sum_{i=1}^{m}\mathbf{r}_i\boldsymbol{\xi}_i, \tag{3.22}$$

where $\boldsymbol{\xi}_i \geq 0$ and $\mathbf{r}_i \geq 0$ are the corresponding Lagrange multipliers. Following the same procedure as previous the dual formulation can be obtained

$$\max_{\alpha\in\Re^m} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle \tag{3.23}$$

$$\text{subject to:} \quad 0 \leq \alpha_i \leq C, \quad i = 1, ..., m \tag{3.24}$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0, \tag{3.25}$$

where the only change turns out to be an upper bound on the α_i's, these constraints are known as box constraints. Predictions for new data points are done using 3.17.

3.5 The Non-linear Support Vector Machine for Non-separable Data

A very important extension of the presented classifier comes with the introduction of kernels. Everything in the setting so far deals with classification of

more or less linearly separable data. In the following kernels are introduced to non-linearly transform the input data, now denoted $x_1, ..., x_m \in \mathcal{X}$, using a map $\boldsymbol{\phi} : x_i \to \mathbf{x}_i$ into a high-dimensional feature space and do the linear separation there. In practice it requires only small modifications of the presented formulation and the transformation leads to a much more powerful classification tool. As with the dot product, the kernel function is used as a similarity measure, and a large class of kernels actually admit a dot product representation in a feature space. Kernels can in general be regarded as generalized dot products and any dot product is in fact a kernel. More formally the class of kernels k that correspond to a dot product in a feature space $\mathcal{H}$ via a map $\boldsymbol{\phi}$ that satisfies

$$\begin{aligned} \phi &: \mathcal{X} \to \mathcal{H} \\ x &\to \mathbf{x} := \boldsymbol{\phi}(x), \end{aligned} \tag{3.26}$$

then

$$k(x, x_i) = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x_i) \rangle. \tag{3.27}$$

There are no constraints on the structure of the domain $\mathcal{X}$ other than it needs to be a non-empty set. Since it is possible to compare similarities between non-vectorial objects such as strings [Haussler, 1999] it makes kernels applicable in situations where vectorial representation is not readily available and expands the field of kernel methods. In this thesis however, only vectorial data is considered. The term kernel originates from the first use of this type of function in the field of integral operators. They were originally introduced since there are many classes of problems that are harder to solve in their original representations. An integral transform maps a function from its original domain into another domain. Solving the equation in the target domain can be easier than in the original domain. The solution can then be mapped back to the original domain using the inverse of the integral transform.

Definition 3.1 [Polyanin and Manzhirov, 2008]
A function k which gives rise to an operator T_k via
$(T_k f)(x) = \int_{\mathcal{X}} k(x, x_i) f(x_i) dx_i$
is called the kernel of T_k.

For a kernel to be valid and describe a dot product in some feature space it generally needs to satisfy Mercer's Theorem, see e.g. [Mercer, 1909, Schölkopf and Smola, 2002]. Usually Mercer's theorem is presented in a form involving L2 functions, but when the input data take values in $\mathcal{R}^n$ as in this thesis, it is equivalent to:

Theorem 3.2 (Mercer) *Let $K : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}^n$ be given.*
Then for K to be a valid kernel, it is necessary and sufficient that for any

$\{x_1, ..., x_m\}, (m < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

In continuation of Mercers Theorem, the following proposition is used:

Proposition 3.3 *[Schölkopf and Smola, 2002]*
If k is a kernel satisfying the conditions of Mercers Theorem, we can construct a mapping $\boldsymbol{\phi}$ into a space where k acts as a dot product,
$\langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x_i) \rangle = k(x, x_i)$,
for almost all (except for sets of measure zero) $x, x_i \in \mathcal{X}$. Moreover, given any $\epsilon \geq 0$, there exists a map $\boldsymbol{\phi}_n$ into an n-dimensional dot product space (where $n \in \mathcal{N}$ depends on ϵ) such that
$|k(x, x_i) - \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x_i) \rangle| \leq \epsilon$
for almost all (except for sets of measure zero) $x, x_i \in \mathcal{X}$.

Positive definite kernels are also called reproducing kernels [Schölkopf and Smola, 2002] and can thought of as a set of dot products in another space. The reproducing kernel property amounts to

$$\langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x_i) \rangle = k(x, x_i), \tag{3.28}$$

which is also the basis of the "kernel trick", which basically states that any algorithm formulated in terms of a positive definite kernel, k, can be reformulated to an alternative algorithm by replacing k by a new positive definite kernel $\tilde{k}$. The Reproducing Kernel Hilbert Spaces theory more precisely states which kernel functions correspond to a dot product and the linear spaces that implicitly are induced by these kernel functions, see [Schölkopf and Smola, 2002].
An example of this is an algorithm where the k is the dot product in the input domain such as the formulation of the optimal separating hyperplane. If the formulation of the optimal hyperplane everything can be rewritten in terms of $\boldsymbol{\phi}(s)$ instead of x and then using the kernel trick we have a way of write a nonlinear operator as a linear one in a space of higher dimension

$$\begin{aligned} &\max_{\alpha \in \Re^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ &\text{subject to: } \; 0 \leq \alpha_i \leq C, \quad i = 1, ..., m \\ &\qquad \sum_{i=1}^{m} \alpha_i y_i = 0, \end{aligned} \tag{3.29}$$

with the corresponding decision function

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b \right). \tag{3.30}$$

For support vectors $\mathbf{x}_j$ for which $\xi_j = 0$ the threshold b can be computed by averaging 3.20 over all support vectors $\mathbf{x}_j$, since they satisfy $0 < \alpha_j < C$.

3.5.1 Kernel

One function satisfying the properties described in the previous section is the Radial Basis Function, RBF, kernel. In general the RBF kernel shows some attractive properties and performs very well for a wide range of problems, see [Caputo et al., 2002]. Any continuous decision boundary can be obtained using the RBF kernel, but with proper parameter selection it makes the SVM behave like a simple linear classifier, see section 5.5.1. The RBF kernel is given by

$$K(x, x_i) = e^{-\frac{||x-x_i||}{2\sigma^2}}. \tag{3.31}$$

Other kernel functions are e.g. the linear kernel, the polynomial kernel, the spline kernel, the Fourier kernel, and the Sigmoid kernel. Except for the linear kernel which is considered briefly, these are not considered further.

3.6 Numerical Optimization

As long as the kernel matrix fits the main memory of modern computers, fast and accurate solutions exist in terms of Quadratic Programming, QP, solvers. In many real life problems the kernel matrix however, is too large to make the full problem tractable.

The Sequential Minimal Optimization, SMO, was introduced by Platt [Platt, 1998], improved by [Keerthi et al., 2001] and a modified version [Fan et al., 2005] is implemented in LIBSVM [Chang and Lin, 2011] in an analogous version of the quadratic formulation presented in previous sections. The SMO algorithm is a widely applied approach for solving the considered QP problem as it has desirable properties for large-scale problems. The SMO algorithm consists in most implementations of an analytical part for optimizing the smallest possible subproblem consisting of two multipliers, and a heuristic part for choosing which two multipliers to optimize.

3.6.1 Subset selection

The quadratic optimization problem in 3.29 is covered by the following QP

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha + c^T\alpha \\ \text{subject to} \quad & A\alpha = d \\ & 0 \leq \alpha \leq u, \end{aligned} \tag{3.32}$$

which can be formulated as a convex program in a subset of the variables. Assume there exists a subset, the working set, $S_w \subset [m]$ which will be used during optimization and a fixed set $S_f = [m] \setminus S_w$ which will not be modified. Then $Q, c, A,$ and u can be split up accordingly into the following permutation matrices

$$Q = \begin{bmatrix} Q_{ww} & Q_{fw} \\ Q_{wf} & Q_{ff} \end{bmatrix}, \quad c = (c_w, c_f), \quad A = [A_w, A_f], \quad u = (u_w, u_f), \tag{3.33}$$

and the QP can be restated as

$$\begin{aligned} \min_{\alpha_w} \quad & \frac{1}{2}\alpha_w^T Q_{ww}\alpha_w + [c_w + Q_{wf}\alpha_f]^T\alpha + [\frac{1}{2}\alpha_f^T Q_{ff}\alpha_f + c_f^T\alpha_f] \\ \text{subject to} \quad & A_w\alpha_w = d - A_f\alpha_f \\ & 0 \leq \alpha_w \leq u_w, \end{aligned} \tag{3.34}$$

where the constant offset produced by α_f is not to be considered in the actual optimization. Solving the subset problem will lead to an improvement of the full problem, and several heuristics have been proposed for choosing the working set.

3.6.2 Sequential Minimal Optimization

The SMO algorithm is the extreme case of the above, where the working set only consist of two variables

$$\begin{aligned} \min_{\alpha_i,\ \alpha_j} \quad & \frac{1}{2}[\alpha_i^2 Q_{ii} + \alpha_j^2 Q_{jj} + 2\alpha_i\alpha_j Q_{ij}] + c_i\alpha_i + c_j\alpha_j \\ \text{subject to} \quad & s\alpha_i + \alpha_j = \gamma \\ & 0 \leq \alpha_i \leq C_i \\ & 0 \leq \alpha_j \leq C_j, \end{aligned}$$

where $s \in \pm 1,\ Q \in \Re^{2x2}$, and $c_i,\ c_j,\ \gamma \in \Re$ are chosen accordingly.
There exist an analytic solution to this optimization problem. The following

shows the derivation and finds the explicit values needed during iterations of the algorithm. By using the equality constraint, $s\alpha_i+\alpha_j = \gamma$, it is possible to express the objective function only in terms of α_i since $\alpha_j = \gamma - s\alpha_i$. Furthermore, due to the constraints on α_j, $s\alpha_i = \gamma - \alpha_j$, the following bound $\gamma \geq s\alpha_i \geq \gamma - C_j$ applies. Combining this with the bound on α_i, $0 \leq \alpha_i \leq C_i$, it is possible to obtain the following constraint: $H \geq \alpha_i \geq L$, where

$$L = \begin{cases} \max(0, s^{-1}(\gamma - C_j)) & \text{if } s > 0 \\ \max(0, s^{-1}\gamma) & o.w. \end{cases}$$

$$H = \begin{cases} \min(C_i, s^{-1}\gamma) & \text{if } s > 0 \\ \min(C_i, s^{-1}(\gamma - C_j)) & o.w. \end{cases}$$

Then, with the new bound on α_i, it is possible to substitute $\alpha_j = \gamma - s\alpha_i$ and the QP can then be stated only in terms of α_i as

$$\begin{aligned} \min_{\alpha_i} \quad & \frac{1}{2}\alpha_i^2(Q_{ii} + Q_{jj} - 2sQ_{ij}) + \alpha_i(c_i - sc_j + \gamma Q_{ij} - \gamma s Q_{jj}) \\ \text{subject to} \quad & L \leq \alpha_i \leq H. \end{aligned} \tag{3.35}$$

By introducing the auxiliary variables

$$\Gamma = sc_j - c_i + \gamma s Q_{jj} - \gamma Q_{ij} \tag{3.36}$$

$$\Lambda = (Q_{ii} + Q_{jj} - 2sQ_{ij}), \tag{3.37}$$

the unconstrained objective function can be written as: $\frac{\Lambda}{2}\alpha_i^2 - \Gamma\alpha_i$. By taking the derivative, the corresponding unconstrained minimum is obtained at $\alpha_i = \Lambda^{-1}\Gamma$. To ensure that the solution is within the constrained interval $\alpha_i \in [L, H]$ the unconstrained solution is cut to the interval, i.e. $\alpha_i = min(max(\Lambda^{-1}\Gamma, L), H)$. In the case of classification it must hold that $\sum_{i=1}^{m} y_i\alpha_i = 0$ and hence $y_i\alpha_i + y_j\alpha_j = y_i\alpha_i^{old} + y_j\alpha_j^{old}$. This gives $\gamma := y_i y_j \alpha_i + \alpha_j = y_i y_j \alpha_i^{old} + \alpha_j^{old}$ and $s = y_i y_j$.
Furthermore from (3.33) it is given that $Q_{ii} = K_{ii}$, $Q_{jj} = K_{jj}$, $Q_{ij} = Q_{ji} = sK_{ij}$ where $K_{ij} := k(x_i, x_j)$ is the kernel matrix and hence

$$\Lambda = K_{ii} + K_{jj} + 2K_{ij}. \tag{3.38}$$

To find Γ, c_i and c_j can be obtained from (3.34)

$$c_i = -1 + y_i \left(\sum_{l \neq i,j}^{m} \alpha_l k(x_i, x_l) \right) = y_i(f(x_i) - b - y_i) - \alpha_i K_{ii} - \alpha_j s K_{ij} \tag{3.39}$$

$$c_j = -1 + y_j \left(\sum_{l \neq i,j}^{m} \alpha_l k(x_j, x_l) \right) = y_j(f(x_j) - b - y_j) - \alpha_i K_{jj} - \alpha_i s K_{ij}. \tag{3.40}$$

Then Γ can be computed using $y_i = y_j s$

$$\begin{aligned}\Gamma =& - y_i(f(x_i) - b - y_i) + \alpha_i K_{ii} + \alpha_j s K_{ij} + y_i(f(x_j) - b - y_j) \\ & + \alpha_j s K_{jj} \alpha_i K_{ij} + (\alpha_i + s\alpha_j)(K_{ij} - K_{jj} \\ =& y_i((f(x_j) - y_j) - (f(x_i) - y_i)) + \alpha_i \Lambda. \end{aligned} \tag{3.41}$$

Plugging back into the original formulation yields the following results:
If $y_i = y_j$

$$L = \max(0, \alpha_i^{old} + \alpha_j^{old} - C_j) \tag{3.42}$$

$$H = \min(C_i, \alpha_i^{old} + \alpha_j^{old}), \tag{3.43}$$

and if $y_i \neq y_j$

$$L = \max(0, \alpha_i^{old} + \alpha_j^{old}) \tag{3.44}$$

$$H = \min(C_i, C_j + \alpha_i^{old} + \alpha_j^{old}), \tag{3.45}$$

then the optimal values are

$$\alpha_i = \min(\max(\bar{\alpha}, L), H) \tag{3.46}$$

$$\alpha_j = s(\alpha_i^{old} - \alpha_i) - \alpha_j^{old}, \tag{3.47}$$

where

$$\bar{\alpha} = \begin{cases} \alpha_i^{old} + \Lambda^{-1}\delta & if \quad \Lambda > 0 \\ -\infty & if \quad \Lambda = 0 \quad \text{and} \quad \delta > 0 \\ \infty & if \quad \Lambda = 0 \quad \text{and} \quad \delta < 0 \end{cases},$$

and $\delta := y_i((f(x_j) - y_j) - (f(x_i) - y_i))$.

From the above, it can be seen that if the constrained and unconstrained solution are identical, i.e. $\alpha_i = \bar{\alpha}$, then the objective function is improved by $\Lambda^{-1}((f(x_j) - y_j) - (f(x_i) - y_i))^2$. Hence it is important to select a working set which makes this term large, see e.g. [Keerthi et al., 2001].

3.6.3 Stopping Criterion

As the decomposition method asymptotically approaches an optimum, it is in practice terminated after satisfying a stopping criterion. Some methods focus on the precision of the Lagrange multipliers α_i, whereas others use the proximity of the primal and the dual objective functions [Schölkopf and Smola, 2002]. It is worth noticing that an improvement in the primal objective does not necessarily imply an improvement in the dual and vice versa. In SMO the dual gap can

fluctuate considerably. In the SMO-algorithm implemented in LIBSVM, [Chang and Lin, 2011], the KKT conditions are checked to be within ϵ of fulfillment: The standard KKT conditions [Nocedal and Wright, 1999] of the dual formulation, Equation 3.32, states that if there exist a scalar, b, and two nonnegative vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ such that

$$\nabla f(\alpha) + b\mathbf{y} = \boldsymbol{\lambda} + \boldsymbol{\mu} \tag{3.48}$$
$$\lambda_i \alpha_i = 0, \quad i = 1..m \tag{3.49}$$
$$\mu_i (C - \alpha_i) = 0, \quad i = 1..m \tag{3.50}$$
$$\lambda_i = 0, \quad i = 1..m \tag{3.51}$$
$$\mu_i = 0, \quad i = 1..m \tag{3.52}$$

where $\nabla f(\alpha) \equiv Q\alpha + c$ is the gradient of $f(\alpha)$. Then a feasible $\boldsymbol{\alpha}$ is a stationary point of 3.32. The conditions can be rewritten as

$$\nabla_i f(\alpha) + by_i \geq 0 \quad \text{if} \quad \alpha_i < C \tag{3.53}$$
$$\nabla_i f(\alpha) + by_i \leq 0 \quad \text{if} \quad \alpha_i > 0. \tag{3.54}$$

Utilizing the fact that $y_i = \pm 1$ this yields that there exists a b such that

$$m(\alpha) = \max_{i \in I_{hi}(\alpha)} -y_i \nabla_i f(\alpha) \leq b \leq M(\alpha) = \min_{i \in I_{lo}(\alpha)} -y_i \nabla_i f(\alpha), \tag{3.55}$$

where

$$I_{hi}(\alpha) \equiv \{t | \alpha_t < C, y_t = 1 \quad \text{or} \quad \alpha_t > 0, y_t = -1\} \tag{3.56}$$
$$I_{lo}(\alpha) \equiv \{t | \alpha_t < C, y_t = -1 \quad \text{or} \quad \alpha_t > 0, y_t = 1\}. \tag{3.57}$$

Hence for an $\boldsymbol{\alpha}$ to be feasible it must hold

$$m(\alpha) \leq M(\alpha), \tag{3.58}$$

which gives the stopping condition employed in LIBSVM

$$m(\alpha) - M(\alpha) \leq \epsilon, \tag{3.59}$$

where ϵ is the stopping tolerance. The SMO algorithm with the described stopping criterion has been shown to converge in a finite number of iterations [Chen et al., 2006, Fan et al., 2005, Keerthi and Gilbert, 2002].
The time required for the SMO algorithm to converge hence depends on the desired accuracy of the output but also on the working set selection. The literature investigating this kind of stopping tolerance for the SMO algorithm is limited, but it generally seems there is consensus that $\epsilon = 10^{-3}$ is the default value to use. [Joachims, 1999, Chang and Lin, 2011, Fan et al., 2005, Hsu and Lin, 2002] use $\epsilon = 10^{-3}$, with the general note that this is an acceptable value, though

without providing any evidence except from Platt's paper [Platt, 1998] which states: *"Recognition systems typically do not need to have the KKT conditions fulfilled to high accuracy: it is acceptable for examples on the positive margin to have outputs between 0.999 and 1.001".* Nevertheless, it is a very interesting parameter to investigate further since it obviously is a trade-off between accuracy and computational effort.

3.6.4 Implementation

LIBSVM [Chang and Lin, 2011] is a library for SVMs written in C++. It implements a version of a SVM for classification problems similar to the one described in the present chapter and solves the optimization problem using the SMO algorithm. LIBSVM has a compiled interface which allows all functions to be called from MATLAB. The library is modified in this thesis to produce a non-standard output used for the feature extraction algorithm described in Section 4.2.3. The code is slow since it, in addition to the suppressible outputs, produces non-suppressible outputs during every iteration. These are removed and a Matlab routine, which utilizes the SUN grid-engine cluster facilities at DTU Informatics, is written, to be able to process high-dimensional jobs in a high performing parallel environment.

Chapter 4

Feature Extraction

Machine learning methods, including SVMs, do not necessarily work well when applied to raw EEG data signal segments. One of the major difficulties in building a classification model based on EEG recordings is to find a good data representation. Feature extraction deals with construction and selection of relevant and informative features.

4.1 Feature Construction

New features can be constructed to get an appropriate data representation. According to [Guyon and Elisseeff, 2003], performance can often be improved using features derived from the original input. Building a new feature representation is also a way of incorporating domain knowledge. There are a number of generic feature construction methods, including clustering, linear transforms of the input variables, e.g. Principal Component Analysis, PCA, Linear Discriminant Analysis, LDA, and spectral transforms e.g. Fourier, multitaper, and wavelet transforms.
As described in Section 2.1, raw EEG signals are time series of voltage fluctuations resulting from ionic current flows within the neurons of the brain. Many applications and analyses does however, as described, generally focus on the spectral content of EEG, i.e. the type of neural oscillations that can be

observed in EEG signals and it has been demonstrated numerous times that assessing specific frequencies can often yield insights into the functional cognitive correlations of these signals. Hence instead of representing the data in the original domain, it can be transformed to the time-frequency domain.

4.1.1 Spectral Decomposition

In theory, every signal can be decomposed into sinusoidal oscillations of different frequencies. Such a decomposition is traditionally computed using a Fourier transform to quantify the oscillations that compose the signal [Nunez and Srinivasan, 2006]. Time-frequency analysis makes it possible to study oscillatory neural activity that appears consistently at particular times, relative to the event of interest, even if this activity is not phase locked to the event and therefore averages out in conventional analyses of evoked responses.
All methods of time-frequency analysis are inherently limited by the fact that the resolution in time is inversely related to frequency resolution, called the uncertainty principle [Percival and Walden, 1993]. Different methods of time-frequency analysis handle this trade-off slightly different and are therefore optimal for certain kinds of signals and suboptimal for others.

4.1.1.1 Wavelets and Multitapers

Wavelet transforms have shown advantageous when handling the trade-off between temporal resolution and frequency resolution in the analysis of EEG signals, [Mørup et al., 2007, van Vugt et al., 2007]. The continuous wavelet transform is very similar to a short time Fourier transform, but instead of having the same window length at all frequencies, it varies the window length over different frequencies. The wavelet length is shorter for higher frequencies than lower frequencies, which is desirable for EEG since high frequencies generally vary more rapidly in time than low frequencies. There are various types of wavelets, but the Morlet wavelet, which compares the signal with short segments of an oscillation multiplied by a Gaussian window function, is widely used in EEG analysis [Mørup et al., 2007, Oostenveld et al., 2011]. The continuous wavelet transform for a sampled signal $x(t_n)$ is defined at time t_0 using the wavelet coefficient

$$X(t_0, a) = \frac{1}{\sqrt{a}} \sum_{n=-\infty}^{\infty} \tilde{\phi}(\frac{t_n - t_0}{a}) x(t_n), \tag{4.1}$$

with scale a and mother wavelet

$$\tilde{\phi}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-i2\pi t} e^{-\frac{t^2}{2\sigma^2}}, \tag{4.2}$$

where the number of oscillations included in the analysis is defined from the width of the wavelet, $2\pi\sigma$. The width of the wavelets is given in number of cycles, where smaller values will increase the temporal resolution at the expense of frequency resolution.
Multitapers, [Percival and Walden, 1993], have also been proposed and adopted to analyses of EEG data because of their properties [Mitra and Pesaran, 1999, Raghavachari et al., 2001, Hoogenboom et al., 2006]. Multitapers differ from wavelets in that the width of the function stays the same in absolute time across frequencies, which is similar to a Fourier transform. In the Multitaper transform, the original signal is multiplied with Slepian windows, which are designed to prevent leakage of power to neighboring frequencies. After multiplication of the window function, a Fourier Transform is performed, and the absolute square is taken of the resulting signal. The convolution is repeated with a number of orthogonal windows to reduce the variance of the estimate. Both transforms produce spectrograms similar to the ones in Figure 4.1.

4.2 Feature Selection

The presented EEG classification problem using a spectral transform comprise numerous input features. The number of features for the transformed EEG data is a function of the number of channels, the time resolution, and the frequency resolution. In the setting described there are hundreds of thousands of initial feature dimensions in contrast to the small number of trials.
Many features do not contain relevant information for the classification problem and some input features are more likely comprised of noise and hence only correlate with the task labels of the training set by chance. A classifier trained on these features might overfit to these false regularities and fail out-of-sample. Figure 4.2 illustrate the trade-off between too many and too few features in relation to out-of-sample classification error. Furthermore, high dimensional input features enlarge the complexity and capacity needed to reach a good separation on the training data. If noisy features can be removed, the capacity is not unnecessarily increased. This can also prevent the classifier from overfitting the training data.
In the current setting with sleeping subjects there are too few training vectors to cope with the original dimension. Only a median of 17 trials (range 7 - 24) in each class (left/right) was recorded during sleep, hence the classification only have few examples in each condition compared to the dimension of the input

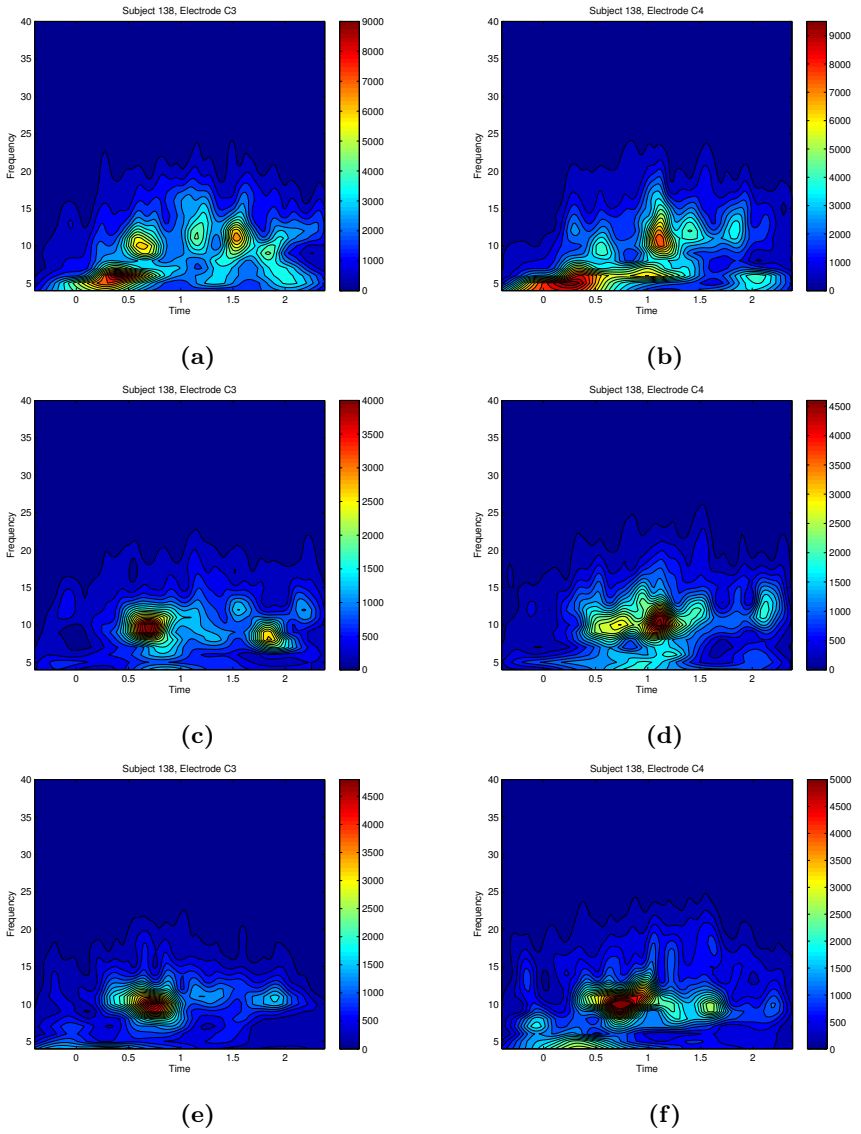


Figure 4.1: Averaged spectrograms in (a) and (b) for one condition and averaged artifact rejected spectrograms in (c) and (d) for the same condition where the k-complex occurring 300 ms after the stimulus is removed. Furthermore averaged artifact rejected spectrograms for the other condition is shown in (e) and (f). Just by visual inspection it seems obvious that some processing takes place after approximately 800 ms, but it is hard to distinguish the two conditions by visual inspection. The spectrograms are constructed using a wavelet transform.

space. One way to obtain more examples is to use the trials across subjects, but it has been shown in BCI settings that "good" features usually cannot be transferred from one person to another without problems. Brain structures are not organized exactly the same way, and hence brain signal characteristics vary between subjects, and some subjects even lack typical mu-rhythm activity, see e.g. [Lal et al., 2004, Pfurtscheller and Lopes da Silva, 1999]. Furthermore, the recording positions can not be controlled in enough detail to transfer directly since the exact positioning of EEG caps is difficult and the re-application at different recording sessions does not result in identical recording positions. This means that an optimal set of features has to be determined for each subject though there might be some overlap.
However, there are also obvious benefits in reducing the number of input features. The interpretability of the recorded signal is improved by extracting a few significant features. For new experimental paradigms, prior knowledge about the importance of features is limited and might even be misleading if transferred directly from other paradigms.
Selecting a subset of features can also lead to improved experiments. If only a subset of channels proves to be relevant future, EEG research can e.g. rely on a reduced set of electrodes.

4.2.1 NIPS 2003 Feature Selection Challenge Summary

One of the most rigorous investigations of different feature selection schemes is the NIPS 2003 Feature Selection Challenge [Guyon et al., 2004]. This is the main inspiration for the selected methods in this thesis. To summarize the results of the NIPS 2003 Feature Selection Challenge [Guyon et al., 2006] it is found that non-linear classifiers generally outperform linear classifiers in the competition. Even for data sets with a high number of features and few examples adequate regularization of nonlinear classifiers leads to better performance than linear classifiers. The SVM is furthermore identified as a versatile classifier.
7 out of 10 of the best performing classifiers use some feature selection strategy where most used forward selection or backward elimination inspired approaches. Forward selection algorithms starts with an empty set of features and then progressively add features based on some measure of improvement of performance. The backward elimination algorithms starts with a full set of unranked features and then progressively removes the least important features. Though some of the top entrants use embedded methods, others perform well using all the available features. Unsupervised dimensionality reductions methods, e.g. Principal Component Analysis, PCA, are shown to work well, as are filter methods like the Pearson correlation coefficient.

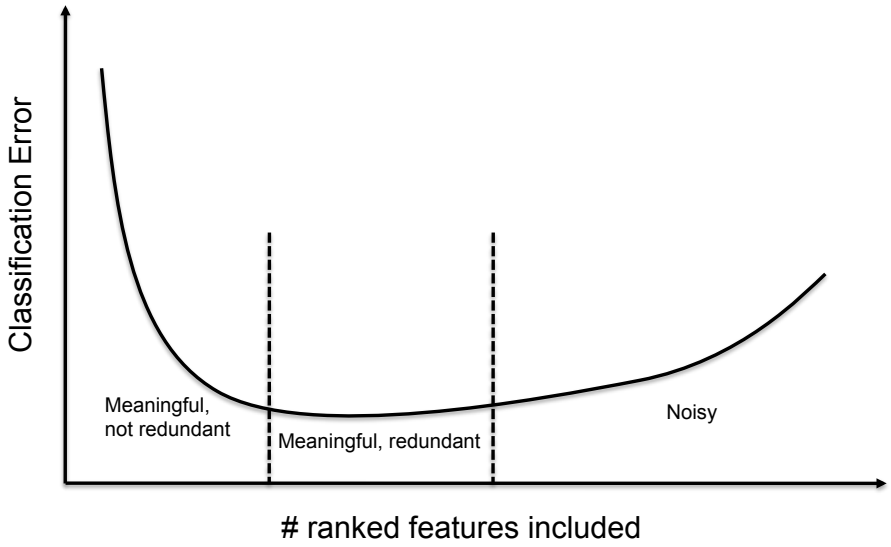


Figure 4.2: The figure shows a typical behavior of the classification error versus the size of the selected feature subset for a setting where the features are ranked according to their relevance to the classification task. (Adapted from [Tangermann, 2007])

4.2.2 Univariate Methods

To test the relevance of individual features simple univariate methods, such as the T-test, F-score etc., can be employed, and they usually perform quite well [Guyon et al., 2006]. A univariate ranking index for a binary classification problem use a test statistic to compare means or variances of two assumed gaussian processes. The T-statistic compares the means of two classes and the realization, t, of the statistic T is given by

$$t = \frac{\mu_A - \mu_B}{\frac{(m_A-1)s_A^2+(m_b-1)s_B^2}{m_A+m_B-2}\sqrt{\frac{1}{m_A}+\frac{1}{m_B}}}, \tag{4.3}$$

where m_A and m_B are the number of trials in each class, μ_A and μ_B are the means of each class, and s_A and s_B are the estimated standard deviation, [Guyon et al., 2006]. The absolute value of the T-statistic can be used directly as a ranking criterion, with the largest value corresponding to the most informative feature. Feature ranking based on such a correlation makes an implicit orthogonality assumption and do not take mutual information between features into

account. However, this assumption is rarely valid, especially not in high dimensional settings such as the presented EEG classification, where features are highly correlated.

4.2.3 Multivariate Embedded Methods

Multivariate embedded methods deal with the shortcomings of univariate methods. The term embedded refers to the relation between the feature selection method and the classification method. Embedded methods integrate the classification algorithm in the feature selection algorithm and this coupling leads to a good feature subset while the classifier is trained. Among the embedded methods used for feature selection, the greedy algorithms based on forward selection and backward elimination are the most popular. A popular embedded method is a backward elimination algorithm proposed by [Guyon et al., 2002]. The Recursive Feature Elimination, RFE, algorithm operates by trying to choose the subset of features, which leads to the largest margin of class separation using a measure given by the SVM classifier itself. The algorithm is mostly used with linear SVMs where the weights directly apply as a measure, but it can be generalized to the non-linear case. In the present thesis the non-linear version is used in a modified version. In the case of a linear SVM, the decision function is given by $f(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$, and the algorithm iteratively removes the feature with the smallest weight, $|w_i|$, and retrains the model until all features are ranked. In the non-linear case it iteratively removes the features leading to the smallest change in the SVM cost function, see Equation (3.29). Hence the RFE is a pruning method according to the smallest change in objective function. At the expense of sub-optimality, several features can be removed at every iteration. For high dimensional data it is common practice, see e.g. [Guyon et al., 2006], to remove chunks of the features at every iteration. The algorithm is described in Algorithm 1. Compared to the original algorithm, it is modified such that the σ and C parameter is re-estimated for every reduced subset of features, this simple improves the performance. Furthermore the full cost function is estimated using a modified output of LIBSVM since it normally does not return this value.

In the present implementation, see Algorithm 1, the number of features that are removed in every iteration of the algorithm is adjusted depending on the remaining number of features. Hence, during the first few iterations, large chunks of features are removed, whereas later in the process only few features are removed every iteration. This will be further discussed in Section 5.5.3. To select the optimal number of features, [Guyon et al., 2002] simply evaluate the accuracy directly on the test set for different sized subsets of genes in a gene selection task. Alternatively it could be based on the ranking criterion itself or by splitting the training set into two new sets, where one is used for validation and on

for training, and then pick the optimal number of features based on the predictive performance on the validation set. This procedure however, proves to be rather intractable and results in this thesis are therefore presented inspired by [Guyon et al., 2002].
The algorithm, see Algorithm 1, is implemented in Matlab and it turns out that it can benefit from the result in Section 3.6.3 and 5.5.2. During evaluation of the single features, the stopping tolerance can be adjusted to reduce computational time. To obtain the the margin based ranking criterion, LIBSVM [Chang and Lin, 2011] is modified to produce the desired values.

Algorithm 1 Modified Recursive Feature Elimination

Input:

Training examples
$\mathbf{X}_0 = [x_1; x_2; \ldots; x_l]$
Class labels
$\mathbf{Y}_0 = [y_1; y_2; \ldots; y_l]$

Initilalize:

Feature subset of all original features
$\mathbf{s} = [1, 2, ..., n]$
Empty subset of ranked features
$\mathbf{r} = [\,]$

while **s** non-empty **do**

Reduce training examples to surviving feature indices
$\mathbf{X} = \mathbf{X}_0(:, \mathbf{s})$

Estimate SVM parameters for the reduced set of features
$(\sigma, C) = crossval(\mathbf{X}, \mathbf{T}_0)$

Train the SVM classifier on the feature subspace defined by **s**
$\mathbf{W} = svmtrain(\mathbf{Y}_0, \mathbf{X})$ (i.e. $W = (\alpha^T - \frac{1}{2}\alpha^T y_i y_j K(x_i, x_j)\alpha)$)

For each feature $m \in \mathbf{s}$ compute the ranking score
for $m = 1 \rightarrow size(\mathbf{s})$ **do**
$W_{\setminus m} = svmtrain(\mathbf{Y}_0, \mathbf{X}_{\setminus m})$
end for
$\mathbf{C}(i) = abs(\mathbf{W}^2 - \mathbf{W}_{\setminus m}{}^2)$

Find the feature with the smallest ranking criterion
$f = argmin(\mathbf{C(i)})$

Update indices
$\mathbf{r} = [\mathbf{s}(f), \mathbf{r}]$
$\mathbf{s}$=$\mathbf{s}$(1:f-1,f+1:length($\mathbf{s}$))

%%% Modification to remove P % of the features at every iteration
Find least important features
$[\sim, \mathbf{f}] = \text{sort}(\mathbf{C}, 1, \text{'ascend'});$
Update the ranked feature list
$\mathbf{r} = [\text{fliplr}(\mathbf{s}(\mathbf{f}(1 : \text{ceil}(\text{size}(\mathbf{X}, 2)/P)))), \mathbf{r}];$
Remove the least important features
$\mathbf{s}(\mathbf{f}(1 : \text{ceil}(\text{size}(\mathbf{X}, 2)/P))) = [\,];$
%%%

end while

Output:

Ranked list of features **r**

Chapter 5

Model Selection

As mentioned earlier, the amount of examples available for training a single-subject classifier is rather limited since only a median of 17 examples (range 7 - 24) are available in each class (left/right). Therefore a resampling strategy is applied to obtain robust estimates of the performance of the classification algorithm. In doing this, the possibility of estimating sound bounds on the error diminishes, but for high-dimensional data sets with few examples this is generally not possible anyway, see [Guyon et al., 1998]. In the following a general description of high dimensional spaces in low sample size settings is given as this is the condition of the used data set. Then a cross-validation approach applicable to this is described, along with a way to test significance of the obtained results. Some small synthetic data sets are introduced to illustrate various aspects of parameter selection. Finally, measures to evaluate the performance are introduced.

5.1 The Curse of Dimensionality

The term "*curse of dimensionality*" is commonly used to describe challenges in high dimensional spaces, [Bishop, 2007]. Especially geometrical intuitions from two and three dimensions are not always valid in high dimensional spaces. [Hall et al., 2005] investigate general properties of high dimensional low sample

size data in a classification setting and present some rather interesting insights. They prove for a finite sample size setting that all points will lose subsequently more of their spatial topology if dimensionality is increased towards infinity. In the presented asymptotic case of infinity, the points will be pairwise orthogonal. Furthermore, they will be asymptotically located on the vertices of a regular simplex where all points have almost the same distances to the origin as well as among each other.
This distance concentration, where all distances in high dimensional space are almost equal, is just one aspect of the special characteristics posed by high dimensional problems. Low sample size high dimensional data additionally gives rise to the phenomenon known as hubness, see [Radovanović et al., 2010]. This phenomenon is related to the number of times a point occurs among the k nearest geometric neighbors. It is shown for a wide range of problems, see [Radovanović et al., 2010], that when dimensionality is increased, the distribution of the number of occurrences becomes considerably skewed. Few points appear to be the nearest geometric neighbor to other points more often. Nevertheless, the phenomena of hubs is related to the distance concentration. As shown by [Hall et al., 2005], the points in low sample size high dimensional settings are almost orthogonal with the same distance to each other and to the origin of the high dimensional simplex. If one point is an "outlier" in the sense that it is a little closer to the the simplex origin than the rest are, it is then also closer to several other points. Hence hubs are outliers in the sense that they are found in low density areas of the distribution, but are close to many other points. Normally outliers are thought of as further away from the center, this is not the case for hubs. These observations can influence classifier performance if not handled with care.

5.2 Cross-validation

A common approach, to avoid overfitting due to the resampling, is to use cross-validation. This is a technique where subsets of data are held out and used for validation, while the model is trained on the remaining data [Bishop, 2007]. This procedure is repeated and the quality of the predictions across the test sets are averaged to yield an overall measure of the predictive power, i.e. the test error. The mean accuracy of the single SVM estimated on the hold-out set is an unbiased estimator of the mean. Furthermore, in a resampling setting, the mean of several unbiased estimates also produces a new unbiased estimate of the mean. The exact strategy for determining the size of the training- and test-subset may vary. One form of cross-validation leaves out a single observation at a time, whereas K-fold cross-validation splits the data into K subsets, which are each used for validation. Both the kernel- and the regularization parameter

are chosen using this strategy within the training set.
In a two class classification task it is common practice to use a leave-two-out procedure, where one example from each class is left out of the training set. In the presented data set the leave-two-out procedure is applied on balanced data sets, i.e. a number of trials are left completely out at every iteration to obtain an equal number of examples in each class. This is done to avoid that the trained classifier just behaves as "majority" classifier. In continuation of the general properties presented for high dimensional spaces, [Hall et al., 2005] show that a basic SVM will predict based on the majority class alone for a dimension towards infinity if the inter-distance between two classes is too small. Furthermore, for unbalanced data, it is shown that the SVM gives asymptotically completely incorrect classification for the population with the smaller sample. Inspired by these results [Klement et al., 2008] show that infinity is not that large in practice, especially not for the soft margin SVM. This means that a leave-one-out procedure will misclassify all examples even for rather low dimensional data if only a small number of examples are available for training.

5.3 Permutation Test

While the mean of the errors obtained via cross-validation is indeed an unbiased estimate of the expected error, the variance is not. The cross-validation trials are not independent and hence it requires thorough modeling of the dependence to avoid a too optimistic estimate of the variance. An alternative is to use a permutation test to estimate how likely it is to obtain the result by chance, due to some random pattern detected by the SVM in the high dimensional data [Golland and Fischl, 2003, Golland et al., 2005, Efron and Tibshirani, 1993]. The result provided by the permutation test provides a weaker answer than standard convergence bounds, since it gives no indication of how well the obtained error rate will generalize. However, it does answer whether the classification result could be obtained by chance.
The null hypothesis of the permutation test is that the SVM cannot learn to predict labels based on the given training set and works under the assumption that the data distribution is adequately represented by the sample data.
For a set of examples, $\{\mathbf{x}_i, y_i\}_{i=1}^{m}$, with all the possible permutations, $\mathcal{Z}_m$ for indices $1...m$, and with the test statistic $\mathcal{T}$ based on the cross-validation error, the permutation test procedure consists of N iterations, see Algorithm 2.
Ideally the full set of permutations should be used to generate the cumulative distribution, $\hat{P}$, but this is not computational tractable, and hence the algorithm relies on sampling from $\mathcal{Z}_m$. This strategy is feasible as long as N is sufficiently large.

Algorithm 2 Permutation Test

for $n = 1 \rightarrow N$ **do**

Sample a permutaion $\mathbf{z}^n = (z_1^n, ..., z_m^n)$ from a uniform distribution over $\mathcal{Z}_m$

Compute the statistic $t^n = \mathcal{T}(\mathbf{x}_1, y_{z_1^n}, ..., \mathbf{x}_m, y_{z_m^n})$

end for

Construct an empirical cumulative distribution

$$\hat{P}(T \leq t) = \frac{1}{N} \sum_{n=1}^{N} \Theta(t - t^m)$$

where Θ is a step function $(\Theta(x) = 1, if x > 0; 0\ o.w.)$

Compute the statistic for the true labels, $t_0 = \mathcal{T}(\mathbf{x}, y, ..., \mathbf{x}_m, y_m)$ and the corresponding p-value p_0 under the empirical distribution $\hat{P}$

Reject the null hypothesis if $p_0 \leq \alpha$, where α is the acceptable significance level

5.4 Synthetic Data

In addition to the main analysis of the presented EEG data some synthetic data sets are created to investigate the performance of the classifier and to illustrate e.g. kernel behavior.
The Exclusive OR, XOR, problem is a classical problem used to illustrate challenges in classification problems. Examples of the XOR problem is the chessboard problem seen in Figure 5.4 and the intertwined circles seen in Figure 5.1. It is clear that neither the input variable x_1 nor x_2 is able to perform the classification independently whereas patterns or regularities in the data can be found if the combination of x_1 and x_2 is used. If the input data is composed of additional noisy features the shown regularity might not be found through classical statistical analysis. In that case more advanced methods can be employed to retrieve the two relevant features thus also simplifying the classification task.
Furthermore the chessboard like patterns require highly non-linear decision boundaries, and hence is a good problem to test non-linear classifiers on.
Likewise, the problem composed of two intertwined semi-circles, see Figure 5.1, is a good XOR problem for illustrating kernel properties.
Finally two high dimensional problems are constructed. An XOR data set with

a more smooth decision boundary is constructed, namely an n-dimensional ball inside an n-dimensional spherical shell - the two having an overlap of 20 % of the data points. For $n = 2$ this corresponds to the circle and annulus shown in Figure 5.2. In addition to this a very high dimensional problem where all features are drawn from two standard normal distributions with different mean is constructed.

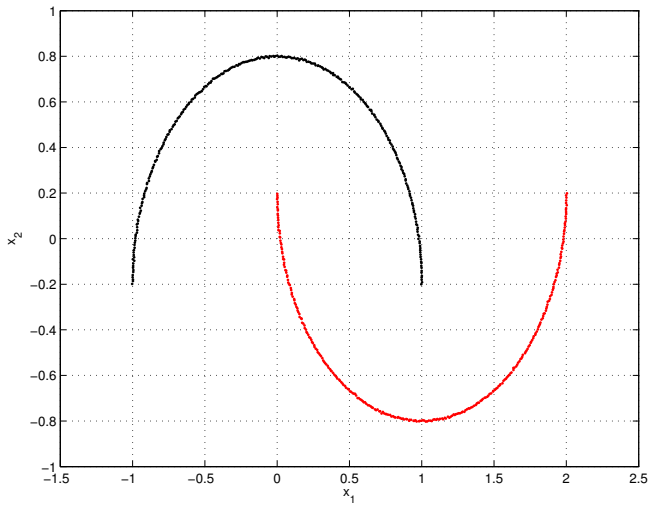


Figure 5.1: An example of the XOR problem.

5.5 Parameter Selection

Both the SVM, SMO and the RFE requires a parameter selection. In the following the structure of the data is used to determine the SVM parameter determining the non-linearity and it is investigated how computational effort can be reduced for the RFE and SMO without jeopardizing accuracy.

5.5.1 SVM Parameters

Generally there is consensus that the RBF kernel is among the best performing kernels [Schölkopf and Smola, 2002]. Actually, [Caputo et al., 2002] find in a

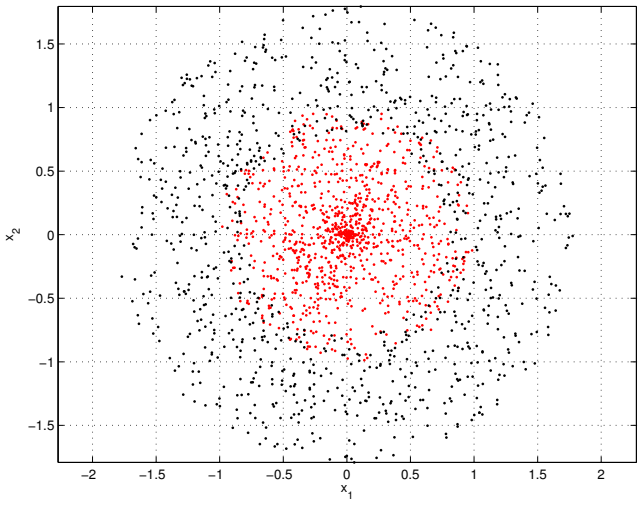


Figure 5.2: A 2-dimensional example of the n-dimensional ball overlapped by an n-dimensional spherical shell.

comparison of standard kernel functions for a wide range of different features, that when the RBF kernel is not the best performing kernel, the error rate is not higher than 1% with respect to the best performing kernel. In many applications of the RBF kernel, the parameters are picked using a coarse search over a very wide range of values [Chang and Lin, 2011] without paying attention to the structure of the problem. [Caputo et al., 2002] propose a heuristic to select the kernel parameter, σ, which is independent of feature type. They find that the best σ-value can be found by searching the interval from the 0.1 - 0.9 quantile of $||x_i - x_j|| = 2\sigma^2$. Figure 5.3 shows an image of the kernel matrix for different σ's in the proposed interval for the XOR problem shown in Figure 5.1. It can be seen that for σ's, corresponding to the 0.1 quantile, only few elements (neighboring points) are non-zero and hence a very sparse representation is obtained, whereas for σs corresponding to the 0.9 quantile, all entries are non-zero. Figure 5.4 furthermore shows how the RBF kernel creates centers, see Figure 5.4b, for each of the squares in the XOR checkerboard problem, see Figure 5.4a.

The regularization parameter, C, is less intuitive and is usually picked from a coarse search over a wide range of values.

Underfitting, where the entire set is assigned to the majority class, generally occurs when σ is fixed and $C \to 0$, when C is fixed sufficiently small and $\sigma \to 0$ and when $\sigma \to \infty$ and C is fixed. Overfitting is seen when small regions around

the training examples of the smallest class are classified to be that class, while the rest is classified as the majority class. This occurs in the case where $\sigma \to 0$ and C is sufficiently large and when σ is fixed and $C \to \infty$ for noisy data since the SVM classifier strictly separates the training examples of the two classes. The linear kernel matrix is shown in Figure 5.3d. Though there is little visual resemblance between the images, it can be shown that for $\sigma \to \infty$ and proper scaling of C the non-linear SVM classifier with RBF kernel converges to the linear SVM classifier, [Keerthi and Lin, 2003].

Theorem 5.1 *For a proper fixed value of $C = \tilde{C}\sigma^2$ and $\sigma \to \infty$ the non-linear SVM with a Gaussian RBF kernel converges to the linear SVM.*

Proof. For $\sigma \to \infty$ the Gaussian RBF kernel function can be written, using a series expansion, as

$$K(x,x) = \exp\left(-\frac{\|x-x\|^2}{2\sigma^2}\right) \tag{5.1}$$

$$= 1 - \frac{\|x-x\|^2}{2\sigma^2} + O\left(\frac{\|x-x\|^2}{\sigma^2}\right) \tag{5.2}$$

$$= 1 - \frac{\|x\|^2}{2\sigma^2} - \frac{\|x\|^2}{2\sigma^2} - \frac{x^T x}{\sigma^2} + O\left(\frac{\|x-x\|^2}{\sigma^2}\right). \tag{5.3}$$

Using this approximation, the quadratic term of the objective function 3.29 in the non-linear version of the SVM can be written as

$$\begin{aligned}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x,x) = \\ \sum_i \sum_j \alpha_i \alpha_j y_i y_j \\ - \frac{\sum_i \sum_j \alpha_i \alpha_j y_i y_j \|x_i\|^2}{2\sigma^2} \\ - \frac{\sum_i \sum_j \alpha_i \alpha_j y_i y_j \|x_j\|^2}{2\sigma^2} \\ + \frac{\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j}{\sigma^2} \\ + \frac{\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \Delta_{ij}}{\sigma^2},\end{aligned} \tag{5.4}$$

where $\lim_{\sigma^2 \to \infty} \Delta_{ij} = 0$.
By defining $\tilde{\alpha}_i = \frac{\alpha_i}{\sigma^2}$ and using the constraint from the QP-formulation in (3.29) ($\sum_{i=1}^m \alpha_i y_i = 0$) in the three first terms of the approximation in (5.4), the

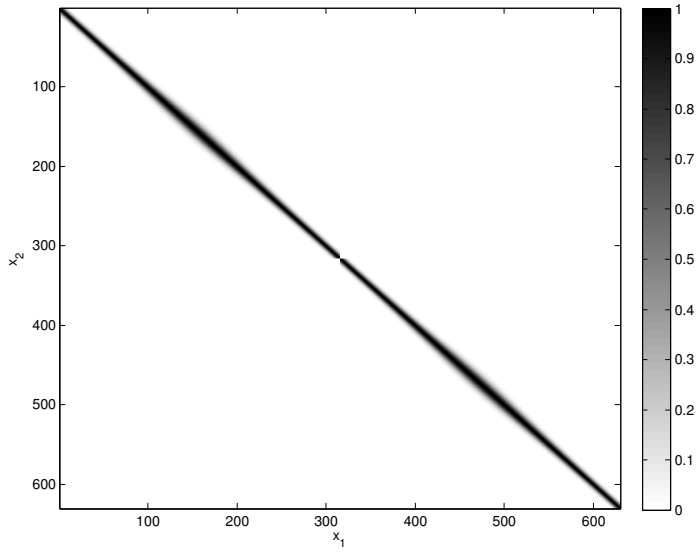
formulation in (3.29) can, if the objective function is divided by σ^2, be written as

$$\begin{aligned} \min_{\tilde{\alpha}} \quad & \frac{F}{\sigma^2} = \frac{1}{2}\sum_{i,j=1}^{m} \tilde{\alpha}_i \tilde{\alpha}_j y_i y_j \tilde{K}_{ij} - \sum_{i=1}^{m} \alpha_i \\ \text{subject to:} \quad & 0 \leq \tilde{\alpha}_i \leq \tilde{C}, \quad i = 1, ..., m \\ & y^T \tilde{\alpha} = 0, \end{aligned}$$

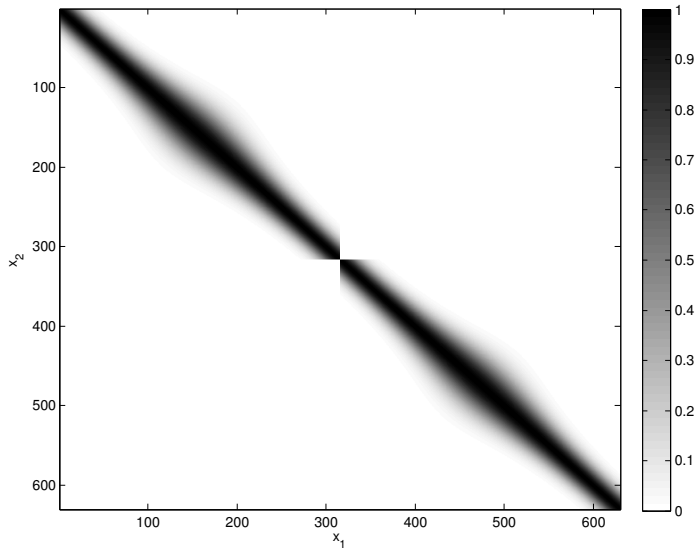
where $\tilde{C} = \frac{C}{\sigma^2}$ and $\tilde{K}_{ij} = x_i^T x_j + \Delta_{ij}$ which is equivalent to the linear kernel since $\lim_{\sigma^2 \to \infty} \Delta_{ij} = 0$, hence there is no need, other than computational effort, to consider the linear SVM.
If C is just fixed, and hence not varied with σ^2, the classifier will, as described earlier, underfit severely.

5.5.2 SMO Parameter

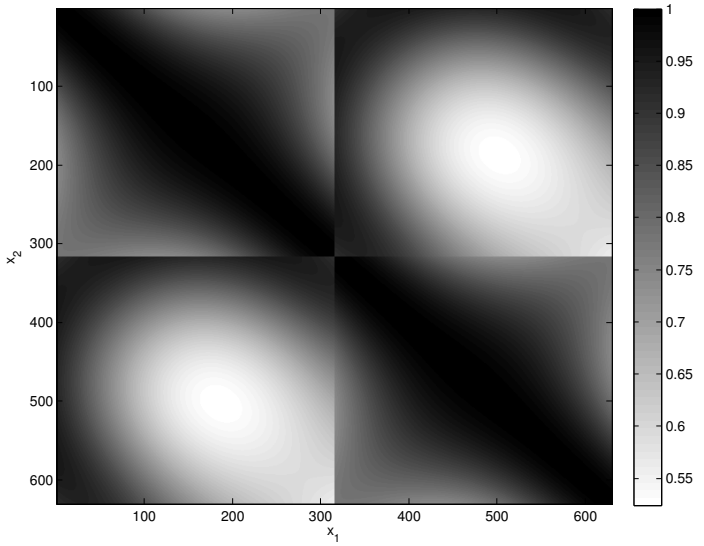
The influence of the optimization stopping tolerance, ϵ, is investigated on the synthetic data sets. As seen in Figure 5.5 - 5.7, and as stated in [Platt, 1998], the tolerance for fulfilling the KKT conditions and hence stopping the optimization need not to be very fine. The accuracy remains unchanged for tolerances up to $10^{-1} - 10^{0}$.
For sparse kernel matrices the number of support vectors changes considerably more than for less sparse kernel matrices, i.e. if σ is chosen around the 0.9 quantile, cf. Section 5.5.1, the number of support vectors does not depend too much on ϵ whereas if σ is chosen around the 0.1 quantile, the number of support vectors varies with ϵ. The C value affects the total computational time and harder regularization generally leads to longer processing time and in some cases reduces the number of support vectors considerably for coarse stopping tolerances, see Figure 5.7a - Figure 5.7b. The chessboard problem requires a higher degree of regularization for values of σ around the 0.9 quantile, this increase the computational time.
Regarding the examples assigned as support vectors in each data set, it seems reasonable to choose $\epsilon = 10^{-3}$ as proposed by Platt. As seen, the accuracy however, does not change significantly even if some borderline vectors are either left out or included as support vectors, whereas the computational time is reduced rather notably. Especially for the very non-linear decision boundaries, it can be seen that a coarse tolerance leads to a noteworthy reduction of computational time. This result comes in handy in the implementation of Algorithm 1, Section 4.2.3.



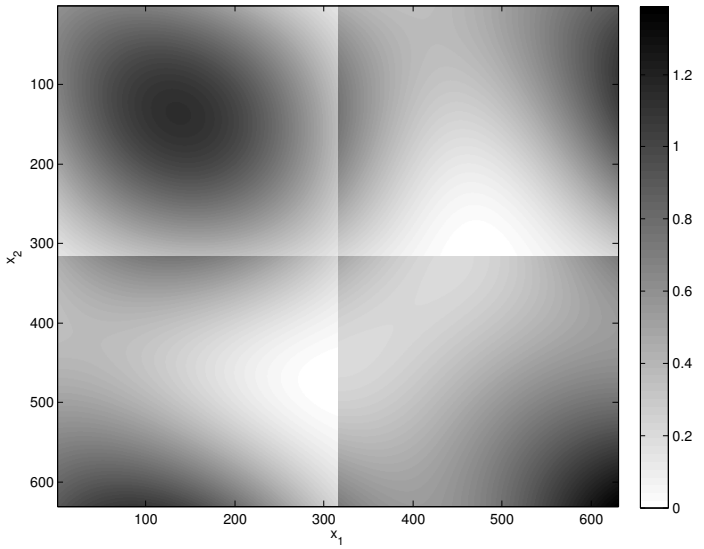
(a) 0.1 quantile.



(b) 0.5 quantile.

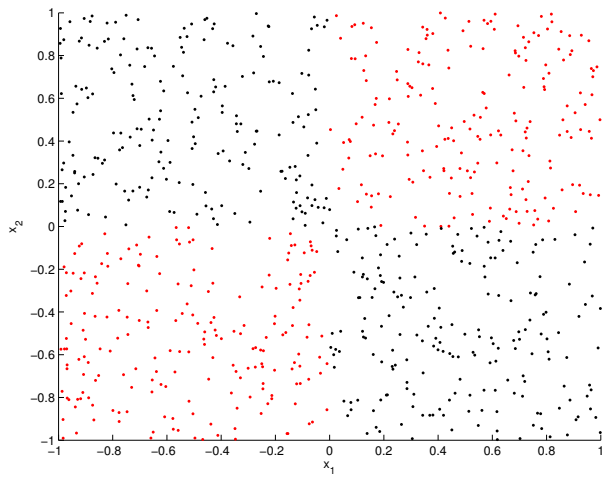


(c) 0.9 quantile.

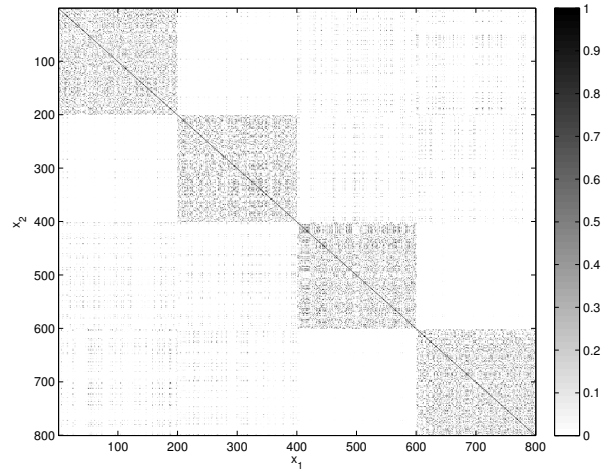


(d) Linear kernel.

Figure 5.3: Selection of kernel parameter, σ, based on different quantiles of the scaled width $||x_i - x_j||$ of the intertwined semi-circles. The figure additionally illustrates that the RBF kernel (a)-(c) is translational invariant, whereas the linear (d) is not. For the RBF kernel especially the clear diagonals should be noticed.

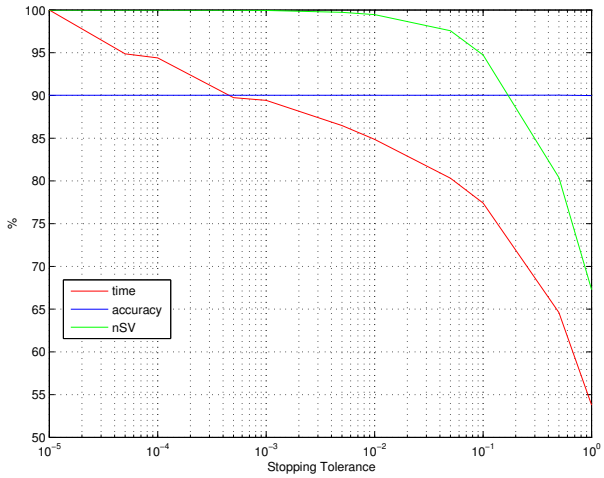


(a) 2 x 2 chessboard XOR problem.

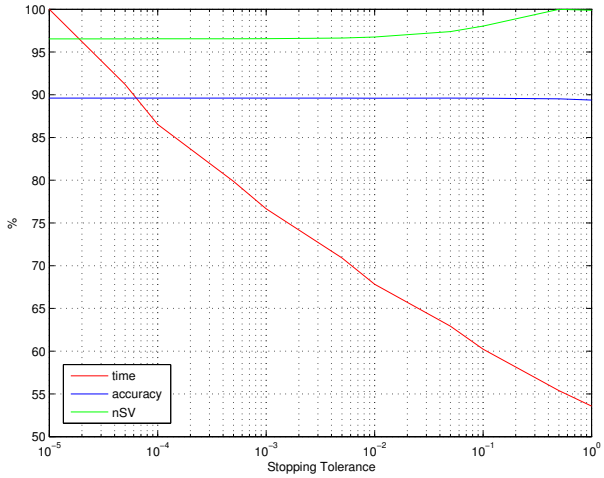


(b) Corresponding kernel for σ based on the 0.1 quantile.

Figure 5.4: The image of the kernel matrix for the 2 x 2 chessboard XOR problem shows how the RBF kernel creates a center for each of the squares.

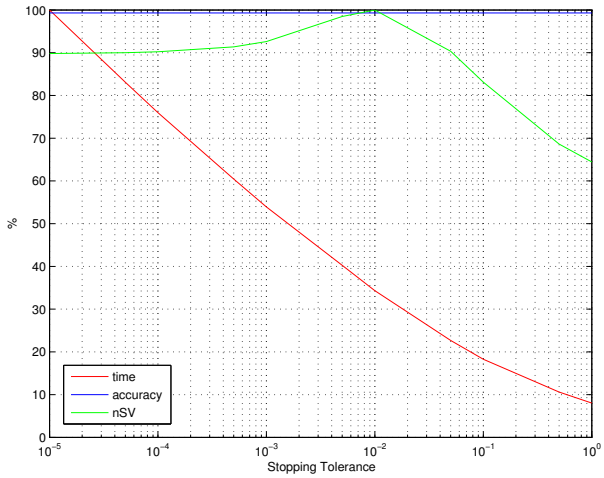


(a)

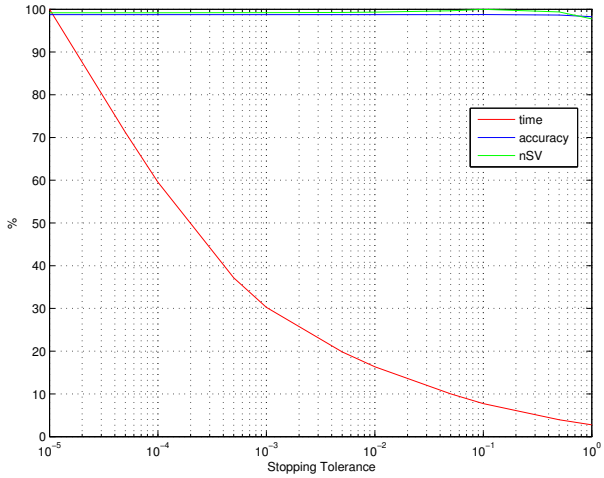


(b)

Figure 5.5: 50-dimensional ball and spherical shell with 1,000 examples in each class for both the test and training set. In (a) σ is chosen as the 0.1 quantile and $C = 10^2$ and in (b) σ is chosen as the 0.9 quantile. The results are averaged over 100 data sets. The two classes are overlapping, hence around 90% classification accuracy is expected.Time and nSV (number of support vectors) are divided by their maximum values.

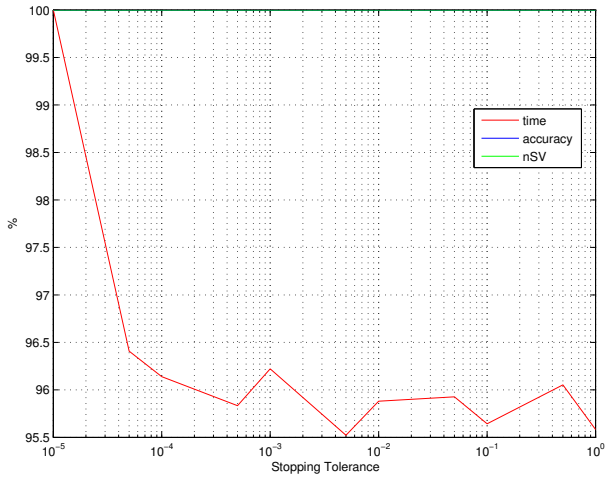


(a)

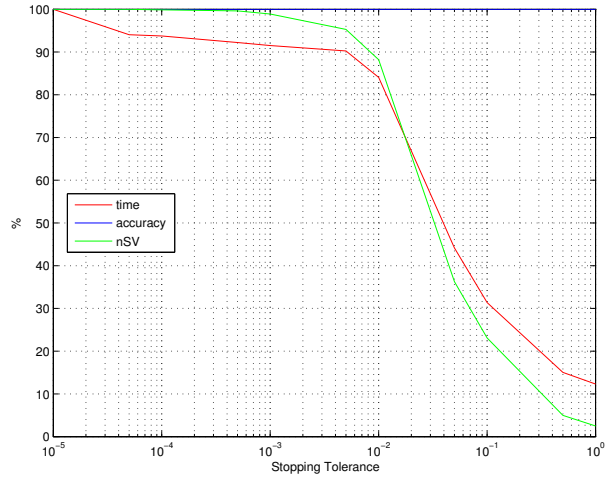


(b)

Figure 5.6: 3 x 3 chessboard problem with 1,000 examples in each square and no overlapping classes. In (a) σ is chosen as the 0.1 quantile and $C = 10^2$ and in (b) σ is chosen as the 0.9 quantile. No overlapping classes, hence 100 % classification accuracy is expected. The results are averaged over 100 data sets. Time and nSV (number of support vectors) are divided by their maximum values.



(a)



(b)

Figure 5.7: 10,000 dimensional examples with all features drawn from normal distributions. One class has mean +1 for all features, the other class has mean -1 for all features. σ is chosen as the 0.9 quantile and $C = 1$ and $C = 10^8$ in (a) and (b) respectively. The results are averaged over 10 data sets. Time and nSV (number of support vectors) are divided by their maximum values. In (a) both nSV and accuracy is 100 and in (b) accuracy is 100.

5.5.3 RFE Parameters

The implementation of the RFE algorithm described in Section 4.2.3 removes 25 % of the features during each iteration while the feature dimension is over 1,000, and removes 10 % during each iteration when a feature dimension below 1,000 is obtained. This approach gives very comparable results to just removing 5 % of the features during all iterations, but reduces the time considerably.
The SVM parameters are re-estimated during every iteration as this improves results significantly. Especially re-estimating the σ-parameter is important as the problem changes size dramatically during the iterations. However, to reduce computational time, only a very coarse grid is used to find values for C and σ. In the final implementation C is found searching a log2 interval and σ is found using a coarse grid from the 0.1 - 0.9 quantile of $||x_i - x_j|| = 2\sigma^2$. Using a very fine grid gives comparable performance for all subjects. The most important parameter of the two proves to be σ and since a rather "narrow" feasible range is given from the inter- and intra class distance, this result is not surprising.
Furthermore a value of $\epsilon = 10^{-2}$ is used as stopping criterion in the RFE algorithm. When evaluating final performance $\epsilon = 10^{-3}$ is used.

5.6 Evaluation of Model Performance

The SVM classifier is evaluated in terms of achieved prediction error (one minus the number of correct predictions divided by the total number of predictions). Additionally the permutation test is conducted corresponding to a 95% significance level. In addition to the permutation test described in Section 5.3 the classifier is also evaluated on the data presented in the awake condition and on data obtained solely over the visual cortex (electrodes O1 and O2). The data from the awake condition is expected to contain enough information for an actual classification in most cases and the data obtained over the visual cortex is expected to show less or no predictive power.

Chapter 6

Results

The following chapter analyses the EEG data obtained from the experiment described in Section 2.3. The EEG data is pre-processed as described in the next section and then analyzed using the setup described in the previous chapters. The SVM described in Chapter 3 is combined with the two feature selection schemes described in Chapter 4. All parameters are tuned as described in Section 5.5. Performance is assessed using the balanced leave-two-out cross-validation scheme described in Section 5.2 and significance is evaluated using the permutation test described in Section 5.3. Everything is implemented in Matlab. The results are reported in this chapter and discussed in the following chapter.

A general overview of the subjects included in the analysis is shown in Table 6. The table shows the number of trials obtained during wakefulness and during sleep. As can be seen, the number of trials included for the different subjects to obtain a balanced training and test set varies between 14 and 46.

6.1 Data Pre-processing

The raw EEG data are pre-processed using a Matlab toolbox developed by Sid Kouider, École Normale Supérieure, Leonardo da Silva Barbosa, École Normale

Subject	Sleep EL	Sleep ER	Balanced no. of trials	Awake L	Awake R	Balanced no. of trials
105	20	22	40	21	15	30
107	17	16	32	20	22	40
109	16	20	32	22	24	44
111	17	7	14	16	20	32
113	20	17	34	21	21	42
117	9	13	18	12	13	24
118	20	17	34	24	24	48
122	17	16	32	15	15	30
127	18	17	34	22	21	42
129	10	10	20	18	19	36
134	21	24	42	18	21	36
137	21	19	38	15	16	30
138	10	9	18	18	19	36
139	10	17	20	21	25	42
144	16	19	32	15	13	26
147	21	22	42	15	15	30
149	23	20	40	23	24	46
150	23	24	46	23	20	40

Table 6.1: Overview of subjects and corresponding number of trials. EL, expected left, and ER, expected right, corresponds to the L, left, and R, right in the awake case and indicates whether the expected response for the subject is left or right (object vs. animal).

Supérieure, and Carsten Stahlhut, Technical University of Denmark. The toolbox is mainly based on other open source toolboxes for Matlab, namely SPM8 [Ashburner et al., 2008], FieldTrip [Oostenveld et al., 2011] and EEGLAB [Delorme and Makeig, 2004].
The pre-processing steps include high-pass filtering, low-pass filtering, epoching, baseline correction and downsampling of the raw data.
In addition to this all features are standardized, as recommended when dealing with SVMs [Chang and Lin, 2011].
The spectral transforms are found using the SPM8 interface to FieldTrip.
Along with the general pre-processing, a more sparse representation of the EEG data is found. This representation is more computational tractable and improves the predictive power of the classifier.

6.1.1 Data Set Dimensionality Reduction

A spectral transform of the EEG data leads to a very high feature dimension. The dimension is a function of the number of channels × the frequency resolution × the time resolution. For a standard wavelet transform of the 64 channel signal with a frequency resolution of 1 Hz, this leads to a feature space with a dimension greater than 10^6 for each trail.
It is of course a trade off between tractability and information resolution when the dimensionality is reduced. However, it proves that reducing the spatiotemporal resolution, i.e. binning frequencies into coarser intervals and reducing the time resolution, not only leads to a reduction in computational time but also improves the predictive power of the SVM. This is not surprising both from a physical and data dimension point of view. Since only a few trials in each condition are present, the physiological variability between trials will lead to too much difference between relevant features. If the time and frequency resolution is too fine and only a few features can be selected by the feature selection algorithm to reduce noise, there will simply not be an overlap of features between the different trials. Additionally, when the resolution is reduced, the power of informative features will increase. On the other hand if the resolution is reduced too much, the informative features will lose information due to too much "averaging" with the neighboring features. As proposed by e.g. [Pfurtscheller et al., 2006], the most important electrodes are in the motor cortex area around C3 and C4, hence a considerable sparse prior to apply, is that only those two electrodes are included in the analysis. The predominant frequencies are expected to lie in the interval 4-40 Hz as described in Section 2.2, and hence the frequency interval from 4 - 40 Hz is used with steps of 4 Hz. The original time resolution is reduced to 50 Hz. This gives a more manageable dimension of the feature space, namely 2,800.
One (sub)optimal configuration is to use a 7 cycle wavelet transform with frequency intervals of 4 Hz in the range 4 - 40 Hz and a reduce the number of time samples, using the subsample option = 10. If nothing else stated, results are obtained using this representation, which is denoted the reduced data set in the following.

6.1.2 Artifact Rejection

An artifact rejected data set is constructed to investigate possible improvements, see Figure 4.1. The artifacts found in EEG data sets recorded during sleep are significantly different from the ones found in a normal EEG signal. Especially some auditory components and K-complexes have been removed in the artifact

rejected data set using Independent Component Analysis, ICA. In the awake condition as well as during sleep, auditory components were removed from virtually all trials, whereas the K-complexes are only present during sleep. Only during wakefulness eye movements are removed, since these disappear during early sleep phases.

6.2 Feature Selection

Both the reduced data set described in 6.1.1 and the full wavelet transform leads to classification accuracies for all subjects around 50 % if no feature selection is performed. Hence the results are not described in further detail. In the following these results obtained using feature selection are presented.

6.3 T-Test Based Feature Selection

Using the T-test described in Section 4.2.2 generally leads to a poor classifier performance around chance level for the full spectral transform. On the reduced data set, the picture changes and it is possible to obtain significant results for at least 11 subjects though in general performance is worse or at best comparable to using the RFE approach, hence the results are not described in further detail. A comparison of the two methods can be seen in Figure 6.1. It shows the obtained error rates averaged over all 18 subjects. A major advantage of the t-test based feature selection is that the computational burden is significantly reduced compared to the RFE. This feature selection approach also leads to anti-learning, which is classification performance consistently worse than chance, in a few subjects, see discussion in Section 7.2 for further details.

6.4 RFE Feature Selection

Using the RFE based features selection requires more computational effort but leads to compelling performance. Results for all subjects are presented in Figure 6.3 - 6.6 and in Appendix A. For a majority of the subjects classification better than chance is possible. The corresponding features selected to obtain the reported results are harder to interpret. There is no single conclusion readily available, but classification performance combined with inspection of the feature selection yields several insights to extract. For each subject a figure showing

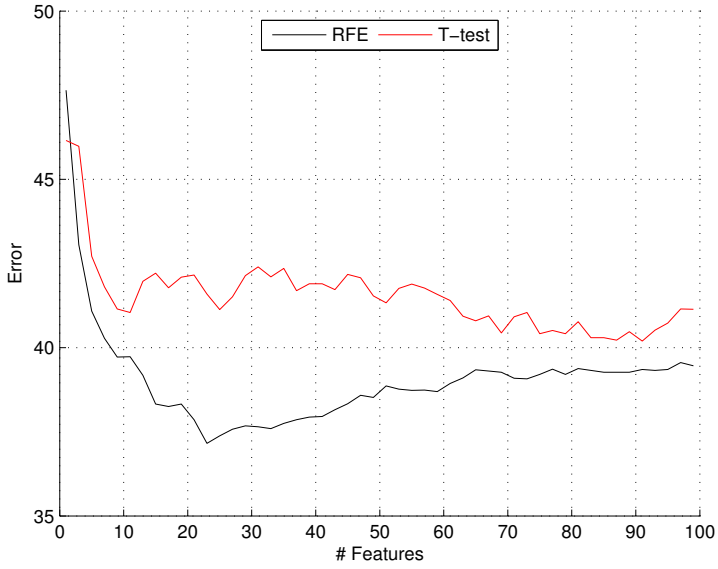


Figure 6.1: Group averaged results for the 18 subjects included in the analysis for the two different feature selection methods described.

error rates vs. the number of features included in the classification model is shown. Additionally two figures showing the features selected from electrode C3 and C4 is shown. The results obtained using the RFE algorithm are examined in the following.

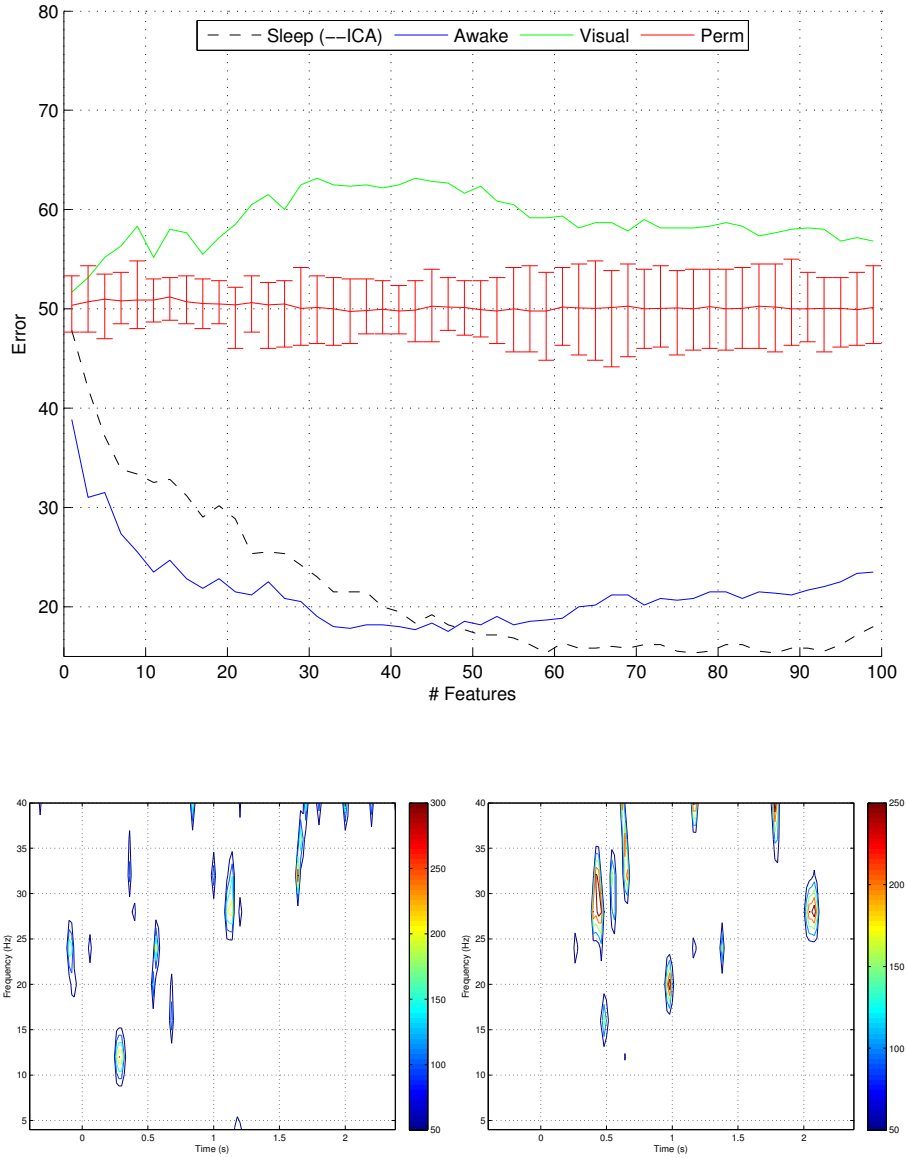
Generally the multitaper transform leads to slightly worse performance than the wavelet transform, see also Chapter 7, and these results are hence not reported in detail. Figure 6.2 shows the group averaged predictive performance for the multitaper and wavelet transform respectively using the RFE algorithm. As seen the overall error is lower using the wavelet transform.

Furthermore, for most subjects the classification error deteriorate if the frequency range is reduced, i.e. including gamma range activity improves the predictive power.

Figure 6.2: Group averaged results for the 18 subjects included in the analysis using the wavelet- and the multitaper transform. The overall picture as well as at a single-subject level, is that a wavelet transform leads to lower error rates than a multitaper transform.

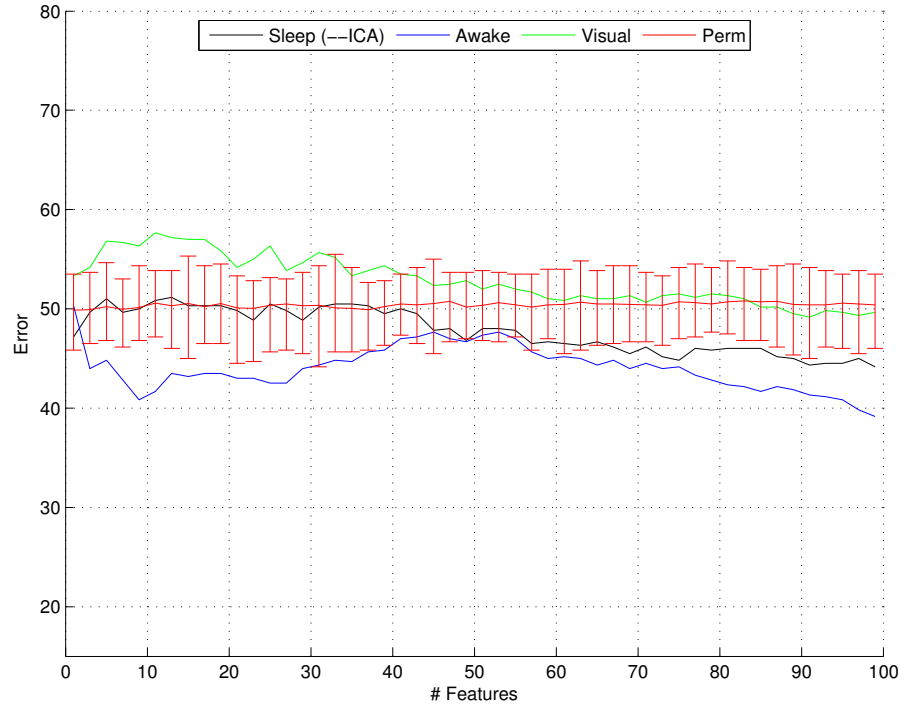
6.4.1 Permutation Test

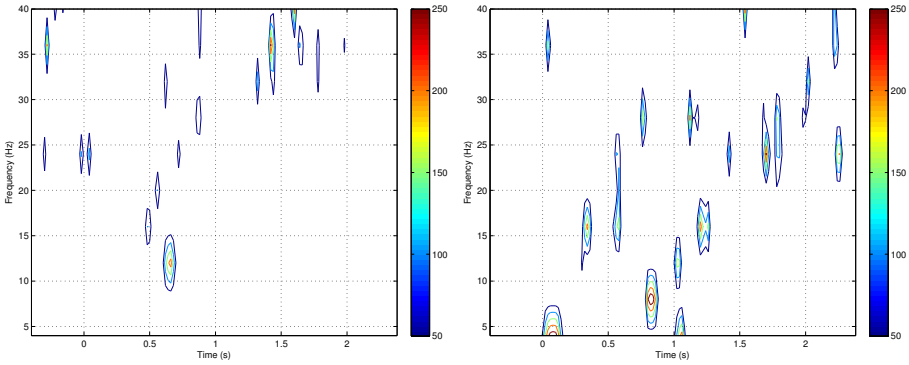
The result of the permutation test is shown as a red band in Figure 6.3 - 6.6 and in Appendix A, indicating a range of the performance for a classifier trained on permuted labels. Generally the classifier trained on data with permuted labels obtains error rates in the range ∼45% - 55% regardless the number of features included. Hence for at result to be significant it should be outside this range. For the permutation test described in Section 5.3, 20 iterations appear to be enough to obtain a significance level of 95 %. However, in addition to this, [Efron and Tibshirani, 1993] state that the estimate of the achieved significance level is affected by the Monte Carlo error. To reduce the influence of the variation to less than 10 %, a conservative estimate of 1,901 permutations are required to obtain an achieved significance level of 95 %. For subject 129 the full 1,901 iterations are carried out and showed together with the result of 20 iterations, see Figure 6.5. As can be seen, the variation is not decisive when testing the hypothesis that labels are exchangeable, and hence $N = 20$ is used for the



(a) Electrode C3 **(b)** Electrode C4

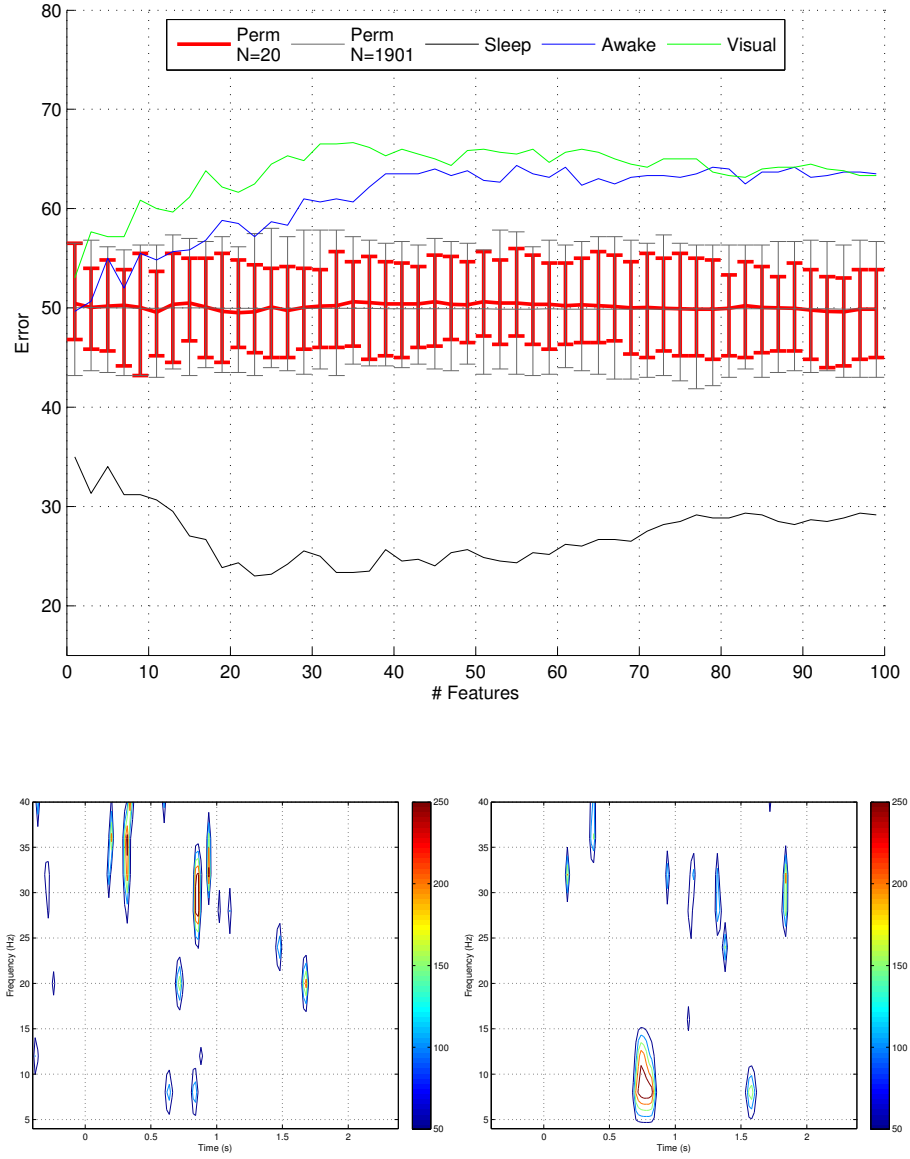
Figure 6.3: Subject 117. Top 50 time-frequency features selected during each leave-two-out iteration for subject 117. The number of times a feature is selected gives an indication of how important the single features are.





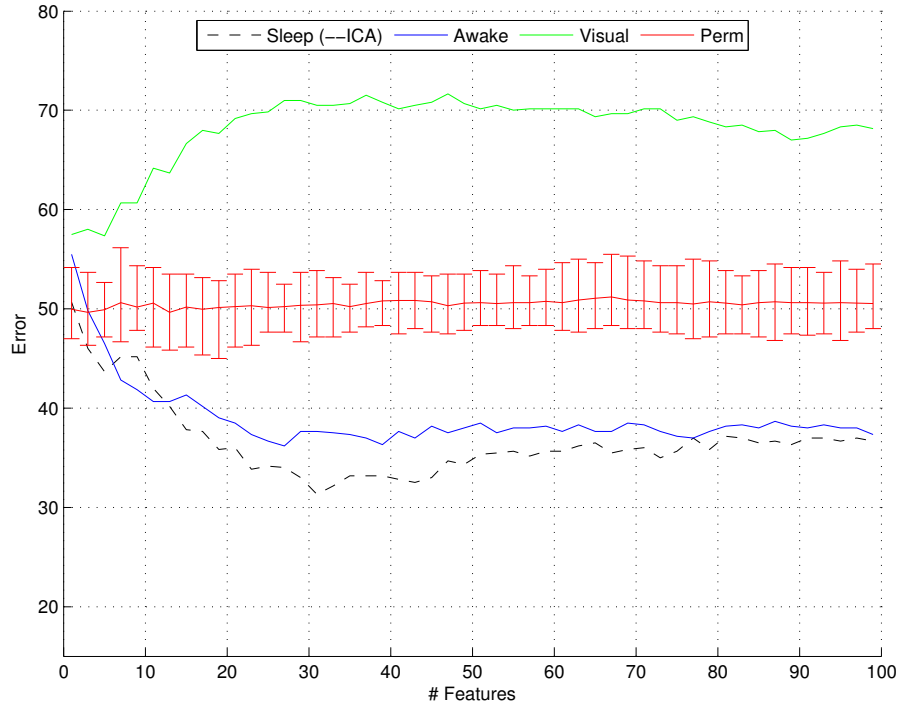
(a) Electrode C3 **(b)** Electrode C4

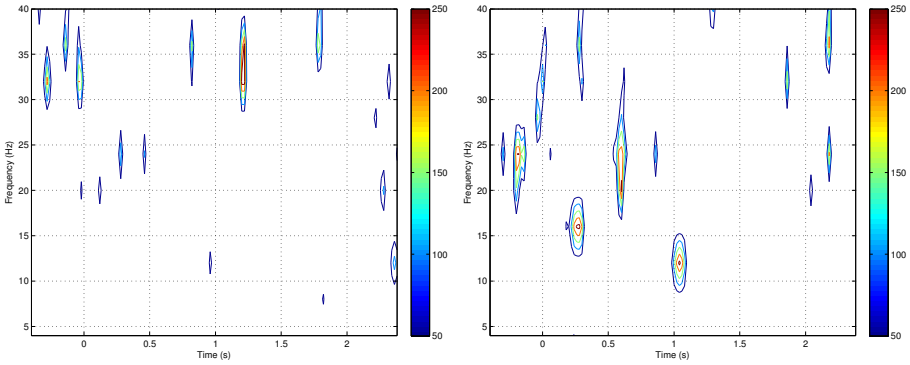
Figure 6.4: Subject 118. Top 50 time-frequency features selected during each leave-two-out iteration for subject 118. The number of times a feature is selected gives an indication of how important the single features are.



(a) Electrode C3 **(b)** Electrode C4

Figure 6.5: Subject 129. Top 50 time-frequency features selected during each leave-two-out iteration for subject 129. The number of times a feature is selected gives an indication of how important the single features are. In addition to the permutation test for N=20, the result of a test with N=1,901 is shown.





(a) Electrode C3 **(b)** Electrode C4

Figure 6.6: Subject 139. Top 50 time-frequency features selected during each leave-two-out iteration for subject 139. The number of times a feature is selected gives an indication of how important the single features are.

remaining subjects.

6.4.2 Data From Pre-motor Cortex

Using data from electrodes C3 and C4 recorded during sleep, yields a classifier performance significantly better than chance for 12 out of the 18 subjects, using either the artifact (subject 105, 113, 117, 134, 139, 144, and 150) or non-artifact (subject 107, 111, 122, 129, and 138) rejected data, see Figure 6.3 - 6.6 and Figure A.1 - A.14 in Appendix A. Some subjects get a much better signal, by removing artifacts, thus the artifact rejected data results are reported for these subjects (labelled ICA). For others the manual artifact rejection does not change the performance of the classifier. In some cases it removes to much of the signal, leading to worse classification performance. For these subjects, non-artifact rejected results are reported.
Five subjects (subject 109, 118, 127, 137, and 147) are borderline cases regarding classification performance, since only small parts of the feature subsets produce significant classification results or classification error rates are close to chance level. In subject 149, see Figure A.13, anti learning behavior is seen. This behavior is discussed in Section 7.2.
Except for the cases where only few features are included in the model, the training error is generally 0 and most trials are treated as support vectors, though not all. Looking at classifier performance shows that a certain number of features are needed to obtain the best classification, but including too many features leads to a decrease in performance, as depicted in Figure 4.2. The plateau where including features neither improves nor worsen performance varies in size for the subjects. For subject 111 where a very low error rate is obtained, the plateau is around 100 features, which is not obvious from the shown plot, see Figure A.4. The best single-subject classification accuracy is exceeding 80%, see Figure 6.3, and several subjects yield accuracies around 70% - 80%. A few subjects, see e.g. Figure 6.4, have no or very little predictive signal. As stated earlier the accuracies are unbiased estimates, but it is not possible to obtain reliable error bars. However, those results not conflicting with the permutation interval (red bars) are significant at a 95% level.

6.4.3 Data From Visual Cortex

Data recorded from electrodes O1/O2 during sleep is used as a proxy for visual cortex. It is expected that data collected from the visual cortex should not lead to good classification. However, it is not entirely the case, see Figure 6.3 - 6.6 and Figure A.1 - A.14 in Appendix A. For subjects 137, see Figure

A.9, and 149, see Figure A.13, the classifier yields results better than chance level. More puzzling are the results obtained for half of the subjects (107, 111, 113, 117, 129, 134, 139, and 150). Here the classifier produces results worse than chance, implicating that flipping all labels would yield a classifier performing significantly better than chance. Running a permutation test on the data collected from O1/O2 yields chance level results as expected. The number of predicted labels in each class are additionally relatively balanced, as are the error rates for the two classes. This anti-learning behavior, where the classifier consistently assign opposite class labels to the test set, indicates that though with a special structure, signal is present in the data. This is discussed further in Section 7.2.

6.4.4 Data From the Awake Condition

Data collected from pre-motor cortex in the awake condition is also shown in Figure 6.3 - 6.6 and Figure A.1 - A.14 in Appendix A. Non-artifact rejected data yields a classifier performance better than chance for 7 out of the 18 subjects (109, 111, 117, 134, 138, 139, and 144) with up to 100 features included in the model. One subject (118) is a borderline case. 6 subjects (107, 122, 127, 129, 137, and 149) exhibit the anti-learning behavior further discussed in Section 7.2. However, several subjects indicate that including more features lead to better classification performance. This is investigated for all subjects including up to a 1,000 features. Additionally, the analysis is done on both artifact rejected and non-artifact rejected data. In that case 10 of the 18 subjects yield good classification performance (109, 111, 113, 117, 118, 134, 138, 139, 144, and 149) where two of the subjects 111 and 113 gives the best performance on artifact rejected data. 6 subjects (107, 122, 127, 129, 137, and 147) show strong anti leaning behavior where subject 122 and 147 gives the result for artifact rejected data. Only 2 out of the 18 subjects yield chance level results for the chosen data representation. The remaining 16 yield results different from chance. Only selected results are shown in Appendix C. For subject 113, see Figure C.1a, it can be seen that building the model based on the top 200 features using artifact rejected data yields improved classifier performance and for subject 118, see Figure C.1b, it can be seen that the model improves significantly by including around 250 features using non-artifact rejected data in the model.
Very interestingly, it can be observed in general for all subjects, see Figure C.2a and C.2b for two examples, that anti-learning is reduced when more (noisy) features are added to the model. This behavior is similar to what is seen for normal learning, see Figure 4.2. But conversely to learning, anti-learning is reversed to learning for some subjects, e.g. subject 149, see Figure C.2b, for the non-artifact rejected data set. Here, if less than 100 features are included in the training set, anti-learning is observed, whereas proper learning is observed

when more than 400 of the top ranked features are included in the training set. It is not always the case that anti-learning is reversed to learning by adding more features. The representation might simply be too noisy compared to the number of available trials.

6.4.5 Spectro-histo-grams

Along with the obtained classification rates, the features selected to obtain these are investigated. For every subject, see Figure 6.3 - 6.6 and Figure A.1 - A.14 in Appendix A, subfigures are shown corresponding to electrode C3 and electrode C4. The figures are combined spectro- and histo- grams, showing the number of times a feature is selected during the 300 iterations of the described cross-validation scheme, see Section 5.2. The histograms are composed by counting the top 50 features. Acceptable inspection of the spectrograms and the corresponding selected features is a trade-off between noise and informative features. An increased number of features significantly increases the scattered noise and makes it harder to see patterns, whereas decreasing the number of included features might remove too much information to actually see any patters. The dominating patterns are present over a wider range of included features, but there is inter-subject variation.
Generally, it can be seen that the classifier uses pre-stimulus features along with post-stimulus features to create the decision boundaries. The pre-stimulus most likely serves like a kind of baseline for the classifier. However, this does not mean that features selected post-stimulus are all discriminative between the two conditions. Post-stimulus features can easily serve as baseline as well. Whether a feature is discriminative or serves as a "reference" is not to say since the selected features together compose one high-dimensional discriminative pattern, separating the two classes.
For a lot of the subjects the patterns of selected features are scattered and it is hard to extract clear information as can be done in standard ERP analysis. It is further to be noted that the classifier uses the actual power levels in each electrode, hence the classifier can choose features from both C3 and C4 even though contralateral patterns are expected for the two classes.
In the following the main patterns found in the subjects with discriminative performance are described.

For **subject 129**, see Figure 6.5, the most prominent features are selected in the alpha band ∼700 ms post-stimulus in electrode C4 and in the high beta ranges at electrode C3 at the same timing. The corresponding classification error is low, around 25 % for several subset of features. For this subject the permutation test is performed with N=1,901 and N=20 to compare the variation. The variation is not substantial, and it is clear that the result is still significant at a

95% significance level.
For **subject 117**, see Figure 6.3, the classifier is also performing very well with error rates as low as 15 %. However, the selected features are rather different. Several early and late, very high beta and low gamma features are selected around ∼500 ms post-stimulus and again 2,100 ms post-stimulus in electrode C4, furthermore beta range features ∼1,000 ms post-stimulus are selected in C4. Features selected in C3 are more scattered and less frequent.
Subject 139, see Figure 6.6, shows a clear pattern in the upper alpha and lower beta around ∼1,100 ms in electrode C4. Early, ∼300 - 600 ms, and pre-stimulus features in the beta band are furthermore selected. In electrode C3, gamma range frequencies are selected ∼1,200 ms post-stimulus.
Subject 105, see Figure A.1, again shows a rather different pattern. The obtained error rates are relatively good, but mostly very early and very late features are selected in the alpha and beta band. Additionally gamma band frequencies dominates the picture. In electrode C4 features in the alpha band are selected ∼900 ms post-stimulus, furthermore activity in the high beta and low gamma band is selected at around the same timing
A somewhat similar pattern is seen for **subject 122**, see Figure A.6 with very early features in lower frequencies and then mainly gamma band features post-stimulus. However, the classification error is slightly higher.
For **subject 107**, see Figure A.2, the picture is the same though some very late alpha activity is selected in C4 and the classifier obtains a relatively low error rate.
In **subject 111**, see Figure A.4, post-stimulus activity is selected in C4 in the beta band from 0 - ∼700 ms and ∼800 ms post-stimulus alpha activity is selected in electrode C3. In both electrodes there is scattered low gamma activity and in electrode C4 a notable amount of features in the gamma range is selected ∼2,000 ms post-stimulus.
For **subject 113**, see Figure A.5, very early alpha and beta features are selected. ∼1,000 ms post-stimulus high alpha and low beta features are selected in electrode C3 and additionally beta activity ∼1,400 post-stimulus is selected in C3. Scattered features are selected in the gamma range.
Subject 134, see Figure A.8, shows from the error rate that only few features are necessary, hence spectro-histo-grams for 6 included features can be seen in Appendix A.15. This shows only two clear patterns, beta activity selected around ∼350 ms post-stimulus and gamma activity selected ∼ 900-1,100 ms post-stimulus.
Subject 138, see figure A.10, shows alpha activity selected 1,100 ms post-stimulus in electrode C4 followed by selection of beta range activity ∼1,300-1,600 ms post-stimulus. Late gamma activity is selected in both electrodes.
Subject 144 shows a very blurry picture with scattered activity above alpha selected over full time interval, the picture remains more or less unchanged for fewer features included. Beta activity selected ∼ 2,100 ms post-stimulus is most dominant.

The above observations give no clear picture, but some common patters can be extracted.

6.4.6 Inter-Subject Learning & Group Level Results

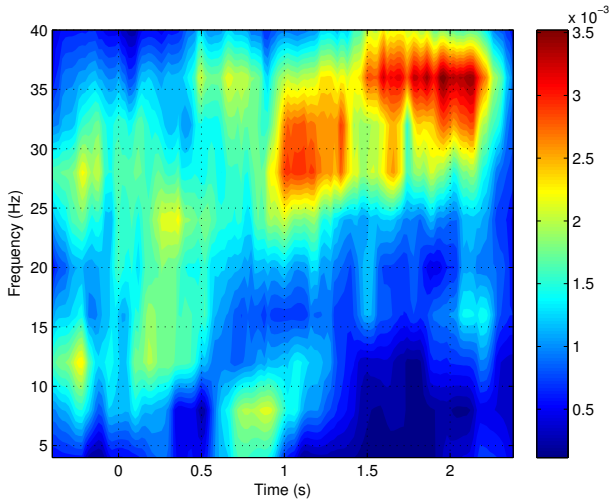
Using the described feature selection and classification framework, it is not possible to classify on an inter-subject basis. The group averaged classification error rates can be seen in Figure 6.2. Summing and scaling all the spectro-histograms, S, obtained for the 12 subjects leading to good classification performance gives a completely scattered image. However, by applying an averaging filter, some general structure is indicated, see Figure 6.7. The corresponding Z-scores, see Figure 6.7b, show that especially high beta and gamma activity transfers across subjects, but also around 1,000 ms post-stimulus, a pattern of low alpha and theta activity is consistently detected. Interestingly, pre-stimulus is selected rather consistently across subjects as described earlier. The Z-scores ($Z = \frac{\frac{1}{n}\sum S}{\sigma(S)}\sqrt{\frac{n}{2}}$) are found using a split-half resampling strategy inspired by the NPAIRS framework [Strother et al., 2002] and the nonparametric analysis of statistical images presented by [Holmes et al., 1996].

6.4.7 Learning Curves

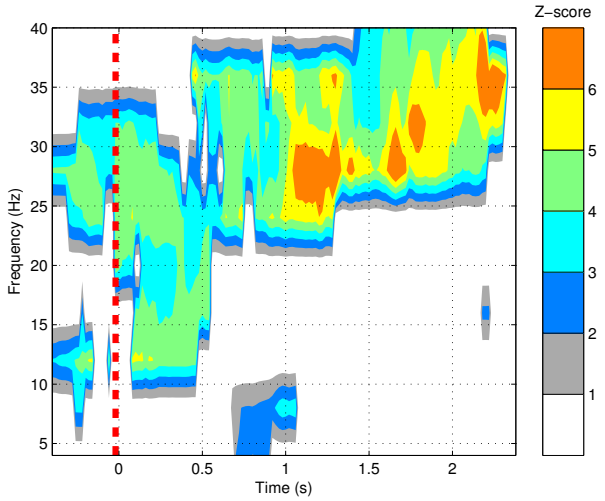
For all subjects learning curves are created where classifier error rate is shown as a function of the size of the training set. For subject 117, see Figure 6.8, both learning curves for different feature size subset and for a cross-section taken at the best performing subset are shown, see Figure 6.8a and B.3a. The learning curves for other subjects are in Appendix B, see Figure B.1a - B.6b. Only cross-sectional learning curves are shown and they generally indicate that convergence in terms of the included number of trials is not obtained for the full set of trials. From the slopes it seems clear that the classifier in most cases would benefit from more trials included.

6.4.8 Included Electrodes

In the main analysis only electrodes C3 and C4 are included. An analysis where the electrodes posterior and anterior to C3 and C4 are included is however also conducted, such that the classification is done based on 6 electrodes instead of 2. For most subjects performance is decreased due to the increased dimensionality,

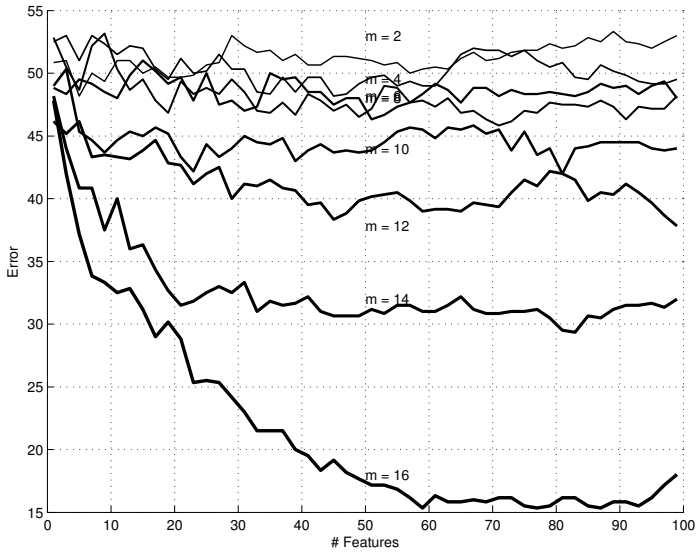


(a) Group average of scaled spectra-histo-grams.

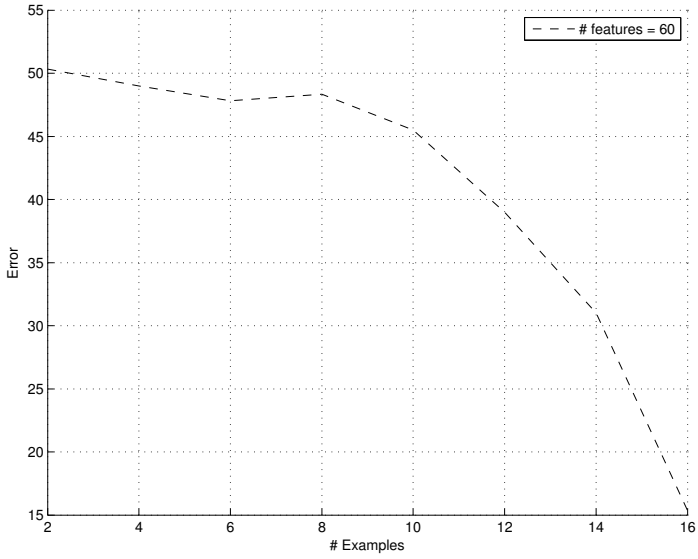


(b) Z-score.

Figure 6.7: Group averaged spectro-histo-gram in Figure 6.7a. 6.7a is a compilation of spectro-histo-grams from electrodes C3 and C4 for the 12 subjects yielding good classification performance. Each spectro-histo-gram is scaled by its sum. To reduce the scattered structure it has been convoluted with an averaging filter. In Figure 6.7b the corresponding Z-scores are calculated using a split-half resampling scheme. Low values in the averages spectrogram are thresholded out.



(a) Learning curves for subject 117.



(b) Learning curve for subject 117.

Figure 6.8: Cross-section of the learning curves with 60 features included, and the full set of learning curves for subject 117, where m indicates the number of trials included in the training set.

but for three subjects performance is increased (122, 127, and 137). Two of the three subjects (127 and 137) are the borderline cases from the previous analysis, which yield significant results when more electrodes are included. Lateralized averaging of neighboring electrodes does not generally improve performance.

6.4.9 Kernel Parameter

The kernel parameter, σ, is found using the described cross-validation scheme and results from Section 5.5.1. It is furthermore investigated which optimal kernel parameter is selected during the cross-validation. Using a confusion matrix, only values where the classifier predicts both labels correctly are investigated. When the classification model is estimated from up to around 20 features the majority of the σ's averaged over subjects are selected within the interval corresponding to the $]0.1; 0.9[$ quantile of the width. Nevertheless, there is a large variability between subjects. When more features are included in the model, only a few percent of the models have σ-values corresponding to an almost non-linear decision boundary. These results are obtained if the cross-validation is set to prefer higher σ-values if equally good. For the reverse case where the cross-validation is set to prefer a smaller value of σ the σ's selected for higher dimensional problems are tend to be smaller.

6.4.10 Number of Cross-validation Iterations

The number of leave-two-out cross-validation iterations included in the analyses are set to 300 for all subjects. A value of 300 leads to convergence within a few percent compared to up to 1,000 cv-iterations included and obviously reduce computational time compared to 1,000 cv-iterations where all subjects produce very stable solutions.

Chapter 7

Discussion

Discriminating brain activity was found in the majority of the subjects. The localization of the patterns are discussed in the following.
Classification errors significantly above 50% might be hard to grasp at first sight in the presented framework, since the classifier could obtain the reverse (and desirable) result by flipping all labels. This anti-learning behavior is discussed in the following.
Finally improvements to the setting and computational issues are discussed.

7.1 Spatiotemporal Cortical Dynamics

For the 12 subjects showing discriminating brain activity, the overall picture is that general patterns do seem to occur though latency, morphology and range varies. Alpha and beta range features show clear patterns between ∼700 and ∼1,400 ms post-stimulus at a single-subject level. Both very early and late patterns occurs as well. For several subjects a considerable number of features are also selected in the gamma band over the course of the epoch. Here latency varies over the full epoch, but late activity is generally more clear. Figure 6.7 showing the averaged result and corresponding Z-scores for the group do catch some of these patterns, but the inter-subject variability blur the general picture. Especially high beta and gamma activity transfer across subjects, but

also around 1,000 ms post-stimulus, a pattern of low alpha and theta activity is consistently detected.

As described in the introductory section, see Chapter 2, voluntary movement is widely acknowledged in the literature [Pfurtscheller and Lopes da Silva, 1999] to induce a desynchronization in the upper alpha and lower beta frequency bands close to the motor areas of the cortex. In addition to alpha and beta rhythms, oscillations are also found in the gamma range frequencies [Pfurtscheller et al., 1993, Pfurtscheller and Lopes da Silva, 1999]. [Andrew and Pfurtscheller, 1996] reports an increase of coherence in phase of 40 Hz oscillations between the contralateral motor and the supplementary motor areas during the performance of unilateral finger movements. The gamma activity can be separated over different parts of the cerebrum but exhibit high correlation and synchrony during the performance of motor tasks. In contrast to the alpha band rhythms, the gamma range activity reflect active information processing. According to [Pfurtscheller and Lopes da Silva, 1999] desynchronization of alpha band rhythms may be a prerequisite for the forming of gamma range frequencies. They furthermore suggest that gamma rhythms indicate active information processing, which may be related to a binding of motor integration.

The results obtained could indicate that the activity is a result of motor planning, but since classification is possible using other electrode pairs than C3/C4, some source localization algorithm would have to be applied to be able to truly determine the origin of the signal. Furthermore, in the current setting it is not possible to determine whether classification is based on the contralateral patterns arising from motor planning or some other discriminating effect, though it is the most plausible explanation. In the described experiment, clear contralateral patterns can be found when analyzing the data obtained during wakefulness. Looking at the lateralized readiness potential, LRP, which is obtained by subtracting ipsilateral ERPs from contralateral ERPs recorded over the scalp, contralateral patterns arise which reflect the preparation of motor activity.

Results (though reversed, see the following discussion of anti-learning) obtained from visual cortex are most likely the true cognitive motor effects, just detected further away due to transcranial volume conduction. The effects can be explained from the fact that conduction is not restricted to one direction and thus electrodes all over the scalp receives some signal though fainter than the electrodes nearer the source. Methods, such as SVMs, not relying on averaging techniques to enhance signal-to-noise ratios are more sensitive and can detect even vague signals if present. This is supported by the actual results, which generally show that electrodes over the visual cortex show less predictive power than those at pre-motor cortex, indicating that the origin of the signal is not visual cortex.

Regarding the data obtained during wakefulness, only 16 of the 18 subjects are found to contain either learnable or anti-learable signal, it is however relatively common in EEG experiments that signal from some subjects is hard to analyze or

7.2 Anti-Learning

In continuation of the remarkable geometric representation found in low sample high dimensional settings, see Section 5.1, it was also described how [Hall et al., 2005] and [Klement et al., 2008] show that for small sample size data, a basic SVM will classify all examples as the same class, even in lower dimensions.
As described earlier, to avoid this kind of behavior, all results are obtained with a balanced resampling scheme on a balanced data set. Hence, the described anti-learning is not a product of this.
Further analysis is hence required to deal with the observed anti-learning behavior since this kind of behavior is very unusual. Best case it is a rare event and therefore rarely reported, though it is more likely that when encountered it is ignored or misinterpreted as a case of over-fitting of noisy data not worth further investigation.
Anti-learning is not a product of choice of classifier. It has been shown, [Kowalczyk et al., 2007], that all standard supervised learning algorithms such as the linear SVM, kernel SVM, naive Bayes, ridge regression, k-nearest neighbors, shrunken centroid, multilayer perceptron and decision trees perform in an unusual way on natural and synthetic data sets containing certain structures. They all classify a randomly sampled training set almost perfectly and perform worse than chance on new unseen validation data. The structure in the synthetic data is the outcome of a winner take all zero-sum game where any two examples of the opposite class are more correlated than any two examples of the same class. This can geometrically be interpreted as that the intra-class distance between examples is greater than the inter-class distance. Based on the structure of the "game" and the anti-learning behavior, [Kowalczyk et al., 2007] states: *Such a simple "Darwinian" mechanism makes it plausible to argue that anti-learning signatures can arise in the biological datasets. However, there are also many other models generating anti-learnable signature, for instance a model of mimicry.* Hence, the anti-learning behavior indicate genuine informative features rather than other phenomena, though the structure of the data is different than usually reported. This is supported by the fact that the permutation test yields chance level results. The anti-learning behavior in a general setting is not very well investigated, and reports of the behavior is mostly related to biological low sample size high dimensional feature space data sets, such as microarray data used for prediction of cancer outcomes and data sets originating from bio-medical research, including heart ECG [Kowalczyk and Chapelle, 2005, Kowalczyk, 2007, Kowalczyk et al., 2007].
There is a US Patent Application by Adam Kowalczyk, Alex Smola, Cheng Soon Ong, and Olivier Chapelle [Chapelle et al., 2005] concerning unlearnable data sets. They propose to use a reverser to apply negative weights to predicted labels if anti-learning is observed, whereby error rates translates directly into accuracy. Furthermore it is described, see also [Kowalczyk et al., 2007], how

anti-learnable data can be transformed using a non-monotonic increasing kernel function to avoid anti-learning behaviour. A transformation like that can increase the within class similarity and decrease inter-class similarities. If the considered data set exhibit perfect anti-learning the transformation can lead to perfect learning, but for data-sets that are not perfectly anti-learnable, the transformation has more limited potential.
As observed in the awake data sets, see Section 6.4, where anti-learning decrease when more (noisy) features are added to the model, [Kowalczyk et al., 2007] also observe that addition of random noise suppresses the symmetries leading to anti-learning. When even more features are added, anti-learning is reversed to learning for the data recorded during wakefulness but not during sleep. This might be another indication that the signals recorded during wakefulness are stronger than those collected during sleep.
Anyway, the 16 subjects in the awake case of the experiment showing either learning or anti-learning are reported to exhibit discriminative brain activity supported. This is supported by a permutation test.
It is not possible consistently to find the structures exhibited by a perfect anti-learning data set in the present EEG data set. However, it is possible for certain feature subsets and certain sub samples to show a somewhat similar structure where the intra-class distances are larger than the inter-class distance. In the data sets exhibiting anti-learning it is generally the case that the classifier predicts approximately equally many labels in each class. Additionally some signs of hubs can be found if a kernel matrix of the nearest neighbors are plotted. Whether these are related to the anti-learning is still to be determined.
Though not reported very often, anti-learning is suspected to occur in many other data sets as well. It would be interesting to make an analysis using the same framework as presented here, but on a data set where it is possible to move away from the low sample size setting and data set characteristics are well known. Such data sets could e.g. be the BCI competition data sets which are publicly available EEG data sets with well known characteristics and many available trials [Tangermann et al., 2012].

7.3 Data Representation

The optimal spectral transform and data representation have probably not been found. It seems that the wavelet transform generally leads to better performance than a multitaper transform for the signals recorded during sleep. This is in agreement with [van Vugt et al., 2007], who find that multitaper methods are less sensitive to weak signals but very frequency-specific compared to wavelets. Generally, the multitaper transform is more applicable for analyses of higher frequencies as well [van Vugt et al., 2007, Raghavachari et al., 2001, Hoogen-

boom et al., 2006].
Of the tested configurations a wavelet transform on down sampled data split into frequency bins of 4 Hz gives good performance using 7 cycles in the transform. Some subjects benefit from fewer cycles whereas some subjects benefit from more. Likewise, both finer and more coarse frequency and temporal resolution leads to better performance for some subjects. Although using signals from electrodes C3 and C4 generally leads to good performance, it is furthermore seen that some subjects benefit from a less sparse representation where 6 electrodes and even up to 18 electrodes are included. For at few subjects the multitaper transform likewise leads to better performance, though overall the wavelet transfer is superior.
It is beyond the scope of the present thesis to evaluate tuning of different spectral transforms relative to each other. Rather it is a goal to find a feasible transform. Hence, the significance of the claims about which spectral transform is better was not tested.

7.4 Computational Issues

Given that data collection itself takes hours for just one subject and days for several subjects and the total planning time of an experiment probably is in the order of months, a total running time of a few hours to do the data analysis is acceptable, it is however desirable that the algorithm is as fast as possible such that it is possible to run several analyses.
Since the full permutation test with 1,901 permutations, see Figure 6.5, takes more than a weekend to run for one subject, the permutation tests in this thesis are run with $N = 20$. 1,901 permutations seems overly conservative and the difference between $N = 20$ and $N = 1,901$ is not decisive. [Strother et al., 2002] uses 10 permutations for each subject in a split-half setting.
It was assessed that 300 cross-validation iterations produced stable enough results when comparing different setups, though around 1,000 iterations are required to obtain very stable results.
Combined with the argument from Section 5.5.1, the investigation of the kernel parameter found from the cross-validation indicates that non-linear decision boundaries yield better performance than linear. This is especially the case for very sparse models with only few input features. Hence it is worth using the extra computational effort required to solve the non-linear version of the SVM. The learning curves give a strong indication that the non-linear SVM could benefit highly from more included examples in the training set. This also elucidate why the test set (leave-two-out) used for cross-validation should be as small as possible.
Since the program runs in parallel on a shared cluster system where recourses

are distributed according to the number of users (which varies a lot), there is no direct way of assessing exact time savings based on the adjustments made to the stopping tolerance. Just running single parts of the program on a local computer shows that during the first few iterations, the main time-consuming part is to build the kernel matrix and not running the actual optimization. These iterations corresponds to most of the evaluations of the SMO-algorithm since all features are included. During later iterations time savings are considerably above 50 % since the kernel matrix is composed of fewer features. Overall the computational time is reduced by using a more coarse tolerance, but additional time savings could be obtained by reducing the number of evaluations of the kernel matrix, in the current setting there is however no direct way of doing that.
The RFE algorithm can be changed to use the assumption that the optimal Lagrange multipliers do not change significantly with one feature left out. Hence the algorithm should only recompute the kernel with one feature left out and then recompute the value of the cost function. However, this does not solve the problem of reducing the number of kernel evaluations.
In the current thesis LIBSVM [Chang and Lin, 2011], which relies on the SMO-algorithm, is used to solve the presented QP. Nevertheless, the kernel matrix is relatively small in the current single-subject setting and the QP could easily be solved directly using standard QP-solvers in MATLAB. LIBSVM does however provide a fast C++ implementation of support vector machines with a Matlab interface, which is believed to be at least as fast. The results obtained for the stopping tolerance in the SMO-algorithm are expected to generalize to some degree for QP solvers in general.

7.5 Future Work - Harnessing the Machine Learning Approach

Anti-learning obviously requires a more thorough investigation, but there are several interesting directions to follow in the described framework to further utilize the machine learning approach.
Automatic electrode selection is on obvious improvement. The EEG-cap is mounted manually on every participating subject, which might lead to some inter-subject variability. Electrodes C3 and C4 may be more or less posterior than desired and therefore other electrodes in the vicinity of C3 and C4 might yield better classifier performance. Furthermore, the cap might move slightly during an experiment. This obviously leads to lower intra-subject classification performance. For all subjects it is tested in the awake condition whether electrodes anterior or posterior to the C3 and C4 electrode pair can improve classification accuracy. For some subjects classification results improved using

more posterior electrodes in the awake case, but it was not consistently a good indication of what electrodes worked best for the sleep data. Combining the relation between the most important electrodes in the awake condition however, could lead to additional insights and further weight could be given to the obtained results if electrodes selected automatically in the awake condition could be transferred to the sleeping condition.
Ideally, a framework where the most important electrodes for the classification are determined automatically using the data should be built. The most informative electrodes could be selected using either a clustering algorithm or a common spatial pattern on the full set of electrodes. This would both reduce the influence of electrode-cap placement and reduce inter-subject variability. FieldTrips clustering algorithm was tested on the data-set, and did indeed find C3, C4, CP3, and CP4 to be the most relevant electrodes in most cases. Hence it would be a good path to follow.
Inspired by the presented approach to feature elimination the method can be expanded to include a recursive channel elimination step where channels are removed iteratively in a similar fashion to the recursive feature elimination. This would reduce the inclusion of prior knowledge, which, as discussed, can be misleading for new experimental paradigms. Furthermore, it would automatically find the most relevant brain areas for the classification. This would require a lower dimensional feature space for each channel, which could be achieved by using e.g. autoregressive coefficients as proposed in [Lal et al., 2004].
The artifact rejection should be automatized to a higher degree to reduce human bias and inconsistency. Using a toolbox like CORRMAP[Campos Viola et al., 2009], would allow the identification and clustering of independent components representing EEG artifacts. Finally, finding a good data representation is key and there are several parameters to investigate further, likewise more advanced methods can be adopted which can deal with variable latencies in data derived from neural activity. Multilinear shift-invariant decompositions can handle this to some degree [Mørup et al., 2008].

Chapter 8

Conclusion

The results obtained in the present thesis indicate that the sleeping human brain is responsive to external auditory input and indeed capable of processing it at a semantic level.
Discriminative activity is found in at least 12 out of the 18 subjects during sleep. Using the proposed method during wakefulness, it is possible to show discriminative activity in 16 out of 18 subjects.
It remains a question whether the observation generalizes to all sleep stages, and to elucidate the degree of motor cortex involvement. Furthermore, it is less clear where in the process from semantic processing to actual execution the obstruction of neural signals occur.

A novel framework for analyzing EEG-recordings has been presented in a challenging context of low sample high dimensional data. The challenge is magnified by the low signal-to-noise ratio from the unexplored experimental paradigm investigating the degree of semantic processing during sleep. The proposed framework gives a way of analyzing whether discriminative patterns are present in the recorded EEG signals, and where the information is located at a single-subject level. The method succeeds regardless of effective behavioral response which is generally absent during sleep.
The engine of the proposed framework is a non-linear SVM with a RBF kernel combined with an RFE algorithm. The presented implementation is a modified and improved algorithm, where parameters are re-estimated during every

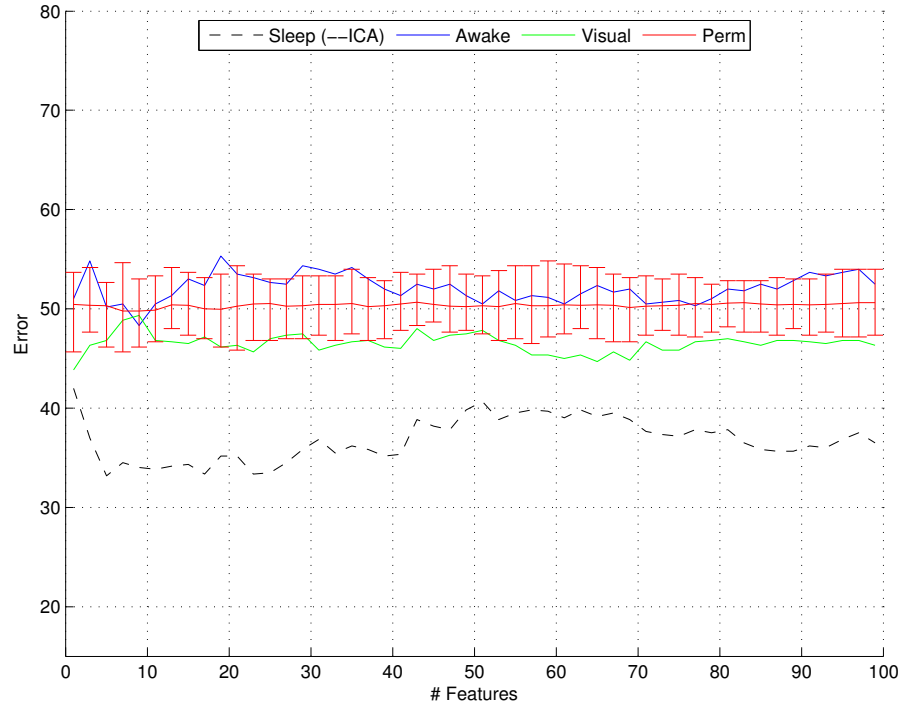
iteration according to a heuristic based on the structure of the feature subset. Even in the low sample size setting, the non-linear capabilities are utilized and non-linear decision boundaries are found to improve classification.
In addition to improved classification performance, the RFE algorithm gives the possibility to investigate a reduced subset of features driving the classifier. This makes the method especially powerful in new paradigms with no or few certain priors established. The results obtained from the feature selection show a large degree of inter-subject variability regarding latency and morphology of brain activity. Anyhow, the results are somewhat similar to what is found in classical ERP and ERD studies of EEG in awake motor-studies. This indicates that the brain prepares a relevant response all the way up to motor preparation during light sleep though execution is clearly suppressed. Since classification was possible using other electrode pairs than those of motor areas, this needs further investigation of source localization. Superior classification performance was however obtained in motor areas indicating a relation to the origin of the signal.

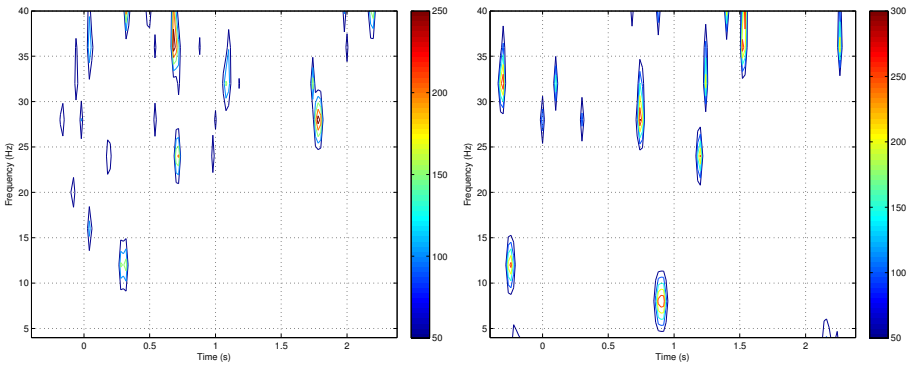
In contrast to the full spectral transform obtained for all channels, a very sparse representation giving good performance is found using a more coarse Morlet wavelet transform only including channels C3 and C4.
Compared to classical statistical analysis, the method excels by the fact that even for the low sample size setting multivariate analysis is possible at the individual subject level where classical analysis can only be done at a group level.
Along with the sparse data representation, computational tractability was improved by adjusting the stopping tolerance of the SMO algorithm during the RFE-loop. This was shown to improve speed without jeopardizing accuracy.
From the classifier learning curves found in the single-subject analyses, it is clear that classification performance in general would benefit from more experimental trials.
Anti-learning was observed, especially in the data obtained from the visual cortex, but plausible explanations have been given. For the awake condition it has been shown how addition of features can reduce anti-learning, and in some cases even change anti-learning to learning.

SVMs combined with embedded feature selection schemes show some general encouraging characteristics in relation to interpretation of EEG recorded neuroscientific data collected in new as well as well-known experimental paradigms. Using a margin-based recursive feature elimination algorithm, it is possible to classify and characterize discriminative brain activity based on full spectrograms at a sensor level for single subjects.

Appendix A

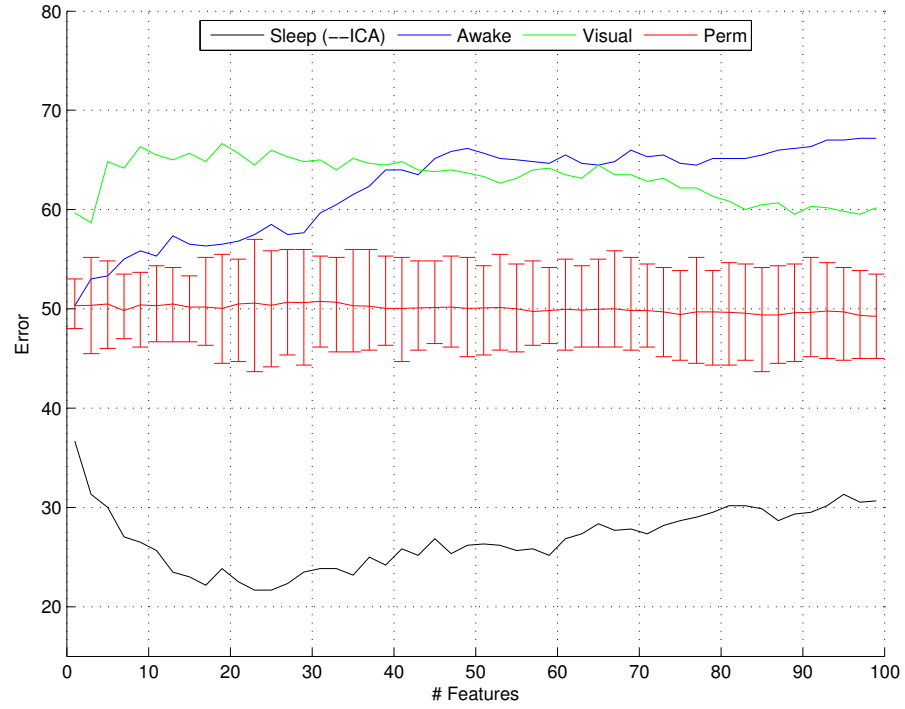
Classifier Performance and Spectro-histo-grams

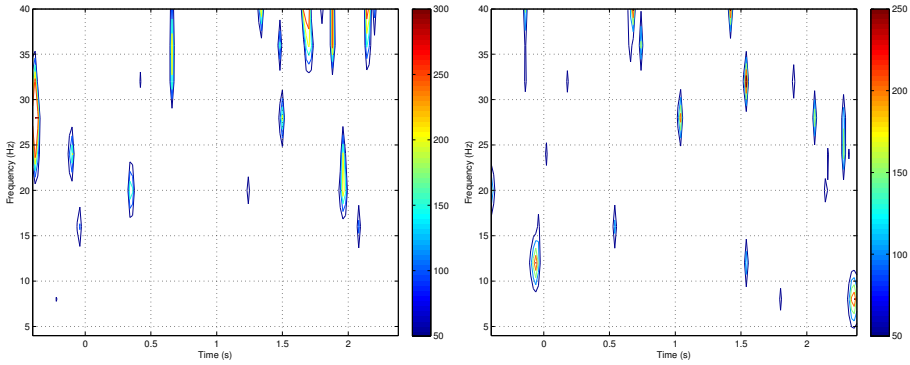




(a) Electrode C3 **(b)** Electrode C4

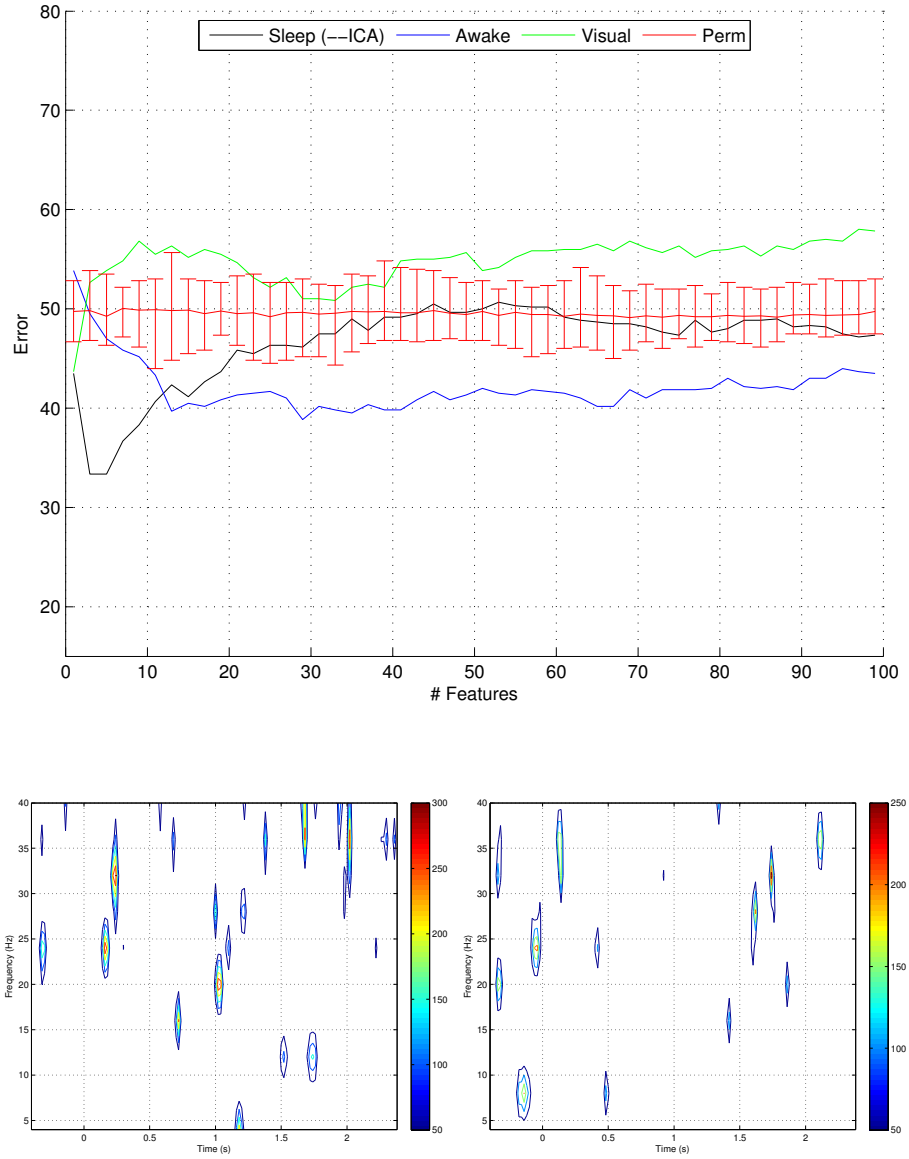
Figure A.1: Subject 105. Top 50 time-frequency features selected during each leave-two-out iteration for subject 105. The number of times a feature is selected gives an indication of how important the single features are.





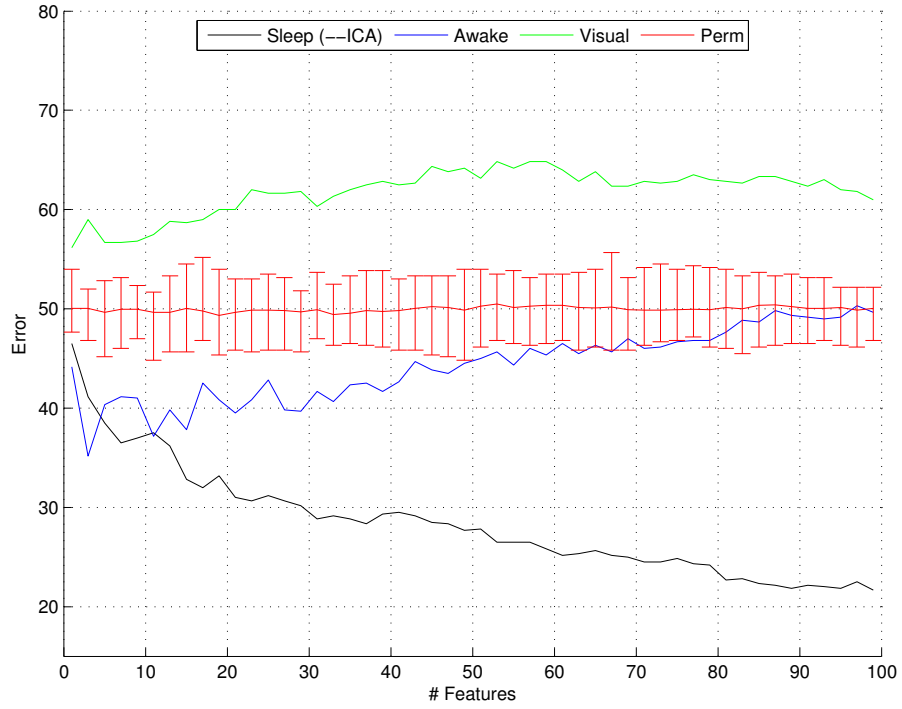
(a) Electrode C3 **(b)** Electrode C4

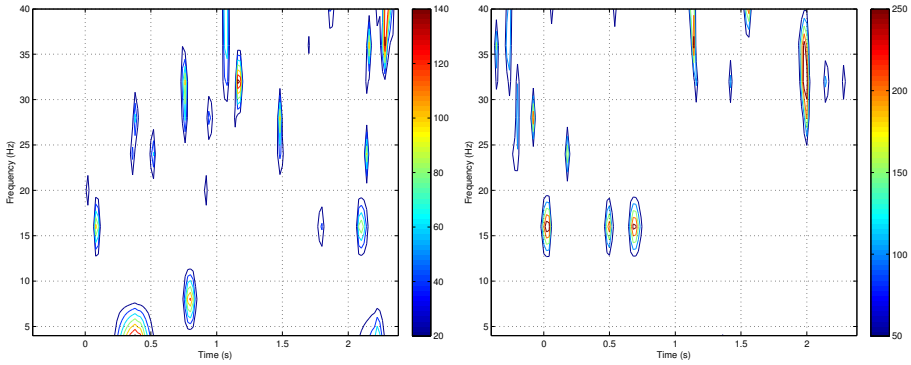
Figure A.2: Subject 107. Top 50 time-frequency features selected during each leave-two-out iteration for subject 107. The number of times a feature is selected gives an indication of how important the single features are.



(a) Electrode C3 **(b)** Electrode C4

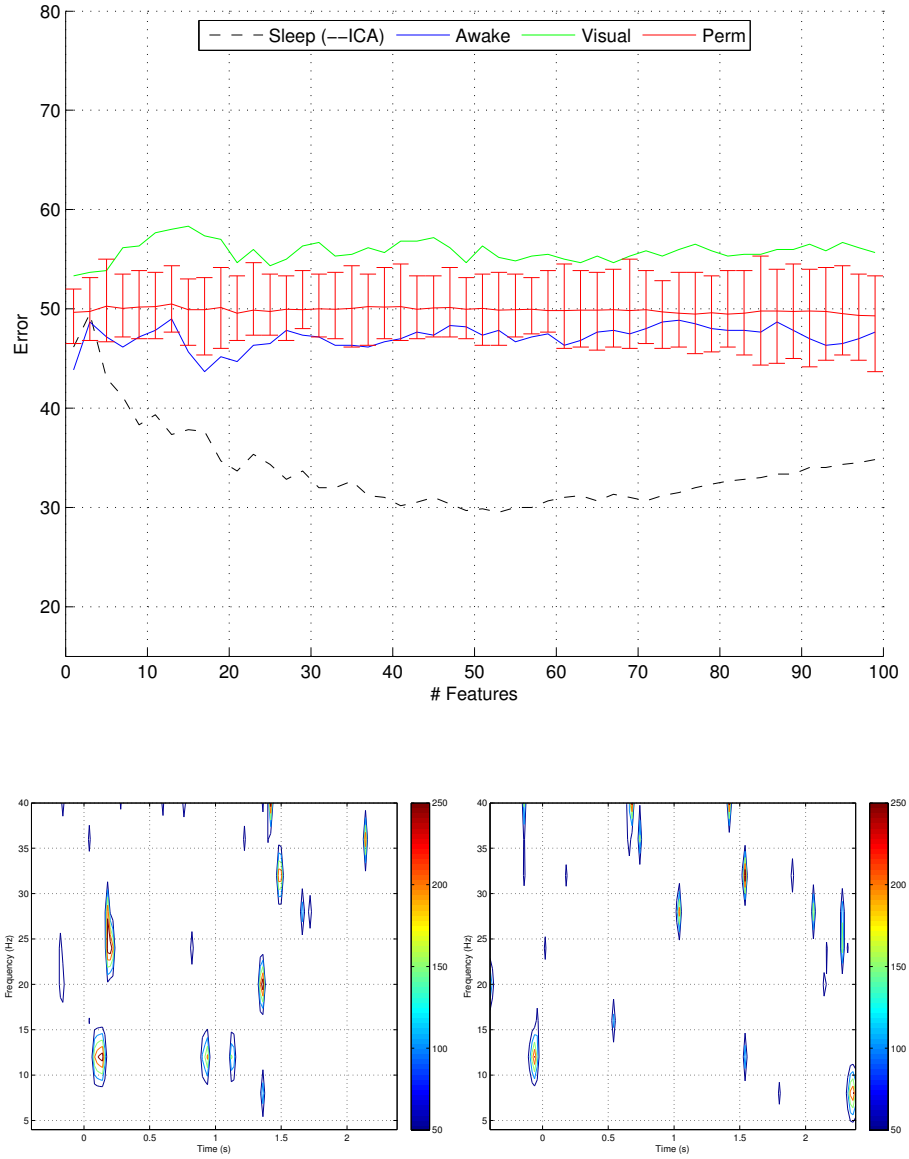
Figure A.3: Subject 109. Top 50 time-frequency features selected during each leave-two-out iteration for subject 109. The number of times a feature is selected gives an indication of how important the single features are.





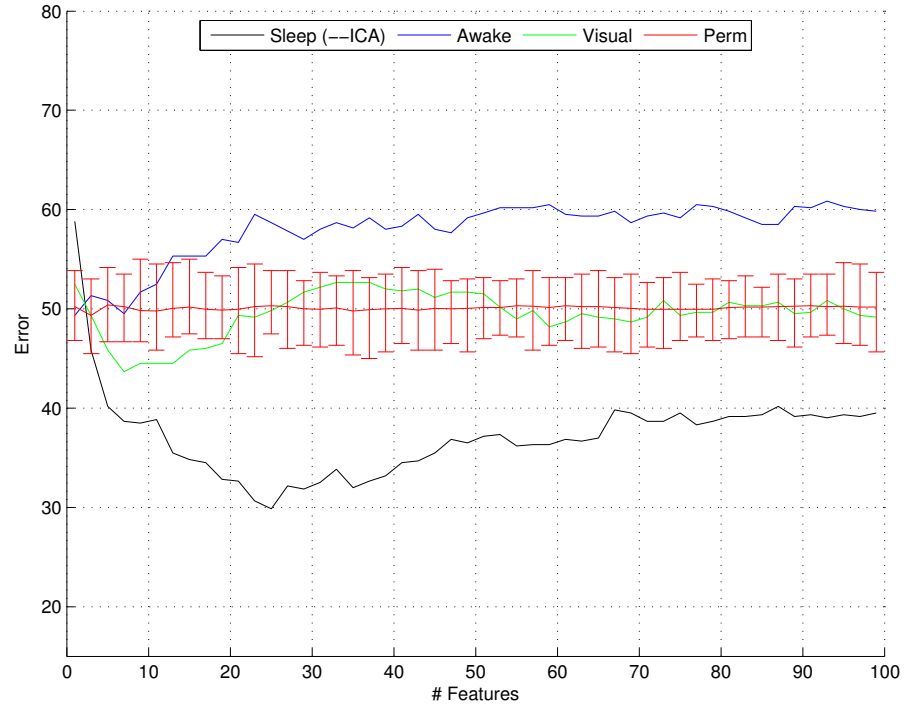
(a) Electrode C3 **(b)** Electrode C4

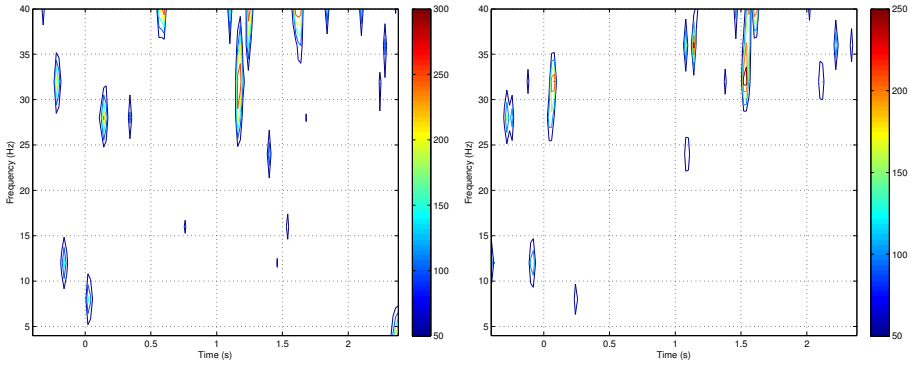
Figure A.4: Subject 111. Top 50 time-frequency features selected during each leave-two-out iteration for subject 111. The number of times a feature is selected gives an indication of how important the single features are.



(a) Electrode C3 **(b)** Electrode C4

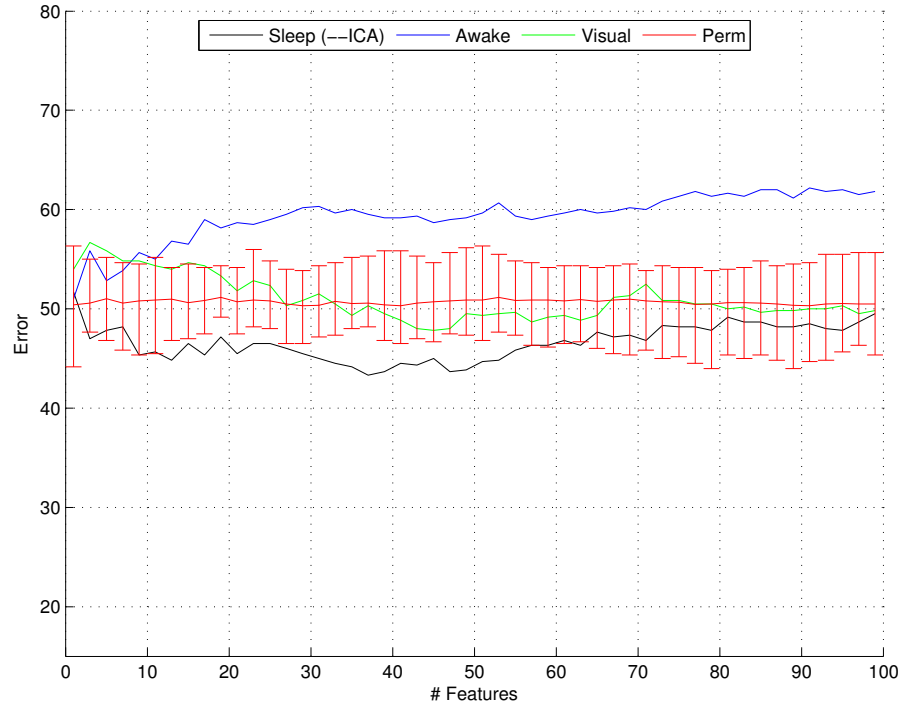
Figure A.5: Subject 113. Top 50 time-frequency features selected during each leave-two-out iteration for subject 113. The number of times a feature is selected gives an indication of how important the single features are.

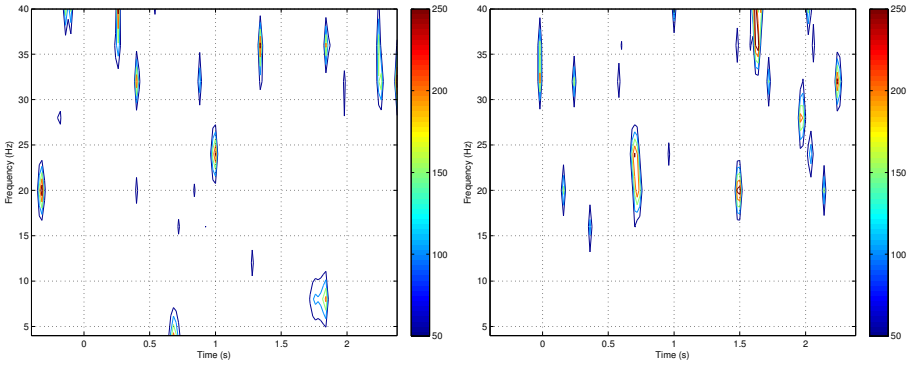




(a) Electrode C3 **(b)** Electrode C4

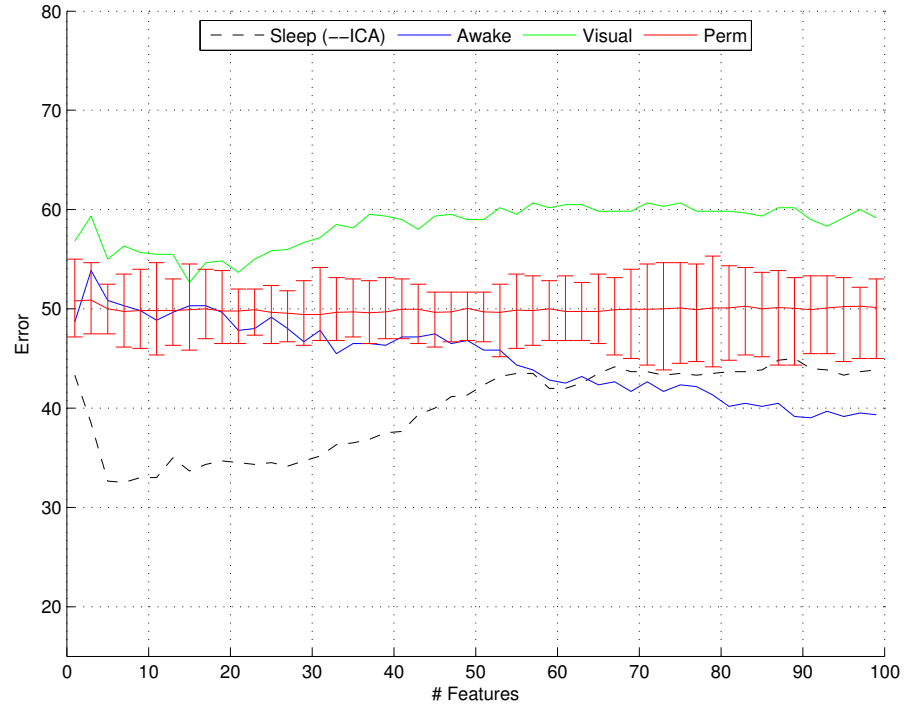
Figure A.6: Subject 122. Top 50 time-frequency features selected during each leave-two-out iteration for subject 122. The number of times a feature is selected gives an indication of how important the single features are.

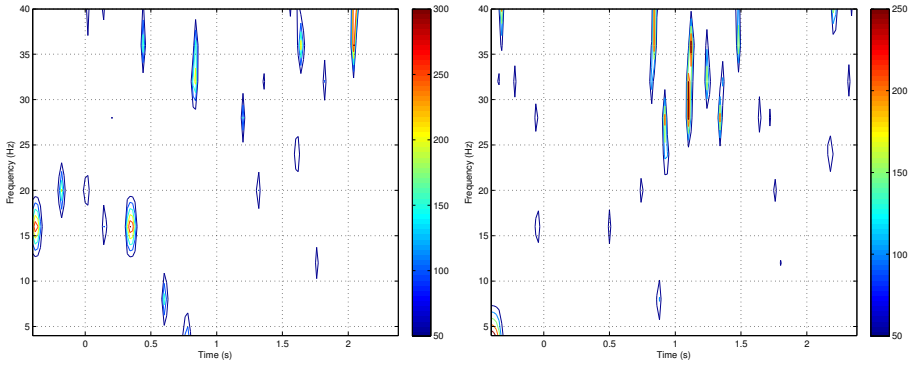




(a) Electrode C3 **(b)** Electrode C4

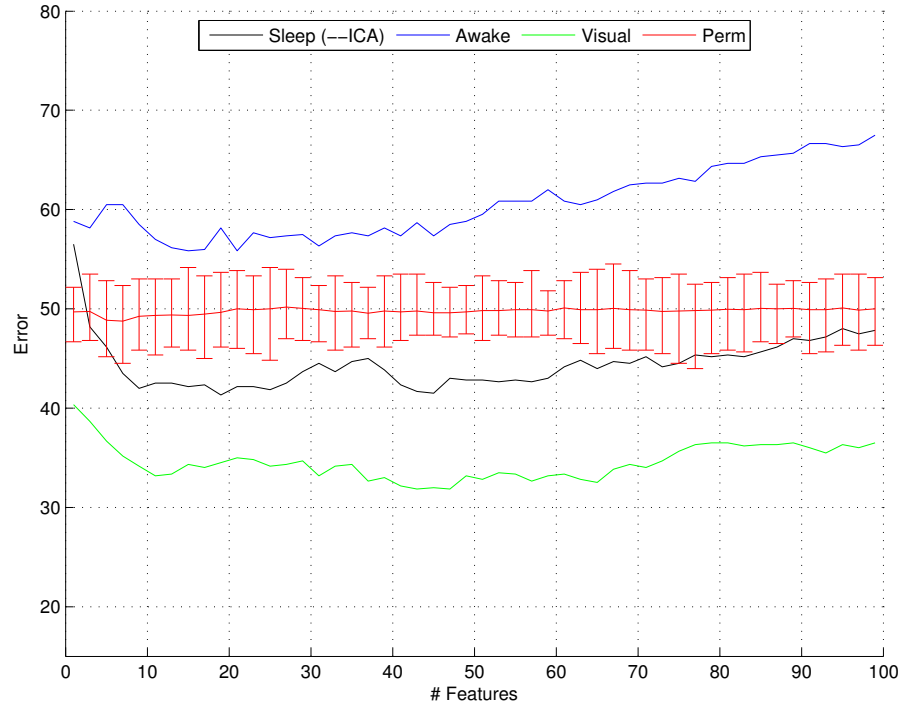
Figure A.7: Subject 127. Top 50 time-frequency features selected during each leave-two-out iteration for subject 127. The number of times a feature is selected gives an indication of how important the single features are.

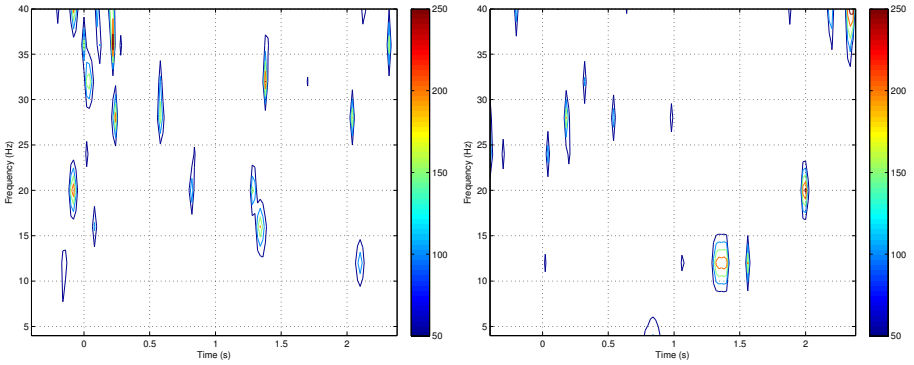




(a) Electrode C3 **(b)** Electrode C4

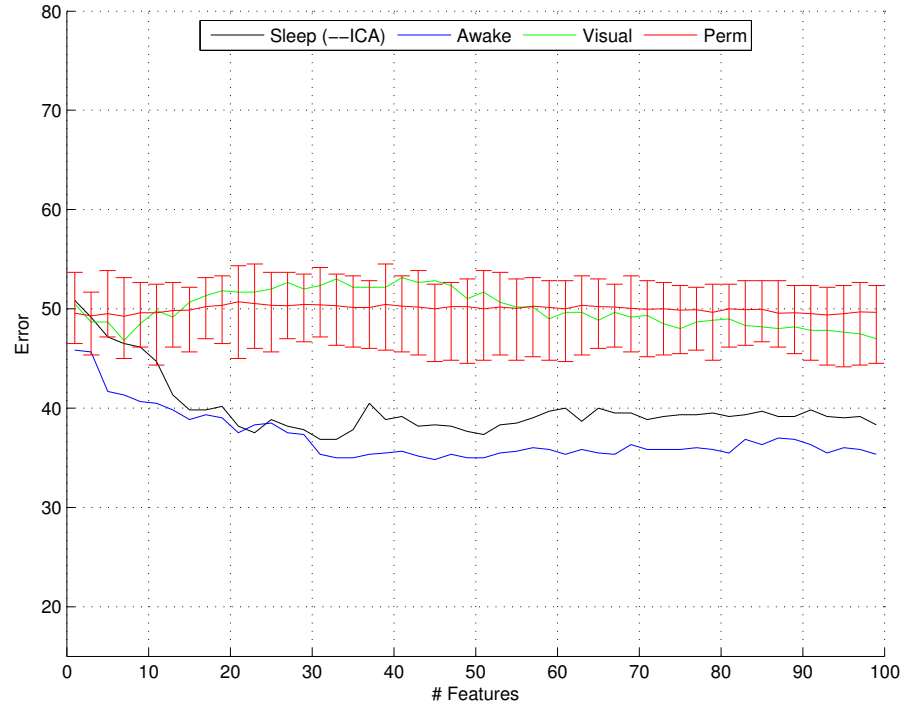
Figure A.8: Subject 134. Top 50 time-frequency features selected during each leave-two-out iteration for subject 134. The number of times a feature is selected gives an indication of how important the single features are.

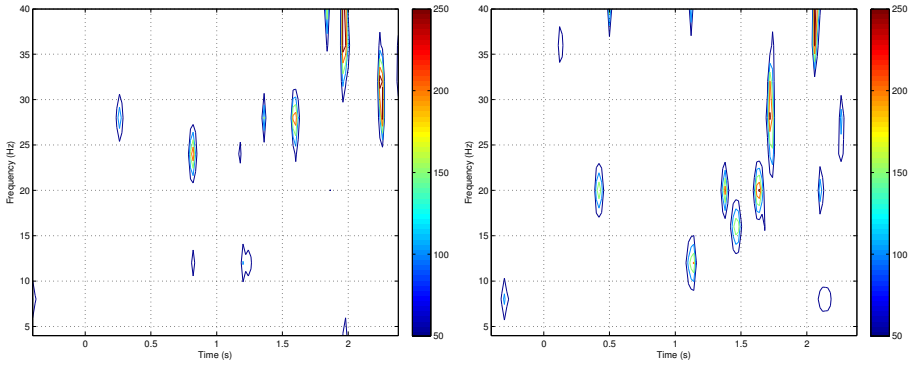




(a) Electrode C3 **(b)** Electrode C4

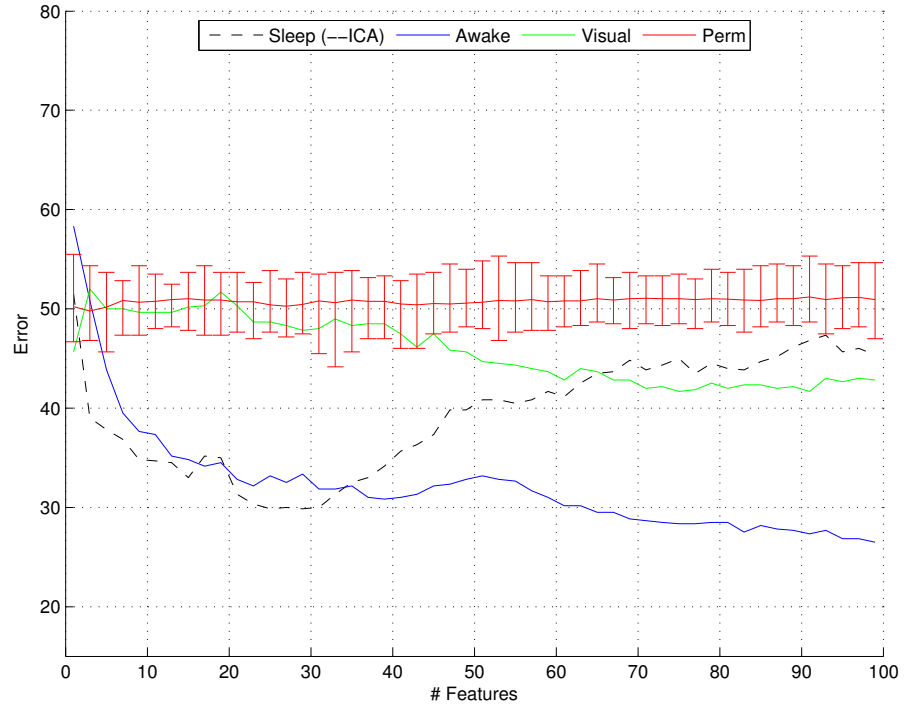
Figure A.9: Subject 137. Top 50 time-frequency features selected during each leave-two-out iteration for subject 137. The number of times a feature is selected gives an indication of how important the single features are.

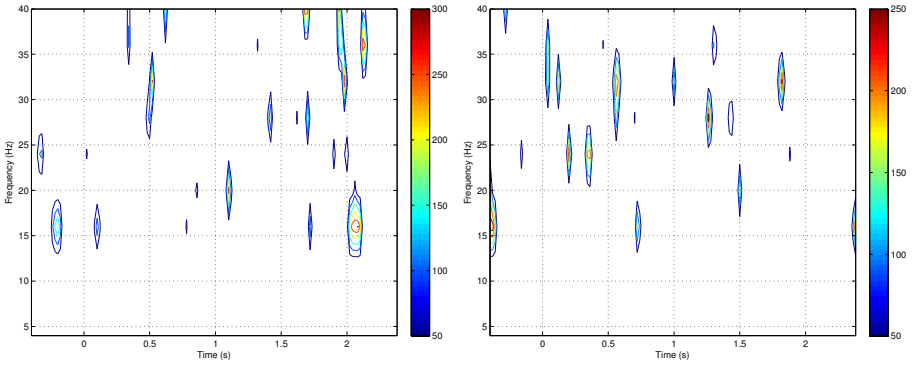




(a) Electrode C3 **(b)** Electrode C4

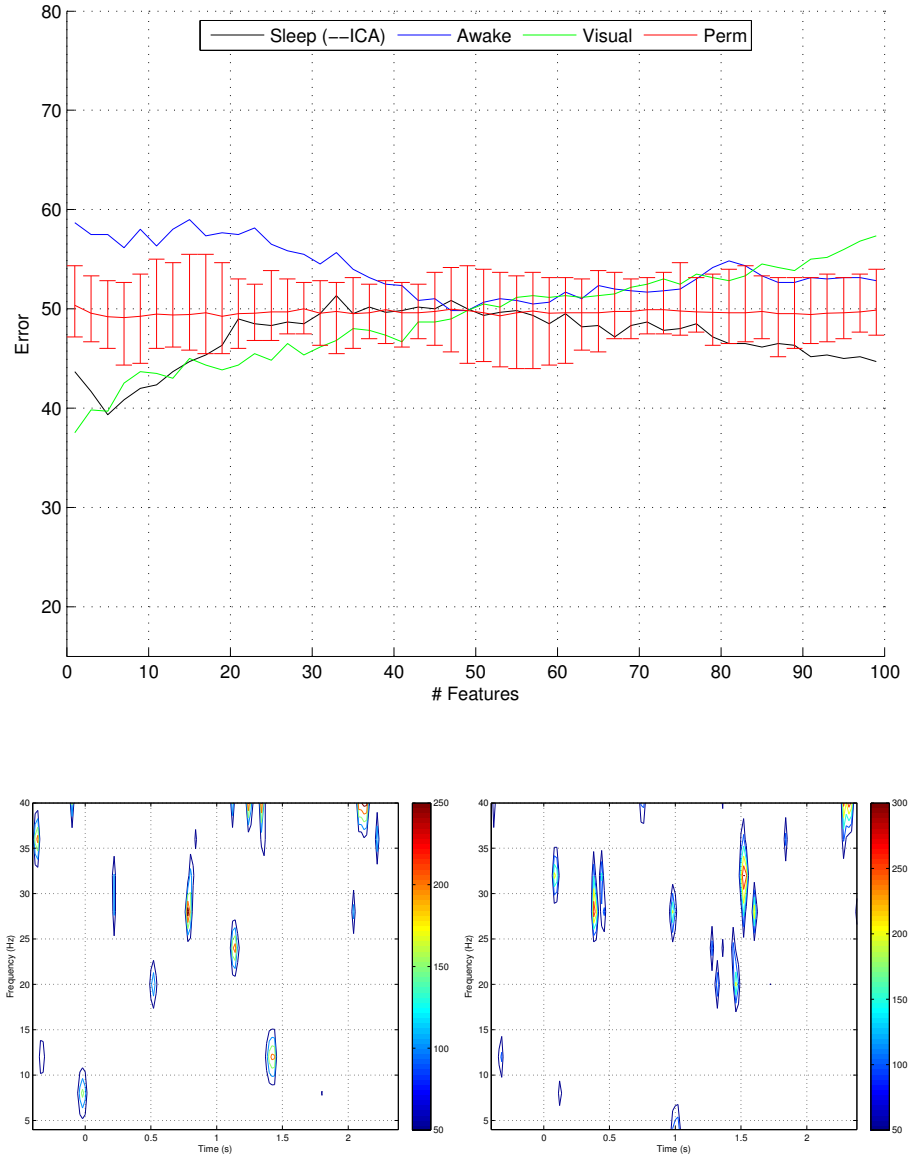
Figure A.10: Subject 138. Top 50 time-frequency features selected during each leave-two-out iteration for subject 138. The number of times a feature is selected gives an indication of how important the single features are.





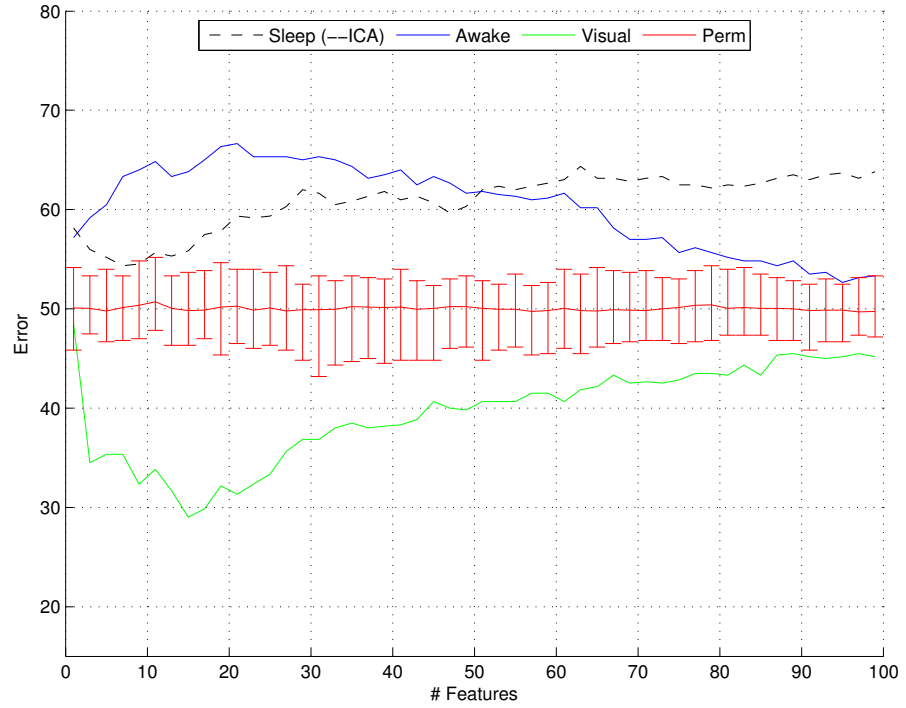
(a) Electrode C3 **(b)** Electrode C4

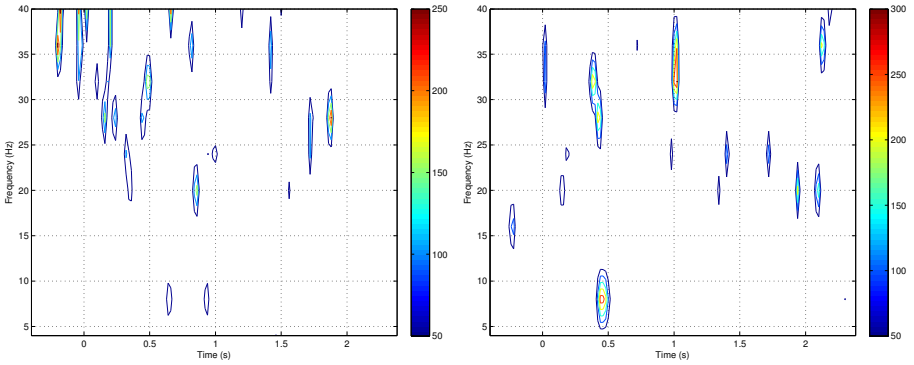
Figure A.11: Subject 144. Top 50 time-frequency features selected during each leave-two-out iteration for subject 144. The number of times a feature is selected gives an indication of how important the single features are.



(a) Electrode C3 **(b)** Electrode C4

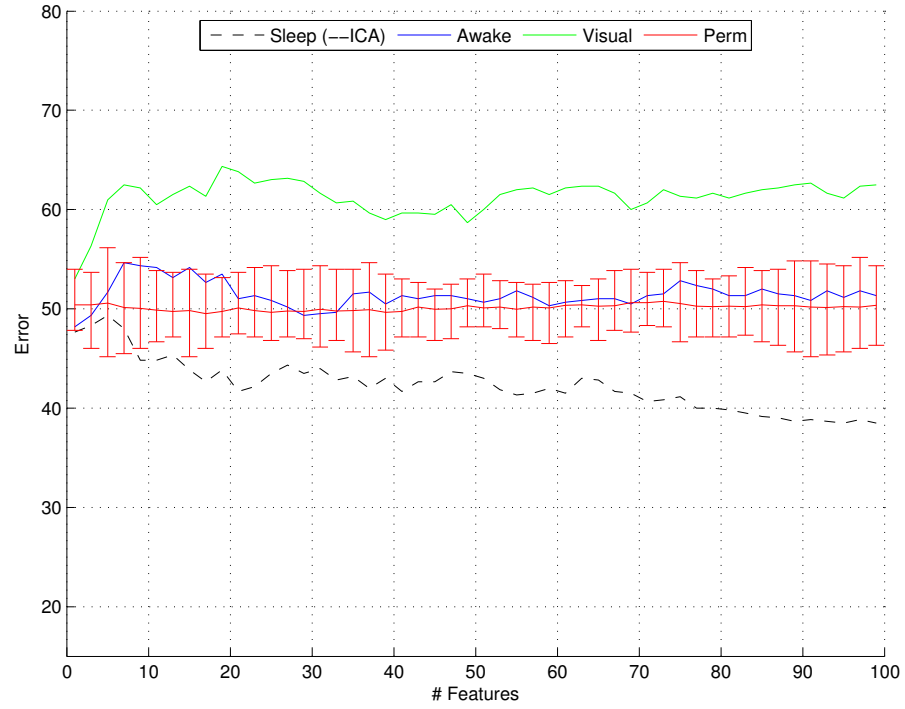
Figure A.12: Subject 147. Top 50 time-frequency features selected during each leave-two-out iteration for subject 147. The number of times a feature is selected gives an indication of how important the single features are.

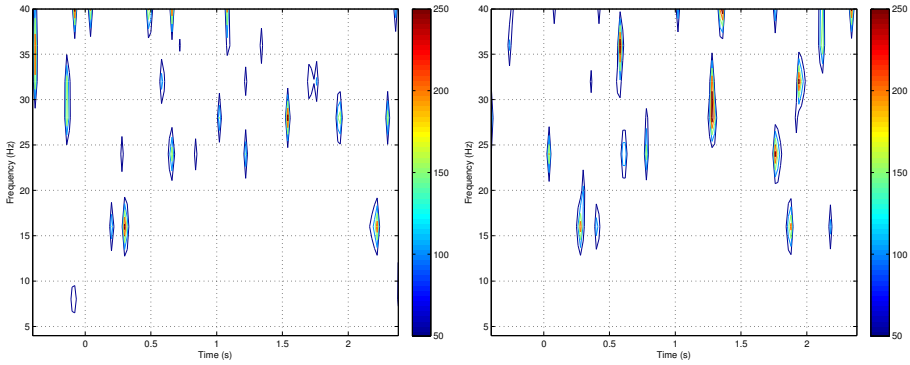




(a) Electrode C3 **(b)** Electrode C4

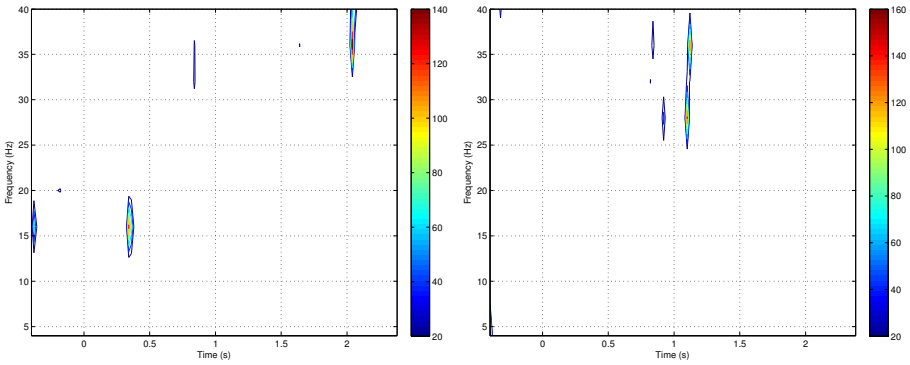
Figure A.13: Subject 149. Top 50 time-frequency features selected during each leave-two-out iteration for subject 149. The number of times a feature is selected gives an indication of how important the single features are.





(a) Electrode C3 **(b)** Electrode C4

Figure A.14: Subject 150. Top 50 time-frequency features selected during each leave-two-out iteration for subject 150. The number of times a feature is selected gives an indication of how important the single features are.

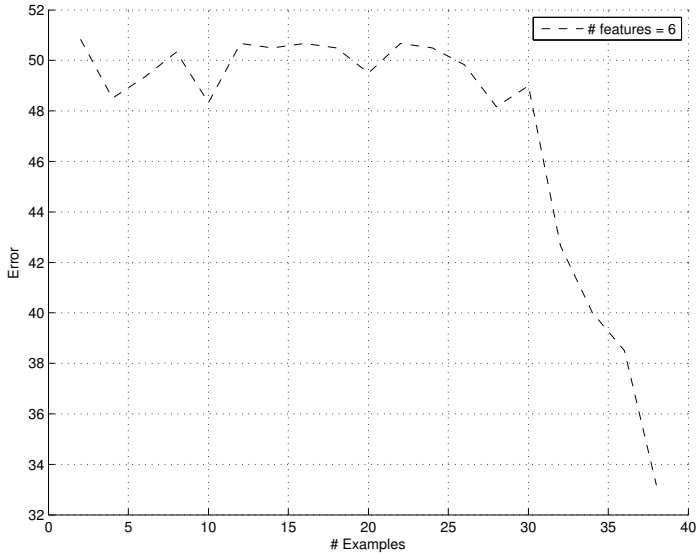


(a) Electrode C3 **(b)** Electrode C4

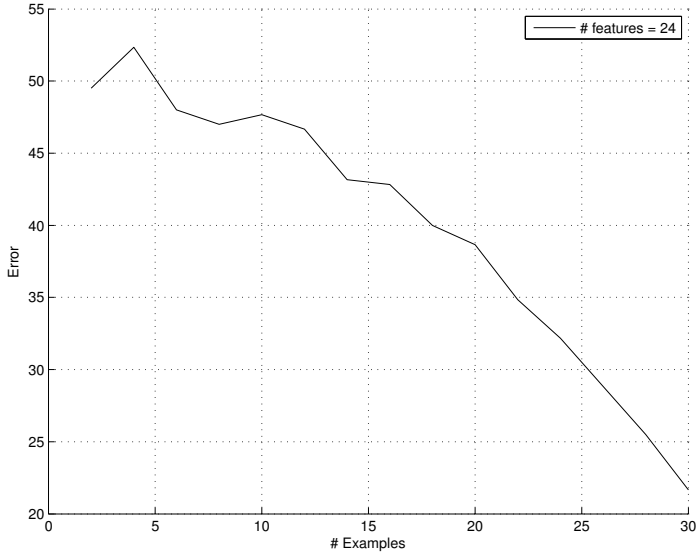
Figure A.15: Subject 134. Top 6 time-frequency features selected during each leave-two-out iteration for subject 134.

Appendix B

Learning Curves

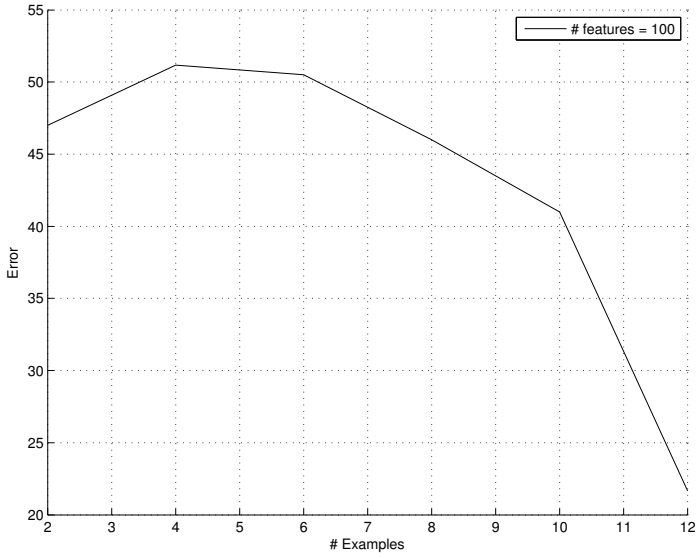


(a) Learning curve for subject 105.

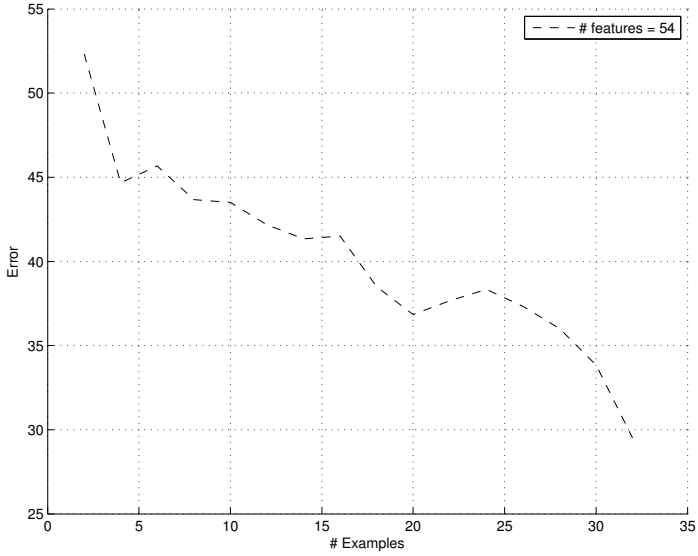


(b) Learning curve for subject 107.

Figure B.1: Learning curves (dashed line indicates that artifact rejected data is used).

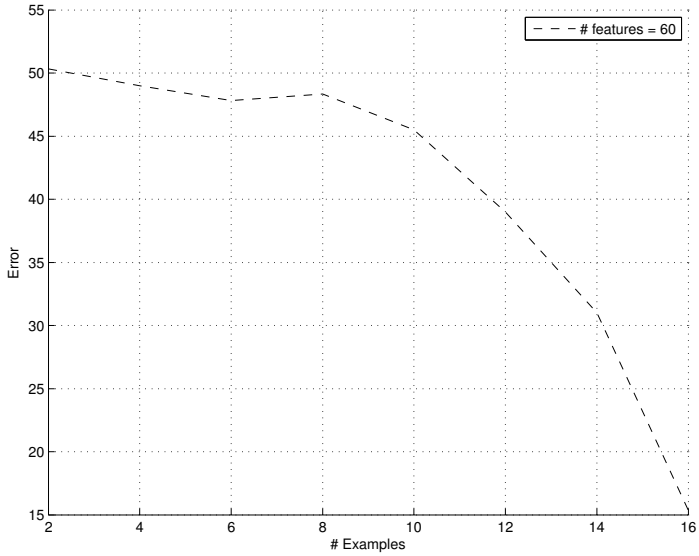


(a) Learning curve for subject 111.

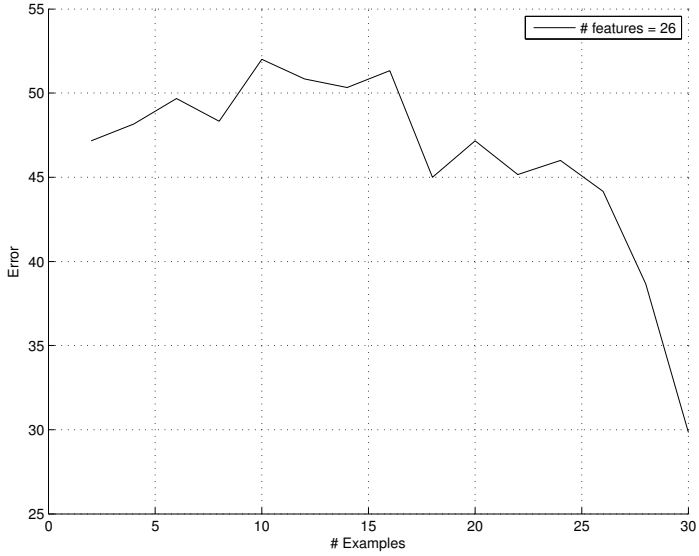


(b) Learning curve for subject 113.

Figure B.2: Learning curves (dashed line indicates that artifact rejected data is used).

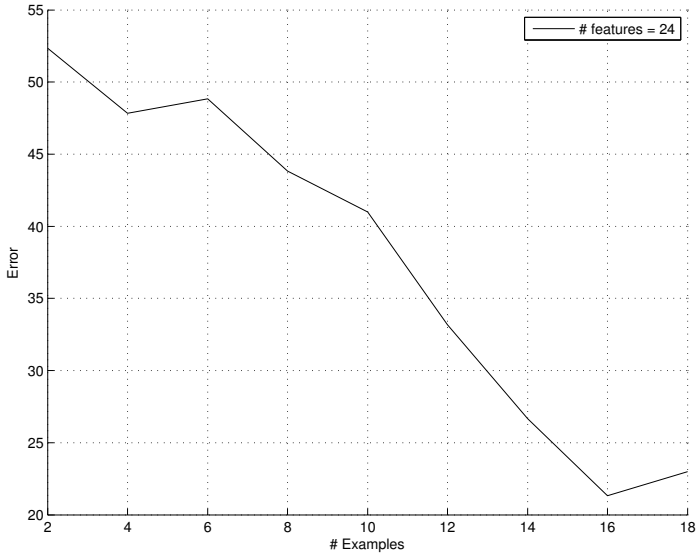


(a) Learning curve for subject 117.

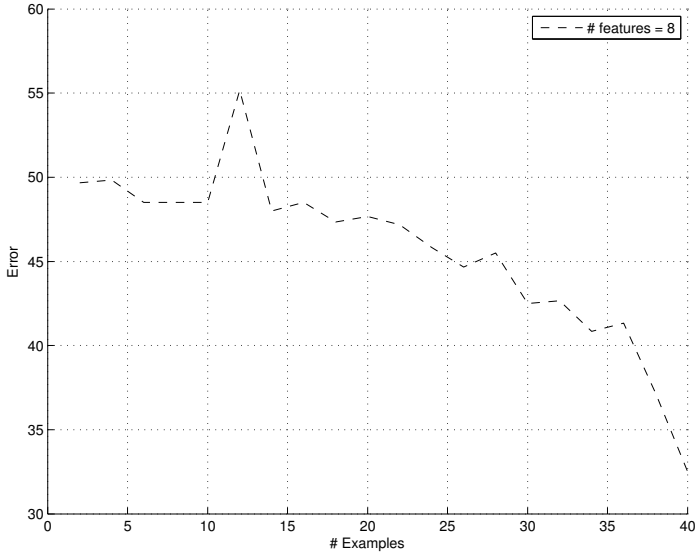


(b) Learning curve for subject 122.

Figure B.3: Learning curves (dashed line indicates that artifact rejected data is used).

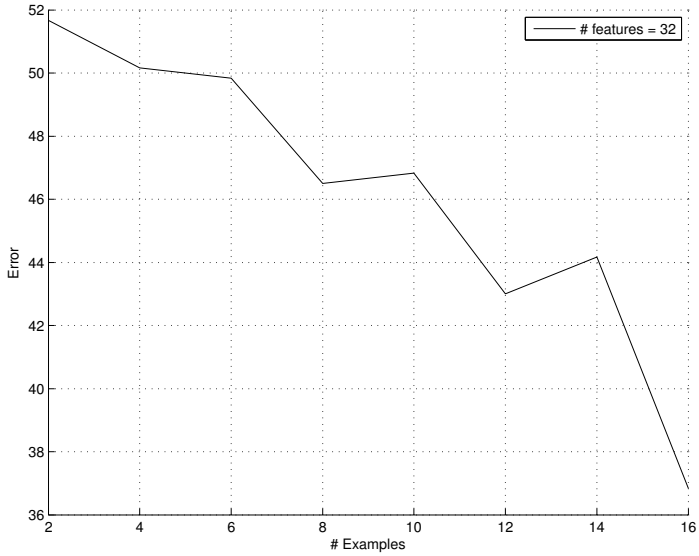


(a) Learning curve for subject 129.

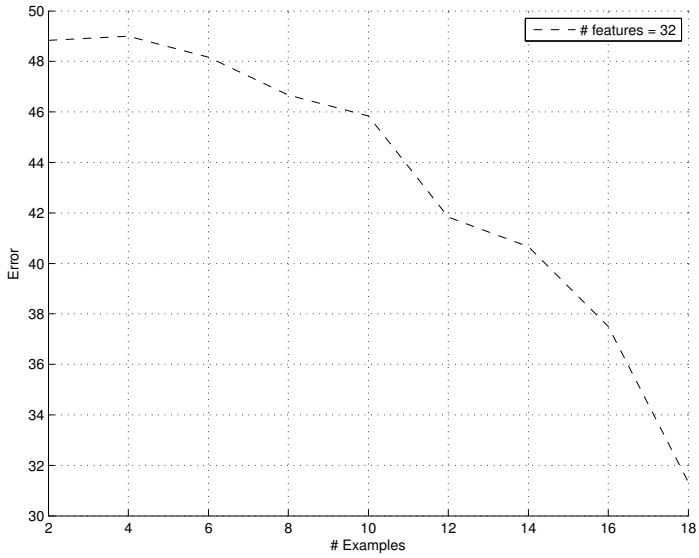


(b) Learning curve for subject 134.

Figure B.4: Learning curves (dashed line indicates that artifact rejected data is used).

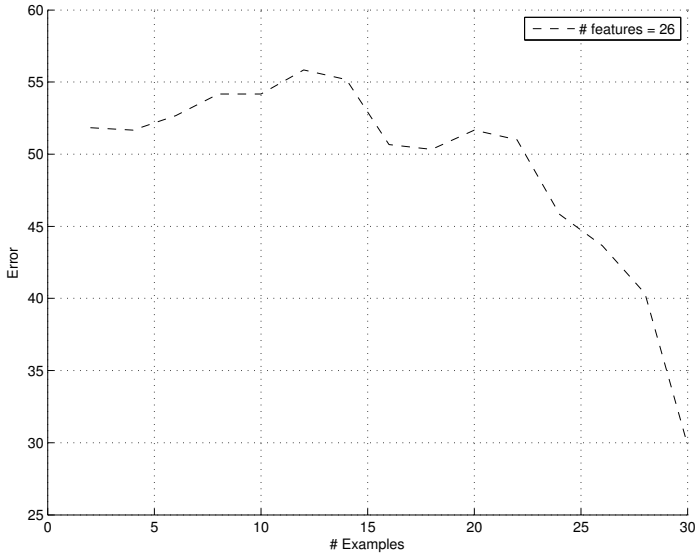


(a) Learning curve for subject 138.

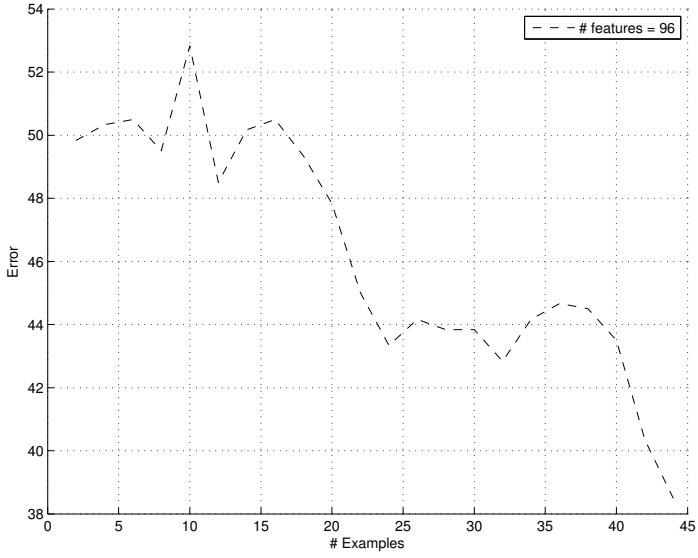


(b) Learning curve for subject 139.

Figure B.5: Learning curves (dashed line indicates that artifact rejected data is used).



(a) Learning curve for subject 144.

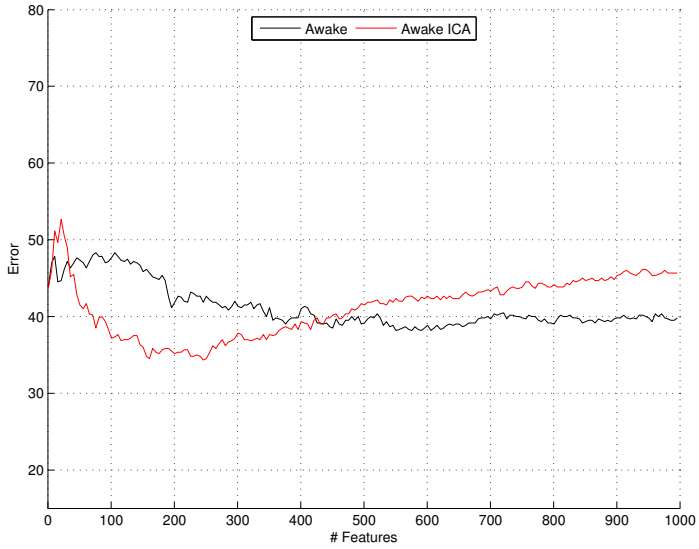


(b) Learning curve for subject 150.

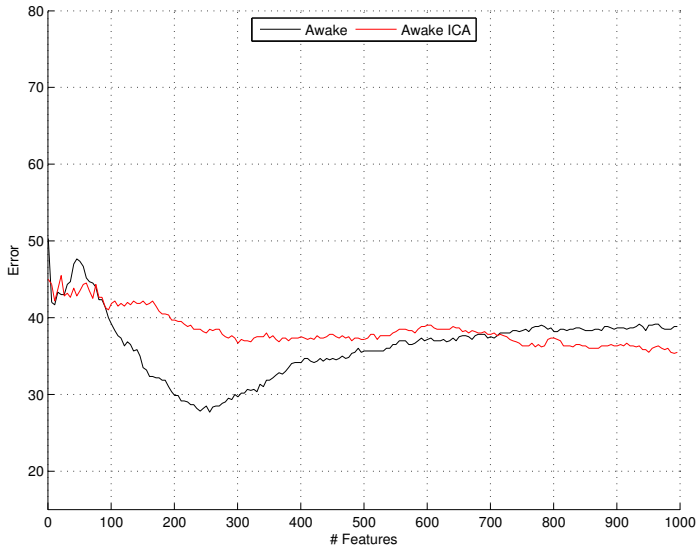
Figure B.6: Learning curves (dashed line indicates that artifact rejected data is used).

Appendix C

Awake Classifier Performance

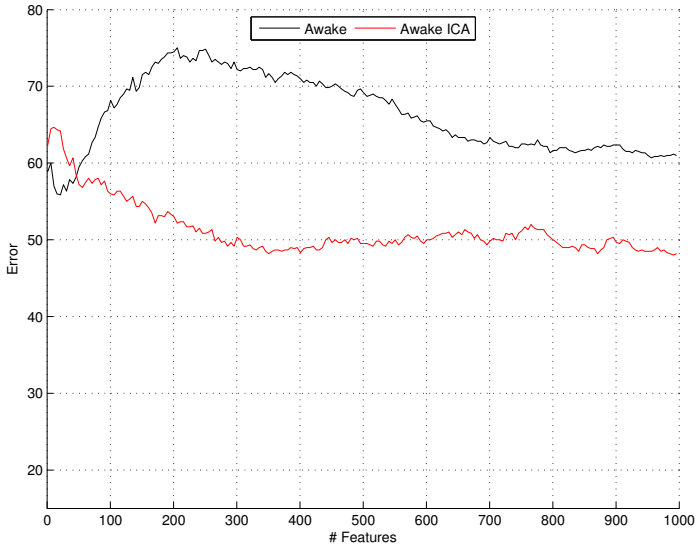


(a) Subject 113.

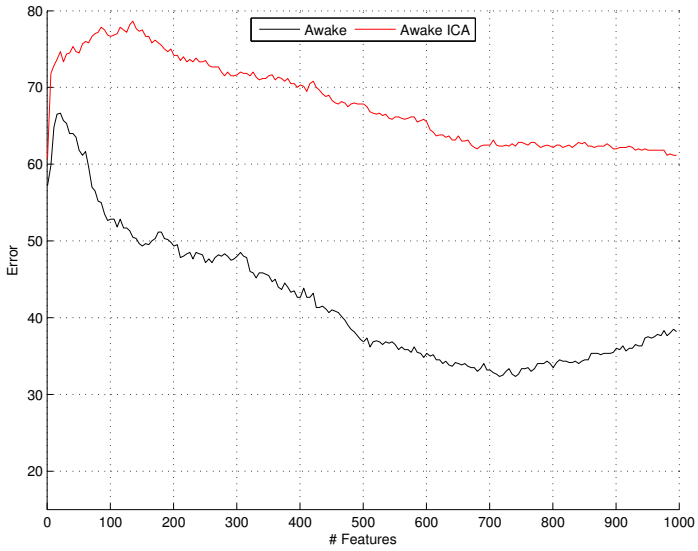


(b) Subject 118.

Figure C.1: Classifier performance with up to 1,000 features included in the model.



(a) Subject 137.



(b) Subject 149.

Figure C.2: Classifier performance with up to 1,000 features included in the model.

Bibliography

[Andrew and Pfurtscheller, 1996] Andrew, C. and Pfurtscheller, G. (1996). Event-related coherence as a tool for studying dynamic interaction of brain regions. *Electroencephalography and clinical Neurophysiology*, 98(2):144–148.

[Antony et al., 2012] Antony, J., Gobel, E., O'Hare, J., Reber, P., and Paller, K. (2012). Cued memory reactivation during sleep influences skill learning. *Nature Neuroscience*.

[Aserinsky and Kleitman, 1953] Aserinsky, E. and Kleitman, N. (1953). Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science*, 118(3062):273–274.

[Ashburner et al., 2008] Ashburner, J., Chen, C., Flandin, G., Henson, R., Kiebel, S., Kilner, J., Litvak, V., Moran, R., Penny, W., Stephan, K., et al. (2008). Spm8 manual. *Functional Imaging Laboratory, Institute of Neurology*.

[Atienza et al., 2001] Atienza, M., Cantero, J., and Escera, C. (2001). Auditory information processing during human sleep as revealed by event-related brain potentials. *Clinical Neurophysiology*, 112(11):2031–2045.

[Bastuji et al., 2002] Bastuji, H., Perrin, F., and Garcia-Larrea, L. (2002). Semantic analysis of auditory input during sleep: studies with event related potentials. *International Journal of Psychophysiology*, 46(3):243–255.

[Bear et al., 2007] Bear, M., Connors, B., and Paradiso, M. (2007). *Neuroscience: Exploring the brain*. Lippincott Williams & Wilkins.

[Berger, 1929] Berger, H. (1929). Über das Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87:527–570.

[Bishop, 2007] Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition.

[Blankertz et al., 2004] Blankertz, B., Muller, K.-R., Curio, G., Vaughan, T. M., Schalk, G., Wolpaw, J. R., Schlogl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schroder, M., and Birbaumer, N. (2004). The bci competition 2003: Progress and perspectives in detection and discrimination of eeg single trials. *IEEE Transactions on Biomedical Engineering*, 51(6).

[Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). Training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152.

[Brown and Singer, 1993] Brown, A. M. and Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, 55.

[Brualla et al., 1998] Brualla, J., Romero, M., Serrano, M., and Valdizán, J. (1998). Auditory event-related potentials to semantic priming during sleep. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 108(3):283–290.

[Bruck et al., 2009] Bruck, D., Ball, M., Thomas, I., and Rouillard, V. (2009). How does the pitch and pattern of a signal affect auditory arousal thresholds? *Journal of Sleep Research*, 18(2):196–203.

[Burton et al., 1988] Burton, S., Harsh, J., and Badia, P. (1988). Cognitive activity in sleep and responsiveness to external stimuli. *Sleep: Journal of Sleep Research & Sleep Medicine*.

[Campos Viola et al., 2009] Campos Viola, F., Thorne, J., Edmonds, B., Schneider, T., Eichele, T., and Debener, S. (2009). Semi-automatic identification of independent components representing eeg artifact. *Clinical Neurophysiology*, 120(5):868–877.

[Caputo et al., 2002] Caputo, B., Sim, K., Furesjo, F., and Smola, A. (2002). Appearance-based object recognition using svms: Which kernel should i use? In *Proc of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision, Whistler*, volume 2002.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

[Chapelle et al., 2005] Chapelle, O., Smola, A., Ong, C. S., and Kowalczyk, A. (2005). Data mining unlearnable data sets. Patent Application Publication, US20080027886.

[Chen et al., 2006] Chen, P., Fan, R., and Lin, C. (2006). A study on smo-type decomposition methods for support vector machines. *Neural Networks, IEEE Transactions on*, 17(4):893–908.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

[Crone et al., 1998] Crone, N. E., Miglioretti, D. L., Gordon, B., Sieracki, J. M., Wilson, M. T., Uematsu, S., and Lesser, R. P. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis: I. alpha and beta event-related desynchronization. *Brain*, 121(12):2271–2299.

[Cruse et al., 2011] Cruse, D., Chennu, S., Chatelle, C., Bekinschtein, T., Fernández-Espejo, D., Pickard, J., Laureys, S., and Owen, A. (2011). Bedside detection of awareness in the vegetative state: a cohort study. *The Lancet*.

[Cruse et al., 2012] Cruse, D., Chennu, S., Chatelle, C., Fernández-Espejo, D., Bekinschtein, T., Pickard, J., Laureys, S., and Owen, A. (2012). Relationship between etiology and covert cognition in the minimally conscious state. *Neurology*.

[Delorme and Makeig, 2004] Delorme, A. and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21.

[Donner et al., 2009] Donner, T. H., Siegel, M., Fries, P., and Engel, A. K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, 19(18):1581–1585.

[Edeline et al., 2000] Edeline, J., Manunta, Y., and Hennevin, E. (2000). Auditory thalamus neurons during sleep: changes in frequency selectivity, threshold, and receptive field size. *Journal of neurophysiology*, 84(2):934–952.

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC.

[Emmons and Simon, 1956] Emmons, W. and Simon, C. (1956). The non-recall of material presented during sleep. *American Journal of Psychology*, 69(1):76–81.

[Fan et al., 2005] Fan, R.-E., Chen, P.-H., and Lin, C.-J. (2005). Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, 6:1889–1918.

[Federmeier and Kutas, 2009] Federmeier, M. and Kutas, K. D. (2009). N400. *Scholarpedia*, 4(10):7790.

[Formby, 1967] Formby, D. (1967). Maternal recognition of infant's cry. *Developmental medicine and child neurology.*, 9(3):293–298.

[Golland and Fischl, 2003] Golland, P. and Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. In *Information processing in medical imaging*, pages 330–341. Springer.

[Golland et al., 2005] Golland, P., Liang, F., Mukherjee, S., and Panchenko, D. (2005). Permutation tests for classification. *Learning Theory*, pages 36–39.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

[Guyon et al., 2004] Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2004). Result analysis of the nips 2003 feature selection challenge. *Advances in Neural Information Processing Systems*, 17:545–552.

[Guyon et al., 1998] Guyon, I., Makhoul, J., Schwartz, R., and Vapnik, V. (1998). What size test set gives good error rate estimates? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):52–64.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

[Guyon et al., 2006] Guyon, I. M., Gunn, S. R., Nikravesh, M., and Zadeh, L. (2006). *Feature Extraction, Foundations and Applications*. Springer.

[Hall et al., 2005] Hall, P., Marron, J., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444.

[Halperin and Iorio, 1981] Halperin, J. M. and Iorio, L. C. (1981). Responsivity of rats to neutral and danger-signaling stimuli during sleep. *Behavioral and Neural Biology*, 33(2):213–219.

[Haussler, 1999] Haussler, D. (1999). Convolution Kernels on Discrete Structures. Technical report, University of California at Santa Cruz.

[Haynes and Rees, 2006] Haynes, J. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534.

[Hennevin et al., 2007] Hennevin, E., Huetz, C., and Edeline, J. (2007). Neural representations during sleep: from sensory processing to memory traces. *Neurobiology of learning and memory*, 87(3):416–440.

[Holmes et al., 1996] Holmes, A., Blair, R., Watson, G., and Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, 16(1):7–22.

[Hoogenboom et al., 2006] Hoogenboom, N., Schoffelen, J., Oostenveld, R., Parkes, L., and Fries, P. (2006). Localizing human visual gamma-band activity in frequency, time and space. *Neuroimage*, 29(3):764–773.

[Hsu and Lin, 2002] Hsu, C.-W. and Lin, C.-J. (2002). A simple decomposition method for support vector machines. *Mach. Learn.*, 46(1-3):291–314.

[Huxley, 1932] Huxley, A. (1932). *Brave New World*. Chatto & Windus.

[Ibáñez et al., 2008] Ibáñez, A., San Martín, R., Hurtado, E., and López, V. (2008). Methodological considerations related to sleep paradigm using event related potentials. *Biological Research*, 41(3):271–275.

[Iber, 2007] Iber, C. (2007). *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine.

[Joachims, 1999] Joachims, T. (1999). Making large-scale support vector machine learning practical. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in kernel methods*, pages 169–184. MIT Press, Cambridge, MA, USA.

[Johns et al., 1991] Johns, M. et al. (1991). A new method for measuring daytime sleepiness: the epworth sleepiness scale. *sleep*, 14(6):540–545.

[Keerthi and Gilbert, 2002] Keerthi, S. and Gilbert, E. (2002). Convergence of a generalized smo algorithm for svm classifier design. *Machine Learning*, 46(1):351–360.

[Keerthi and Lin, 2003] Keerthi, S. and Lin, C. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural computation*, 15(7):1667–1689.

[Keerthi et al., 2001] Keerthi, S., Shevade, S., Bhattacharyya, C., and Murthy, K. (2001). Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649.

[Klement et al., 2008] Klement, S., Madany Mamlouk, A., and Martinetz, T. (2008). Reliability of cross-validation for svms in high-dimensional, low sample size scenarios. *Artificial Neural Networks-ICANN 2008*, pages 41–50.

[Kouider and Dehaene, 2007] Kouider, S. and Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):857–875.

[Kowalczyk, 2007] Kowalczyk, A. (2007). Classification of anti-learnable biological and synthetic data. *Knowledge Discovery in Databases: PKDD 2007*, pages 176–187.

[Kowalczyk and Chapelle, 2005] Kowalczyk, A. and Chapelle, O. (2005). An analysis of the anti-learning phenomenon for the class symmetric polyhedron. In *Algorithmic Learning Theory*, pages 78–91. Springer.

[Kowalczyk et al., 2007] Kowalczyk, A., Greenawalt, D., Bedo, J., Duong, C., Raskutti, G., Thomas, R., and Phillips, W. (2007). Validation of anti-learnable signature in classification of response to chemoradiotherapy in esophageal adenocarcinoma patients. In *Proc. Intern. Symp. on Optimization and Systems Biology, OSB (to appear*. Citeseer.

[Kramer et al., 1982] Kramer, M., Kinney, L., and Scharf, M. (1982). Dream incorporation and dream function. *Sleep*, 82:369–371.

[Kutas and Federmeier, 2011] Kutas, M. and Federmeier, K. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647.

[Lal et al., 2004] Lal, T., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., and Scholkopf, B. (2004). Support vector channel selection in bci. *Biomedical Engineering, IEEE Transactions on*, 51(6):1003–1010.

[Lotte et al., 2007] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1.

[Manganotti et al., 2004] Manganotti, P., Fuggetta, G., and Fiaschi, A. (2004). Changes of motor cortical excitability in human subjects from wakefulness to early stages of sleep: a combined transcranial magnetic stimulation and electroencephalographic study. *Neuroscience letters*, 362(1):31–34.

[Maquet et al., 2000] Maquet, P. et al. (2000). Functional neuroimaging of normal human sleep by positron emission tomography. *Journal of Sleep Research*, 9(3):207–232.

[Mercer, 1909] Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Royal Soc. (A)*, 83(559):69–70.

[Mitra and Pesaran, 1999] Mitra, P. and Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophysical journal*, 76(2):691–708.

[Morash et al., 2008] Morash, V., Bai, O., Furlani, S., Lin, P., and Hallett, M. (2008). Classifying eeg signals preceding right hand, left hand, tongue, and right foot movements and motor imageries. *Clinical Neurophysiology*, 119(11):2570–2578.

[Mørup et al., 2007] Mørup, M., Hansen, L., and Arnfred, S. (2007). Erpwavelab: A toolbox for multi-channel analysis of time–frequency transformed event related potentials. *Journal of neuroscience methods*, 161(2):361–368.

[Mørup et al., 2008] Mørup, M., Hansen, L., Arnfred, S., Lim, L., and Madsen, K. (2008). Shift-invariant multilinear decomposition of neuroimaging data. *NeuroImage*, 42(4):1439–1450.

[Muzur et al., 2002] Muzur, A., Pace-Schott, E., and Hobson, J. (2002). The prefrontal cortex in sleep. *Trends in Cognitive Sciences*, 6(11):475–481.

[Nocedal and Wright, 1999] Nocedal, J. and Wright, S. (1999). *Numerical optimization*. Springer.

[Nofzinger et al., 2002] Nofzinger, E., Buysse, D., Miewald, J., Meltzer, C., Price, J., Sembrat, R., Ombao, H., Reynolds, C., Monk, T., Hall, M., et al. (2002). Human regional cerebral glucose metabolism during non-rapid eye movement sleep in relation to waking. *Brain*, 125(5):1105–1115.

[Norman et al., 2006] Norman, K., Polyn, S., Detre, G., and Haxby, J. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430.

[Nunez and Srinivasan, 2006] Nunez, P. and Srinivasan, R. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA.

[Oostenveld et al., 2011] Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. (2011). Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011:1.

[Oswald et al., 1960] Oswald, I., Taylor, A., and Treisman, M. (1960). Discriminative responses to stimulation during human sleep. *Brain*, 83(3):440–&.

[Percival and Walden, 1993] Percival, D. and Walden, A. (1993). *Spectral analysis for physical applications: multitaper and conventional univariate techniques*. Cambridge University Press.

[Pereira et al., 2009] Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209.

[Perrin, 2004] Perrin, F. (2004). Auditory event-related potentials studies of information processing during human sleep. *Psychologica belgica*, 44:43–58.

[Perrin et al., 1999] Perrin, F., García-Larrea, L., Mauguičre, F., and Bastuji, H. (1999). A differential brain response to the subject's own name persists during sleep. *Clinical Neurophysiology*, 110(12):2153–2164.

[Pfurtscheller, 1977] Pfurtscheller, G. (1977). Graphical display and statistical evaluation of event-related desynchronization (erd). *Electroencephalography and Clinical Neurophysiology*, 43(5):757–760.

[Pfurtscheller, 1992] Pfurtscheller, G. (1992). Event-related synchronization (ers) - an electrophysiological correlate of cortical areas at rest. *Electroencephalography and Clinical Neurophysiology*, 83(1):62–69.

[Pfurtscheller et al., 2006] Pfurtscheller, G., Brunner, C., Schlögl, A., and Lopes da Silva, F. H. (2006). Mu rhythm (de)synchronization and eeg single-trial classification of different motor imagery tasks. *Neuroimage*, 31(1):153–159.

[Pfurtscheller and Lopes da Silva, 1999] Pfurtscheller, G. and Lopes da Silva, F. H. (1999). Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857.

[Pfurtscheller et al., 1993] Pfurtscheller, G., Neuper, C., and Kalcher, J. (1993). 40-hz oscillations during motor behavior in man. *Neuroscience letters*, 164(1-2):179–182.

[Platt, 1998] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances in Kernel Methods - Support Vector Learning.

[Polyanin and Manzhirov, 2008] Polyanin, A. D. and Manzhirov, A. V. (2008). *Handbook of Integral Equations*. Chapman & Hall/CRC Press, 2 edition.

[Radovanović et al., 2010] Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *The Journal of Machine Learning Research*, 9999:2487–2531.

[Raghavachari et al., 2001] Raghavachari, S., Kahana, M., Rizzuto, D., Caplan, J., Kirschen, M., Bourgeois, B., Madsen, J., and Lisman, J. (2001). Gating of human theta oscillations by a working memory task. *The journal of Neuroscience*, 21(9):3175–3183.

[Rodenbeck et al., 2006] Rodenbeck, A., Binder, R., Geisler, P., Danker-Hopfe, H., Lund, R., Raschke, F., Weeß, H., and Schulz, H. (2006). A review of sleep eeg patterns. part i: A compilation of amended rules for their visual recognition according to rechtschaffen and kales. *Somnologie*, 10(4):159–175.

[Roth et al., 1956] Roth, M., Shaw, J., and Green, J. (1956). The form, voltage distribution and physiological significance of the k-complex. *Electroencephalography and Clinical Neurophysiology*, 8(3):385–402.

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. The MIT Press, 1st edition.

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge Univ Pr.

[Strother et al., 2002] Strother, S., Anderson, J., Hansen, L., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The npairs data analysis framework. *NeuroImage*, 15(4):747–771.

[Tangermann, 2007] Tangermann, M. (2007). *Feature Selection for Brain-Computer Interfaces*. PhD thesis, Universitätsbibliothek Tübingen.

[Tangermann et al., 2012] Tangermann, M., Müller, K., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K., Müller-Putz, G., et al. (2012). Review of the bci competition iv. *Frontiers in Neuroscience*, 6.

[van Vugt et al., 2007] van Vugt, M., Sederberg, P., and Kahana, M. (2007). Comparison of spectral analysis methods for characterizing brain oscillations. *Journal of neuroscience methods*, 162(1):49–63.

[Vapnik, 1998] Vapnik, V. N. (1998). *Statistical learning theory*. Wiley, 1 edition.

[Vapnik, 2000] Vapnik, V. N. (2000). *The nature of statistical learning theory*. Springer-Verlag New York Inc.

[Vapnik, 2006] Vapnik, V. N. (2006). *Estimation of dependences based on empirical data*. Springer-Verlag New York Inc.