SINGLE TRIAL ANALYSES OF ENCEPHALOGRAM DATA

vorgelegt von
Diplom-Mathematiker

Steven Lemm

von der Fakultät IV-Elektrotechnik und Informatik
der Technischen Universität Berlin

zur Erlangung des akdemischen Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)
genehmigte Dissertation

Promotionsausschuss

Vorsitzender:	Prof. Dr. Manfred Opper
Berichter:	Prof. Dr. Klaus-Robert Müller
Berichter:	Prof. Dr. Gabriel Curio
Berichter:	Univ.-Prof. DI Dr.techn. Gert Pfurtscheller

Tag der wissenschaftlichen Aussprache 29. November 2007

Berlin 2007
D83

Brains are able to process environmental stimuli in a 'single-trial' mode. Accordingly, human cortical neurophysiology should be approached describing 'single-trial' behavior as well.

Thesis advisor
Prof. Dr. Klaus-Robert Müller

Author
Steven Lemm

SINGLE TRIAL ANALYSES OF ENCEPHALOGRAM DATA

Abstract

In this thesis, inspired by the development of the Brain-computer-interface (BCI) technology, we present novel methods for the analysis of macroscopically recorded brain signals. Here the focus is put on improved feature extraction methods, the detection of mental states and the analysis of variability of brain responses.

Conditional event-related (de-)synchronization The fluctuation of signal power in a narrow band induced by an event is conventionally termed event-related (de-)synchronization and is quantified as the relative deviation from the mean baseline activity. We extend the ERD terminology with respect to a generalized reference. To this end, we oppose the time course of the event-related activity against those obtained from single trials without specific stimulus processing. From this generalized approach we derive a method to determine the dependencies of the ERD response on initial cortical states. A comparative study of surrogate and real ERD data validates this approach.

Spatio–spectral filters The common-spatial-pattern algorithm (CSP) determines optimally discriminative spatial filters from multivariate broad-band signals. We extend the conventional algorithm such that it additionally obtains simple frequency filters. This enables adaptation to the individual characteristics of the power spectrum and thus improves feature extraction. An empirical comparison with the conventional CSP method reveals the advantages of our approach in the context of the classification of imaginary unilateral hand movements.

Extraction of event-related potentials (ERP) Independent component analysis (ICA) is a tool for statistical data analysis that is able to linearly decompose multivariate signals into their underlying source components. We present an ICA method that uses prior knowledge about the phase-locked property of ERPs for their improved extraction from single trial EEG. The application on artificially generated and real world data validates this approach in terms of an improved signal-to-noise ratio of the extracted ERPs.

Adaptive feature combination across time Lateralized μ-rhythm ERD and lateralized movement-related potentials are the most commonly used discriminative features for the classification of imaginary hand movements. In the context of real time classification we present a method that efficiently combines these temporally differently accentuated features. To this end, we first train weak classifiers for each time instance and each feature separately. Subsequently we combine these weak classifiers in a strictly causal, probabilistic manner. The effectiveness of this approach was proven by its successful application to data from the international BCI competitions in 2003 and 2005.

Betreuer
Prof. Dr. Klaus-Robert Müller

Autor
Steven Lemm

SINGLE TRIAL ANALYSES OF ENCEPHALOGRAM DATA

Zusammenfassung

Inspiriert, nicht zuletzt durch die Erforschung der Brain-computer-interface (BCI) Technologie präsentieren wir in dieser Dissertation neue Methoden zur Analyse makroskopisch gemessener Hirnsignale. Der Fokus liegt hierbei auf Methoden zur verbesserten Merkmalsextraktion, der Detektion mentaler Zustände und der Analyse der Variabilität von Reizantworten.

Bedingte ereigniskorrelierte (De-)Synchronization Die durch ein Ereignis induzierte Leistungsschwankung in einem Frequenzband wird konventionell als ERD bezeichnet und als relative Veränderung gegenüber der mittleren Grundaktivität gemessen. Wir erweitern den ERD-Begriff in Bezug auf eine verallgemeinerte Referenz. Dafür kontrastieren wir den zeitlichen Verlauf ereigniskorrelierter Aktivität mit dem gemessener *single trials* ohne spezifische Reizverarbeitung. Aus diesem verallgemeinerten Ansatz leiten wir eine Methode zur Bestimmung der Abhängigkeit der ERD-Antwort von initialen kortikalen Zuständen ab. Vergleichende Untersuchungen auf künstlichen und realen Daten validieren diesen Ansatz.

Räumlich-spektrale Filter Der Common-Spatial-Pattern Algorithmus (CSP) bestimmt für multivariate breitbandige Signale diskriminative räumliche Filter. Wir erweitern den klassischen Ansatz, so dass zusätzlich eine Optimierung einfacher Frequenzfilter erfolgt. Dies ermöglicht eine Adaptation an das individuelle EEG-Frequenzspektrums und somit eine verbesserte Merkmalsextraktion. Ein empirischer Vergleich mit dem klassischen CSP Algorithmus belegt die Vorteile unseres Verfahrens im Kontext der Klassifikation vorgestellter unilateraler Handbewegungen.

Extraktion ereigniskorrelierter Potenziale (EKP) *Independent component analysis* (ICA) ist ein Werkzeug der statistischen Datenanalyse und Signalverarbeitung, welches multivariate Signale linear in ihre Quellkomponenten zerlegen kann. Wir präsentieren eine ICA Methode zur Extraktion von *single trial* EKP, welche unter Ausnutzung der Phasengebundenheit des EKP verbesserte räumliche Filter bestimmt. Simulationen mit künstlichen und echten Daten validieren diesen Ansatz in Bezug auf ein verbessertes SNR der extrahierten EKP.

Zeitlich adaptive Merkmalskombination Lateralisierte ERD des μ-Rhythmus und bewegungskorrelierte Potenziale sind die gebräuchlichsten diskriminativen Merkmale zur Klassifikation vorgestellter Handbewegungen. Wir präsentieren eine Methode diese zeitlich unterschiedlich ausgeprägten Merkmale für die Echtzeit-Klassifikation zu verwenden. Hierzu trainieren wir zunächst separat zu jedem Zeitpunkt einfache Klassifikatoren für jedes Merkmal und kombinieren diese anschließend adaptive in einem strikt kausalen, probabilistischen Ansatz. Die Leistungsfähigkeit dieses Algorithmus wurde durch seine erfolgreiche Anwendung in den BCI-Wettbewerbe 2003 und 2005 zur Klassifikation vorgestellter unilateraler Handbewegungen nachgewiesen.

Citations to Previously Published Work

Large portions of Chapters 4 have appeared in the following paper:

> "Spatio-Spectral Filters for Improving the Classification of Single Trial EEG", S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, IEEE Transactions on Biomedical Engineering, 52(9), pp. 1541–48, Sep. 2005,

Most of Chapter 5 has been published as:

> "Enhancing the Signal to Noise Ratio of ICA-Based Extracted ERPs", S. Lemm, G. Curio, Y. Hlushchuk, and K.-R. Müller, IEEE Transactions on Biomedical Engineering, 53(4), pp. 601–7, April 2006.

Finally, the methods of Chapter 6 have been published in the following:

> "Probabilistic Modelling of Sensorimotor μ-Rhythms for Classification of Imaginary Hand Movements", S. Lemm, C. Schäfer, and G. Curio, IEEE Transactions on Biomedical Engineering, 51(6), pp. 1077–80, June 2004.
>
> "Aggregating Classification Accuracy across Time: Application to Single Trial EEG", S. Lemm, and G. Curio, In Advances in Neural Information Processing Systems (NIPS 06), volume 19, 2007.

Electronic preprints of the IEEE publications are available via the Internet at the following URL: `http://ieeexplore.ieee.org`.
Please refer to `http://books.nips.cc` for electronic versions of the NIPS publications

Acknowledgments

Writing this thesis would not have been possible without the invaluable help and support from others, let it be financial, intellectual or personal. First of all I would like to thank my supervisors Prof. Dr. Klaus-Robert Müller and Prof. Dr. Gabriel Curio. Without their encouragement, this thesis would not have been written. I appreciate in particular Prof. Müllers deep scientific knowledge and his sustained scientific curiosity and enthusiasm. Likewise I owe gratitude to Prof. Dr. Gabriel Curio who guided my way through the field of neurophysiology and pointed me at many interesting research questions. Moreover, I enjoyed several occasions of getting acquainted with astrophysical issues.

This dissertation has been done at the "Intelligent data analysis" (IDA) group at the Fraunhofer institute FIRST in Berlin, Germany. I would like to thank all people at this institute and in particular all current and former members of our group for creating a pleasant and open-minded research atmosphere, including A. Schwaighofer, F. Meinecke, A. Ziehe, P. Buenau, S. Harmeling, J. Laub, S. Sonnenburg, M. Kawanabe, R. Tomioka, P. Laskov, G. Blanchard, G. Nolte, K. Rieck. Very cordial thanks go to my former roommates Sebastian Mika, Gunnar Rätsch and Christin Schäfer for sharing their thoughts and chocolate with me and for a sustained friendship. A very special and mathematical thank goes to Mikio "Ω" Braun, with whom I shared the office and the preference for mathematical stringency. Moreover, I owe a particular recognition to the members of the Berlin Brain-computer interface (BBCI) team with whom I had the pleasure to frequently work in common and who made major contributions to this work. So many thanks go to Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Michael Tangermann, Volker Kunzmann and Florian Losch. And last but by no means least, I am indebted to Jens Kohlmorgen who encouraged me to go my own way.

There are some other people who accompanied me from time to time along the scientific road: thanks to R. Schubert, Y. Hlushchuk. S. Liehr, and K. Pawelzik. Moreover, for a scientifically excellent atmosphere I thank all members of the Bernstein Center for Computational Neuroscience Berlin and the members of the collaborative research center for theoretical biology "SFB618". Especially I would like to thank Vadim Nikulin for many inspiring discussions on various aspects of computational neuroscience.

Financially, I gratefully acknowlegde partial support from the BMBF grants # FKZ 01GQ0415, # SFB 618-B4 from the Deutsche Forschungsgemeinschaft.

Notably, I thank my colleagues that have proof-read parts of this thesis, Benjamin Blankertz, Guido Dornhege, Andreas Ziehe, Mikio Braun - all remaining errors are mine!

Last but not least, I would like to express my thanks to those who non-scientifically, but to the same degree, contributed to the completion of this thesis. I would like to thank my parents for their love and support, Anja, Ronja and Thorben for adding a meaning to everything. Cheers to Bettina & Jesko, Rolf, Ulf, Benjamin, Tini & Martin, Kristin & Arend, to the denizen of the K12, and finally to Quelli for continuously sustaining our friendship.

Steven Lemm, December 2007.

Notation

Symbol	Explanation
$(\Omega, \mathcal{F}, P)$	Probability space, equipped with a σ-algebra and a probability measure
ω	Elementary event in Ω
$\mathcal{P}(\Omega)$	Power set of Ω
$(\mathbb{R}, \mathcal{B})$	Set of real numbers, equipped with the Borel σ-algebra
$\mathbb{Z}$	Set of integer numbers
X, Y, Z, C	Random variables
$\mathbb{X}, \mathbb{Y}, \mathbb{O}$	Stochastic processes
$f_X(x)$	Probability density function of the random variable X
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance matrix Σ
$U_{[a,b]}$	Uniform distribution on the interval $[a, b]$
$\mathbb{E}[X]$	Expectation value of the random variable X
$\mathbb{E}[X \mid Z]$	Conditional expectation of X given the random variable Z
$\mathcal{D}$	Set of observations
x^k	k^{th} realizations of the random variable X
$X_\cdot^k$	k^{th} realizations (path) of the stochastic processes $\mathbb{X}$
$\mathcal{T}$	Temporal index set of a stochastic process
$\lvert\mathcal{T}\rvert$	Number of elements of the set $\mathcal{T}$
Φ	Kernel function
h	Bandwidth parameter of a kernel function
$A, A^\top$	Generic matrix and its transposed
w	Spatial filter
b	Impulse response filter
λ	Regularization parameter
$\operatorname{diag}(\boldsymbol{x})$	Diagonal matrix with the elements (x_i)
I_n	$(n \times n)$-Identity matrix
τ	Delay parameter
δ^τ	Delay operator
$*$	Convolution operator
$\mathbb{1}_B$	Indicator function on the set B
$\lVert\cdot\rVert_p$	p-norm

Contents

Preamble

Understanding the brain is a challenge that is attracting scientist from various disciplines and defines the aim for the interdisciplinary field of computational neuroscience that serves as the primary theoretical method for investigating the function and mechanism of the nervous system. The interest in modelling *single trial* behavior of the human brain has rapidly grown in the past decades. Nowadays the scope of modern neuroscience has been widened to decoding *single trial* encephalogram data with respect to the identification of mental states or human intentions. This branch of research is strongly influenced by the development of an effective communication interface connecting the human brain and a computer [97], which finally also attracted the machine learning community to the field.

Accordingly, this thesis has a dual scope and therefore addresses an interdisciplinary audience: On the one hand we are going to strive for robust feature extraction methods that are particularly intended for enhancing the signal-to-noise ratio of *single trial* brain responses prior to their analysis. Here we will primarily concentrate on improving the classification of *single trial* encephalogram data in the context of BCI applications. This part will be of interest to practitioners and applied scientists. One the other hand parts of this thesis will concentrate on methods for variability analyses of brain responses to identical stimuli. More precisely, we will introduce an elaborate framework that allows for the variability analysis of spectral perturbations in dependence on internal or external factors. This part therefore aims for a fundamental issues of computational neuroscience and thus primarily addresses pure neuroscientists. Nevertheless, variability of brain responses has a direct impact on the design of classification models, e.g., for detecting mental states, consequently this part of the thesis should be of interest to applied research as well.

However, this thesis is the work of a mathematician who decided to do applied neuroscience. Hence it attempts to strike a balance between mathematical stringency and practical clearness, sufficient to annoy many readers. However, I tried to split the nuisance equally.

Roadmap

CHAPTER 1. The first chapter gives a brief introduction to encephalogram data, including acquisition techniques and the neurophysiological background. Moreover, we present the conventional concept of quantifying event-related (de-)synchronization (ERD/ERS) and the fundamentals about the primary area of application, i.e., brain-computer interfacing (BCI).

CHAPTER 2. The second chapter introduces the necessary mathematical preliminaries. Basically it recalls the definitions of a random variable and a stochastic process and elaborates on the concept of conditional expectation. Moreover, we introduce kernel density estimators and derives an empirical estimator for conditional expectation. At the end of this chapter a few summarizing examples are given.

CHAPTER 3. In the third chapter we derive a novel framework for the analysis of ERD. To this end we first generalize the conventional framework with respect to the commonly used reference condition. Secondly, based on the generalized model we derive the novel framework of *conditional ERD* that allows for the analysis of dependencies of the ERD characteristics on internal or external factors.

CHAPTER 4. The fourth chapter is concerned with an improved feature extraction method in the context of BCI. Here we present a new algorithm, the so-called *Common Spatio-Spectral Pattern* (CSSP) algorithm that extends the well known common-spatial-pattern (CSP) algorithm. Additionally to optimal discriminative spatial filters derived by CSP, the CSSP method determines simple frequency filters which enable the adaptation to the individual characteristics of the signals, i.e., to their individual power spectra. A comprehensive comparison of various methods on a broad set of BCI experiments proves the efficiency of the proposed method for the classification of imaginary hand movements.

CHAPTER 5. In the fifth chapter we introduce an approach for the extraction of

event-related potentials (ERP) from *single trial* data. To this end we will particularly incorporate prior knowledge about the phase-locked nature of ERPs into a Independent Component Analysis framework. The application on artificially generated and real world data validate this approach in terms of an improved signal-to-noise ratio of the recovered ERPs.

CHAPTER 6. In the sixth and last chapter we report on our winning algorithm that has been successfully applied to data from the international BCI competitions in 2003 and 2005. In particular we present a Bayesian classification framework which combines sequences of features efficiently across time.

A brief discussion concludes this thesis.

Chapter 1

Introduction

At the beginning of the last century Hans Berger detected the human electroencephalogram (EEG). His most outstanding finding was the existence of prominent oscillations in the frequency range between 8 and 12 Hz, which he called alpha wave rhythm. He also studied and described for the first time its suppression (substitution by the faster beta waves) when the subject opens its eyes (the so-called alpha blockade). Since this early work, the processing of external or internal stimuli and its accompanying modulation of ongoing neurocortical activity has been studied extensively. These investigations were mostly focused on the analysis of induced neural activity, such as evoked potentials and event-related spectral perturbations. However, evoked brain responses are typically weaker than the accompanying ongoing neural activity. Moreover, recordings of brain activity are usually contaminated with artifacts that disadvantageously exceed the signals of neural origin by an order of magnitude, and often reside in the same spectral, temporal or spatial domain as the evoked components. Consequently, raw recordings of brain activity just provide evoked *single trial* responses at a low signal-to-noise ratio (SNR) and mostly limited neurophysiological investigations to the analysis of averaged responses to repeated identical stimuli. Averaging responses to several repetitions of a certain stimulus improves the signal-to-noise ratio of the evoked responses, but also coincides with masking the variability of the *single trial* responses and hence inhibits the understanding of human cortical neurophysiology on the basis of *single trial* behavior.

To facilitate the analysis of *single trial* responses, this thesis presents novel concepts for their analysis, as well as novel feature extraction methods to enhance their SNR.

1.1 Data acquisition

There exists a variety of different methods to measure brain activity. The question of which recording technique appears most appropriate in order to investigate a specific neurophysiological issue, depends on several factors. All available techniques focus on specific physical or physiological correlates of neural activity. At a first distinction level, these methods divide into direct and indirect measurements. Here, the first group comprises recordings of the electromagnetic field, while the indirect measurements basically focus on the detection of changes in the blood flow and the blood oxygenation, which are closely linked to neural activity. Changes in the electrical or magnetic field can be measured in the range of milliseconds. On the opposite the hemodynamic response rises to a peak over 4-5 seconds, before falling back to baseline. Consequently the indirect measurements of vascular activity, for instance by functional magnet resonance imaging (fMRI), are hampered by a low temporal resolution that typically lies in the range of seconds. Another differentiation is given by the degree of invasiveness of the different measurements. Here the available techniques to access the neural activity can be ordered, starting at the most invasive methods where electrodes are directly implanted within the brain, going on to methods that place electrodes subdurally (below the skull but still above the brain) and finally to non-invasive methods, such as the EEG or magnetoencephalogram (MEG) that measure neural activity from a macroscopic perspective. Generally speaking, the quality of the obtained signals improves along with a higher degree of invasiveness. However, this comes at the cost of medical risk for the patient, hence it is impractical to use invasive techniques on healthy subjects. Another important difference is given by the spatial resolution of the individual methods. On the one hand there are multielectrode arrays, which can be used to record action potentials of single neurons in the cerebral cortex. In contrast a moderately good spatial resolution is provided by the fMRI, where the voxels in the resulting image typically represent cubes of tissue about 2-4 millimeters on each side, while the spatial resolution of the EEG and the MEG are worst. Here the resolution directly depends on the number and the placement of the sensors.

1.1.1 Electroencephalogram

Throughout this thesis we will restrict ourselves to EEG recordings. However, as the recorded signals of EEG and MEG are quite similar with respect to temporal and spatial resolution, most algorithms developed for analyzing EEG signals can be transferred smoothly to MEG data.

The EEG is an extra-cranial noninvasive recording technique that is sensitive to changes in the electrical fields generated by neural activity. It was first discovered by Hans Berger in 1924 and published in 1929 [6]. More precisely, EEG signals are electrical potentials that are determined at particular positions on the scalp relative to one or more reference electrodes. Often the electrode which serves as reference is placed at the earlobe or at the bridge of the nose. Commonly the EEG is simultaneously recorded at many locations by a set of electrodes placed at different positions on the scalp. Ordinarily the term channel is used to refer to a single recording position. The distance between neighboring electrodes usually is in the range of one to a few centimeters. The currently available EEG caps provide up to 256 channels. An exemplary montage scheme of a 128 channel EEG recording system using an electrode placed on the nose as reference is shown in Fig. 1.1. The position of electrodes and the nomenclature of the corresponding channels follow international standards [16].

As the recorded signals are in the order of $\pm 100\mu$V an amplifier enlarges the signal range to, e.g., ± 5 Volt before it is sampled for the computer. Typical sampling rates for EEG are in the range of 100 Hz and up to 1000 Hz. This is quite sufficient to track the typical EEG patterns that are reported to have frequency components approximately below 80 Hz. Only occasionally higher sampling rates up to 5000 Hz are used, for instance to study very high frequency responses at approximately 600 Hz accompanying early evoked somatosensory potentials [4].

As previously stated, EEG recordings have a very good temporal resolution but suffer from spatial insufficiencies that are mostly caused by the scull bone, the meninges and the intra-cerebral liquor. These layers act as a spatial low-pass filter. Thus spatial resolution is not necessarily limited by the distance between electrodes (usually approximately 2cm) but additionally by this smearing effect. The EEG as well as any other data acquisition method, records not only the signal of interest, but also a variety of specific artifacts, which are signal components picked up by the sensors that are not

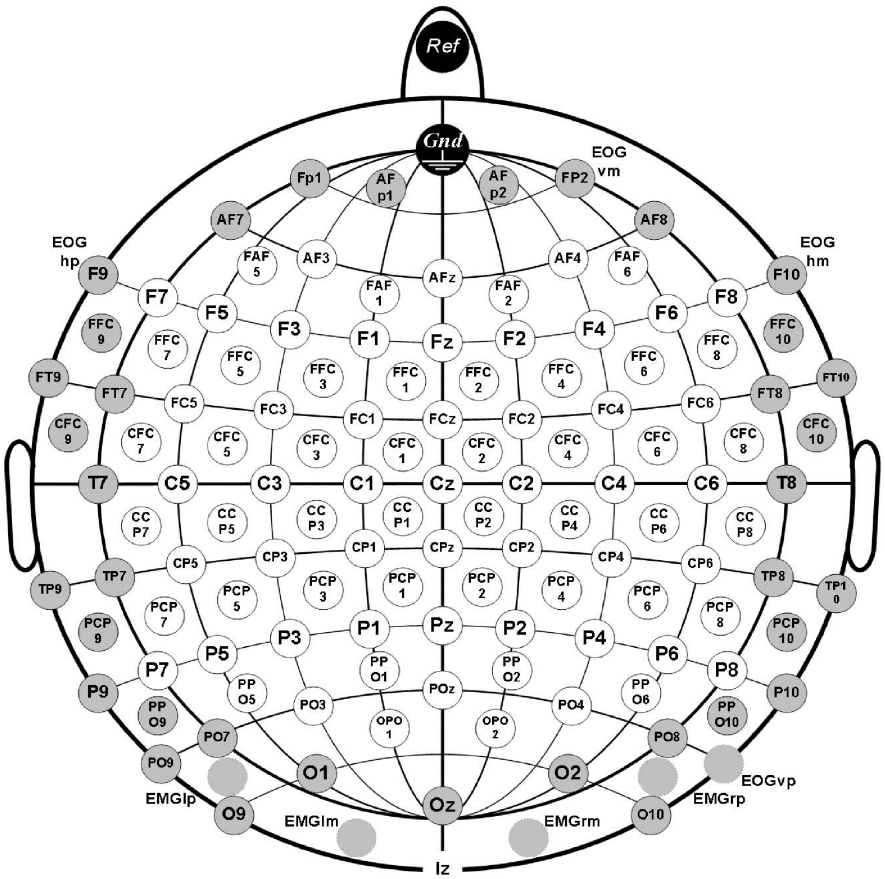


Figure 1.1: Exemplary EEG montage scheme indicating the position of electrodes and nomenclature of the corresponding channels according to international standards. This particular montage uses a few selected electrodes for the additional acquisition of electro-oculogram (EOG) and electro-myogram (EMG). [Figure taken from [44] with the consent of the author]

caused by neural activity. Typical artifacts in EEG recordings comprise:

- *Muscle activity:* This class of artifacts is caused by e.g., clenching jaws, facial and neck muscles. They can easily be recognized and at least partially filtered due to prominent frequency components above 50 Hz and high amplitudes of up to a few hundred μV. To ensure accuracy the subjects are instructed to avoid any kind of movements during the recordings.

- *Movements of the eyeball:* The eye movement of a healthy subject typically exhibit saccades. Since each eyeball represents a strong dipole at a fixed location, movements of the eyes typically result in a rotation of the corresponding electric

fields, which causes large artifacts affecting all electrodes. To avoid the masking of neural signals by that kind of artifact, the subjects are asked to keep a sustained gaze at a marked spot during the course of the experiment.

- *Eye blinks:* In EEG recordings an eye blink is represented as a peak with an higher signal amplitude than the neural activity. Moreover, eye blinks have a fixed scalp pattern that exhibits a distinct gradient pointing from the occipital towards the frontal areas. Thus their impact can be reduced by appropriate spatial filters. Additionally, the subjects are instructed to avoid or postpone eye blinks until the occurrence of an explicitly reserved time interval.

- *External strays*: This class of artifacts is always present in unshielded environments. The most prominent jamming sources are nearby electrical devices, such as power supplies, rectifiers, electric bulbs, etc. that are basically reflected by a prominent 50/60 Hz (and its harmonics) component.

The influence of most artifacts can be removed or at least reduced either by proper instruction of the subjects, or a posteriori by appropriately (spectrally and spatially) filtering the recorded signals.

1.2 Neurophysiological background

Throughout this thesis we study brain responses during the processing of either sensory information or motor commands. The processing of such stimuli mainly causes responses in the somatosensory and the motor cortex, respectively. Consequently we restrict our analysis to neural activity originating from these locations. Both cortical areas reside in the central (perirolandic) region and are separated by the central sulcus. In particular the primary motor area occupies the precentral gyrus, while the primary somesthetic area occupies the postcentral gyrus. To give a rough overview, Fig. 1.2 depicts a coarse topography of the human brain.

Beside a differently distinct spatial representation, the human perirolandic sensorimotor cortices show rhythmic macroscopic EEG oscillations (μ-rhythm) [32], with spectral peak energies around 10 Hz (localised predominantly over the postcentral somatosensory cortex) and 20 Hz (over the precentral motor cortex). These oscillations are

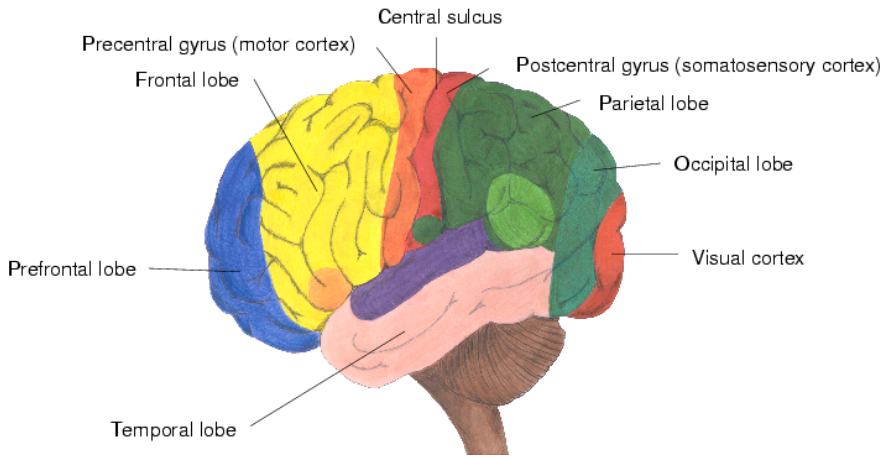


Figure 1.2: Coarse overview about the topography of some important cortical areas in the human brain. [taken from [44] with the consent of the author]

generated by large populations of neurons in the respective cortex that fire in rhythmical synchrony. However, the macroscopically observed oscillations exhibit fast inherent fluctuations as they are limited to brief periods (bursts) of $0.5-2\,\mathrm{s}$ duration [71], which appear to occur in the absence of processing sensory information or motor commands. Hence the μ-rhythm was originally conceived to reflect a cortical idling or "nil-work" state [81]. In contrast recent studies argue for a relationship between brain rhythms and higher cognitive functions, reporting short- and long-lasting suppression of ongoing activity with cognitive processing, but a precise relationship is not yet clearly established. For an elaborated review of the functional significance especially of μ-rhythm see [82]. However, modulations of the μ-rhythm have been reported for different physiological manipulations, e.g., by motor activity, both actual and imagined [37, 77, 88], as well as by somatosensory stimulation [72]. As the attenuation effect of rhythmic activity is due to a loss of synchrony in the neural population, it is termed event-related desynchronization (ERD) [78], while the dual effect of enhanced rhythmic activity is called event-related synchronization (ERS).

1.3 Event-related (de-)synchronization

The power spectra of human brain activity measured by EEG or MEG exhibit a characteristic $1/f$ shape, with some intermediate spectral peaks. The $1/f$ part of the signal is usually considered as noise or ongoing background activity, and is basically

generated by the predominant amount of neurons firing in an asynchronous fashion. On the opposite a peak in the power spectrum indicates the presence of a pronounced rhythmicity that is caused by a neural oscillator. Such an oscillator is composed of neurons from a certain cortical area, where these neuron fire in synchrony. If a stimulus is processed by that cortical area, then a subset of these neurons is recruited for this task and is thereby detached from synchronously firing. It is the goal of ERD analysis to quantitatively determine the amount of disturbance of such a neural oscillator.

To this end, ERD is conventionally defined as the relative difference in signal power of a certain frequency band, between two conditions, i.e, a reference period and an immediate event-related period. Hence ERD and ERS describe the power modulation of the ongoing activity, induced by a certain stimulus or event. Expressed as a formula, let $(P_t)_{t\in\mathcal{T}}$ denote the time course of the event-related power in a narrow frequency band covering the spectral peak and let P_{REF} denote the power at the reference condition, then ERD is quantified as the relative difference in power between both conditions, i.e.,

$$\text{ERD}[t] := \frac{P_t}{P_{\text{REF}}} - 1. \tag{1.1}$$

By convention an ERD corresponds to a decrease in power, while ERS refers to an increase in signal power [77]. Up to now there have been basically two very similar methods for estimating the ERD, i.e., the power method [77] and the intertrial variance method [40]. A schematic representation of both approaches is depicted in Fig. 1.3. Both methods express ERD as the change of the averaged (expected) rhythmic activity relative to an (averaged) reference activity of the unperturbed dynamic. The essential difference between both techniques lies in the fact that the intertrial variance method compensates for the amount of power modulation, which is introduced by the phase-locked components. To this end it removes the averaged response from the narrow bandpass filtered signal.

Notable, in this context, *the term ERD is only meaningful if the baseline measured some seconds before the event represents a rhythmicity seen as a clear peak in the power spectrum. Similarly, the term ERS only has a meaning if the event results in an appearance of a rhythmic component and consequently in a spectral peak that was initially not detectable* [78].

However, the conventional concept of ERD and ERS just describe the relative difference between the averaged power of the unperturbed and the perturbed oscillatory

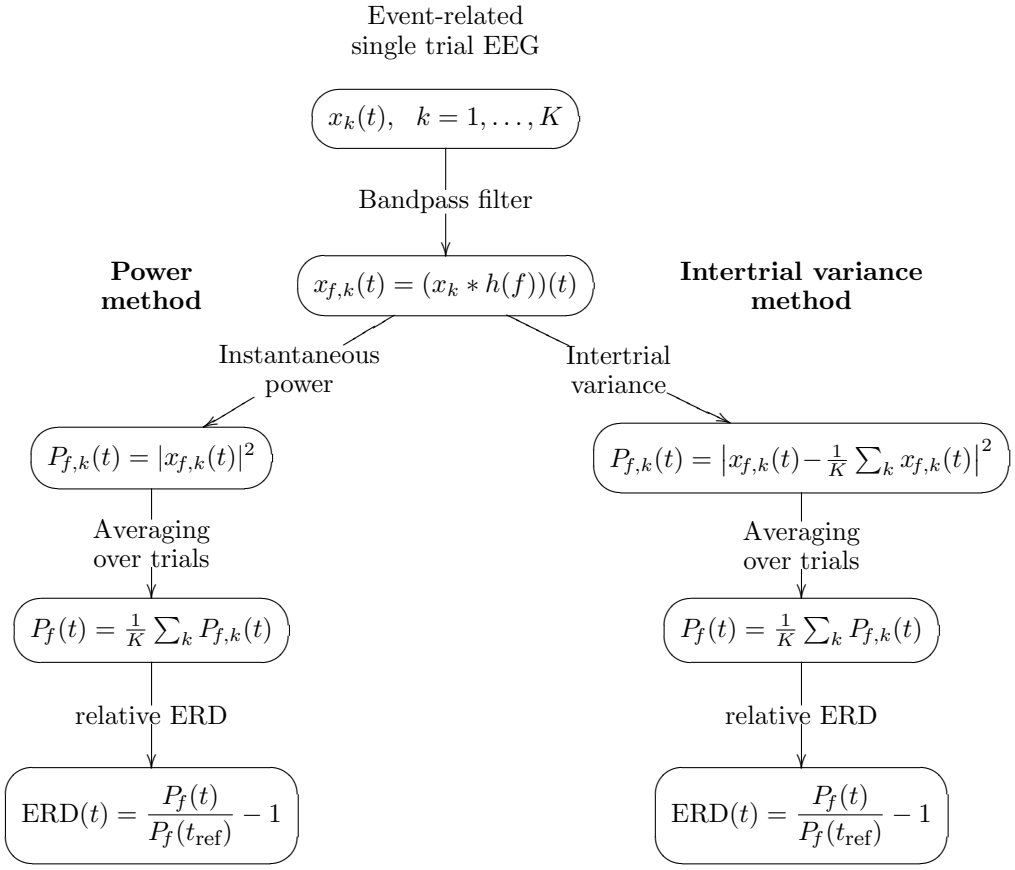


Figure 1.3: The two conventional procedures for estimating the relative ERD [40]. The left path in the diagram illustrates the power method while the right path explains the inter-trial variance method which adjusts for the averaged evoked potential.

.

system and lacks of an appropriate *single trial* model, which allows for studying dependencies on initial mental or cortical states.

Moreover, the ERD estimator is affected by additive noise and therefore tends to underestimate the genuine ERD. To give an illustrative example let us consider the suppression of μ-rhythmic activity caused by processing sensory stimuli. Suppose further that the narrow band signal comprises the superposition of two oscillatory processes, i.e., the μ-process itself and the ongoing background activity. Let us intuitively denote the power of the individual signals by μ_t^2 and B_t^2, respectively (see Fig. 1.4). Assuming mutually independent phasing of both processes it follows that the expected instantaneous band power of the composed signal equals the sum of the individual signals

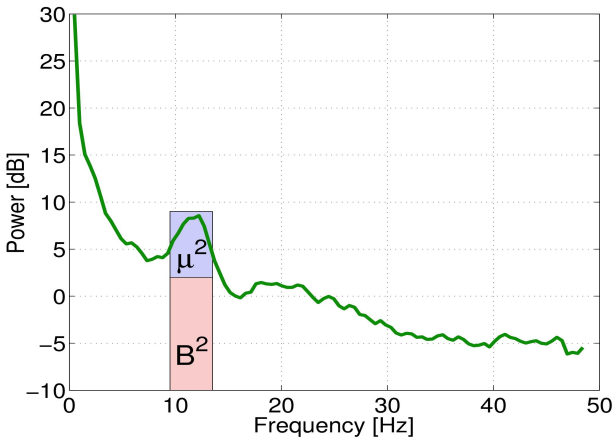


Figure 1.4: Power spectrum of an EEG signal obtained at electrode "CP3" over the left somatosensory cortex.

powers, i.e., $P_t = \mu_t^2 + B_t^2$. Consequently in the case of an exclusive perturbation of the μ-rhythm (1.1) yields:

$$\begin{aligned} \left|\mathrm{ERD}[t]\right| &= \left|\frac{\left((\mu_t^2 + B_t^2) - \mu_{\mathrm{REF}}^2 + B_{\mathrm{REF}}^2\right)}{\mu_{\mathrm{REF}}^2 + B_{\mathrm{REF}}^2}\right| && (1.2) \\ &\simeq \left|\frac{\mu_t^2 - \mu_{\mathrm{REF}}^2}{\mu_{\mathrm{REF}}^2 + B_{\mathrm{REF}}^2}\right| && (1.3) \\ &\leq \left|\frac{\mu_t^2 - \mu_{\mathrm{REF}}^2}{\mu_{\mathrm{REF}}^2}\right| = \left|\frac{\mu_t^2}{\mu_{\mathrm{REF}}^2} - 1\right|. && (1.4) \end{aligned}$$

Note that in (1.3) we explicitly used the assumption that the power of the background activity is approximately equivalent at both conditions, i.e., $B_t \simeq B_{\mathrm{REF}}$. This assumption appears to be plausible, as the observed $1/f$ shape of the power spectrum does not exhibit any significant change during stimulus processing at other frequencies. However, the final term in (1.3) corresponds to the exclusive, genuine ERD of the μ-rhythm, which is clearly underestimated by the conventional ERD measure. Moreover, adding further task-unrelated rhythmic activity to the system, e.g., occipital α-oscillations, the absolute value of the estimated μ-rhythm ERD will continuously decrease. This preliminary thought emphasizes the importance of a preprocessing of the data, e.g, by advanced feature extraction method that spatially and spectrally filter the signals in order to enhance the SNR of the μ-rhythm.

1.4 Brain computer interfacing

The ambitious goal of brain computer interface (BCI) research is to develop a novel augmented communication system that translates user intentions – reflected by suitable brain signals – into control signals [97, 104]. Such a signal transduction pathway may give disabled people direct control over neuro-prostheses or computer applications as tools for communicating solely by their intentions. In practice the user is behaving according to a well-defined paradigm (e.g., movement imagination) that allows for effective discrimination between different brain states which are used to encode information.

Currently there are plenty of different approaches and realizations of a BCI system and the number is continuously growing. In principle they can be grouped in several ways. With respect to the recording technique we distinguish between invasive [33, 45, 50, 51, 94] and non-invasive BCIs [7, 8, 64, 58, 80]. A further distinction is given by either using the training capabilities of the human [7, 105] or those of the computer [8, 64, 76]. Finally there are different ways to provoke heterogeneous brain states. Some BCIs are based on externally evoked potentials [17, 62], while others uses the changes in the ongoing dynamic which accompany human intentions [7, 8, 64, 58, 80]. In the latter case, many approaches uses imaginary movements of the foot and the left and the right hand to provoke different states. In order to distinguish between the individual limbs they exploit differently lateralized μ-rhythm ERD and ERS effects that accompany the imaginary movements.

From the signal processing perspective a BCI requires the definition of appropriate features that can be effectively translated into a control signal, either by simple threshold criteria (cf. [104]), or by means of machine learning. For the latter, the task is to infer a possibly nonlinear discriminating function that distinguishes the different states [9, 65, 74, 75]. However, the accuracy of any BCI system is linked to the achievable SNR of the *single trial* features that represent the different states. Moreover, as a mandatory requirement for a BCI, it has to operate in an online fashion, i.e., it has to instantaneously process the recorded neurocortical activity, which is in fact the analysis of *single trial* data. Consequently the appropriateness of models describing the *single trial* behavior also determines the performance of a BCI. Accordingly, advances in BCI research are currently characterized by improvements in feature extraction and refinements to the *single trial* models [24].

Chapter 2

Preliminaries

In this chapter we will introduce the necessary fundamentals of the theory of probabilities and estimation along with their notations[1]. We will start with the mathematical definition of a random variable and will introduce the concept of random variables sharing a common probability space. Based on that, we will extend the definition of a random variable to those of a stochastic process, as a family of random variables defined on a common probability space. The main part of this chapter will elaborate on the introduction of conditional expectation and its estimation. Usually conditional expectation is used in the context of regression or classification with the goal of explicitly modelling dependencies between random variables. In chapter 3 we will exploit the concept of conditional expectation to analyze the dependencies of a dynamic system, such as ERD, on environmental (initial) states/conditions. To this end we will extend the definition of conditional expectation from random variables to stochastic processes. The chapter will conclude with two examples which estimate the conditional expectation of a random variable and of a stochastic process.

2.1 Random variables and stochastic processes

In the following, let $(\Omega, \mathcal{F}, P)$ always denote a probability space, equipped with a probability measure P that maps each set of the σ-algebra $\mathcal{F}$ over Ω to a value in $[0, 1]$.

[1]Note, it is not within the scope of this chapter to give a complete introduction into the field of probability theory. For this purpose, we would like to refer to an introductory textbook for further reading, e.g., [27].

Definition 2.1.1 (Random Variable). Let $(\Omega', \mathcal{F}')$ be an arbitrary measure space, where random variable X is a function $X : \Omega \to \Omega'$, such that for every $F' \in \mathcal{F}'$ the pre-image $X^{-1}(F')$ is an element of $\mathcal{F}$, i.e., any $\mathcal{F}-\mathcal{F}'$ measurable function defines a random variable on Ω'.

In case of $(\Omega', \mathcal{F}')$ equals $(\mathbb{R}, \mathcal{B})$ or $(\mathbb{R}^d, \mathcal{B}^d)$, we briefly refer to the random variable as a (d-dimensional) real-valued random variable, where $\mathcal{B}$ is the Borel σ-algebra, i.e., the σ-algebra that contains all open sets.

Remark 2.1.1. In the setting of the previous definition PX^{-1} defines a probability measure on $(\Omega', \mathcal{F}')$ and thereby completing the measure space to a probability space $(\Omega', \mathcal{F}', PX^{-1})$.

In order to study dependencies between random variables or to introduce the definition of a stochastic process, it is necessary to understand the basic concept of random variables sharing a common probability space. This can be easily understood from an illustrative example of two random variables, for instance two dice. Thus let's define a common probability space as $\Omega := \{(i,j) : i,j \in 1,\ldots,6\}$, equipped with the σ-algebra $\mathcal{F} := \mathcal{P}(\Omega)$ along with an arbitrary probability measure P. (Here $\mathcal{P}(\Omega)$ denotes the power set of Ω.) Furthermore we assume two identical measure spaces for each die, i.e., $\Omega_1 = \Omega_2 := \{1,\ldots,6\}$, equipped with $\mathcal{F}_1 = \mathcal{F}_2 := \mathcal{P}(\Omega_1)$, see Fig. 2.1. Based on this particular setting, we define two random variables X and Y as the canonical projections of an element $\omega = (i,j) \in \Omega$ onto its first and second component respectively. Note that each elementary event $\omega \in \Omega$ simultaneously determines both random variables, i.e., $X(\omega)$ and $Y(\omega)$. According to the conventional definition of statistical independence,

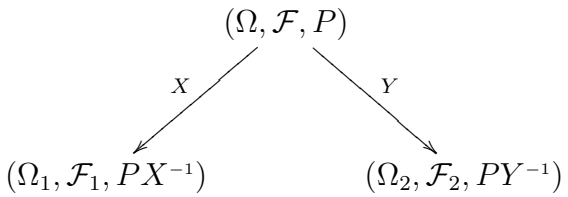


Figure 2.1: Schematic of two random variables X and Y, sharing a common probability space $(\Omega, \mathcal{F}, P)$.

the two random variables X and Y are mutually independent, if and only if for all $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$

$$PX^{-1}(F_1) \cdot PY^{-1}(F_2) = P\left(X^{-1}(F_1) \cap Y^{-1}(F_2)\right). \tag{2.1}$$

Remark 2.1.2. We will use $\mathcal{F}_X \subseteq \mathcal{F}$ as the smallest σ-algebra, for which X is measurable, i.e., $\mathcal{F}_X$ contains all sets of the form $X^{-1}(F), F \in \mathcal{F}_1$ and $\mathcal{F}_Y$ correspondingly. Then the statistical independence of the two random variables can by equivalently expressed as the independence of the two σ-algebras $\mathcal{F}_X$ and $\mathcal{F}_Y$ over Ω, i.e., for all $F_1 \in \mathcal{F}_X$ and $F_2 \in \mathcal{F}_Y$ the probability of the intersection $P(F_1 \cap F2)$ is equal to the product of the probability of the individual events $P(F_1) \cdot P(F_2)$.

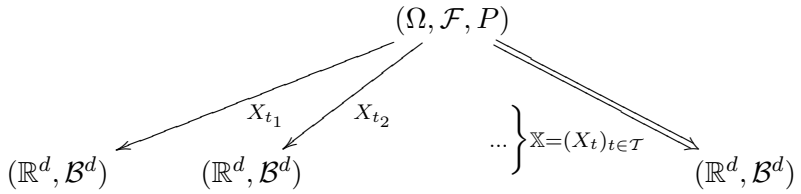
However, the concept of random variables sharing a common probability space can be easily extended to an infinite number of random variables, which directly leads to the definition of a stochastic process.

Definition 2.1.2 (Stochastic Process). For a given index set $\mathcal{T}$, a stochastic process $\mathbb{X}$ is a family of $d-$dimensional real valued random variables $(X_t)_{t\in\mathcal{T}}$ defined on a common probability space $(\Omega, \mathcal{F}, P)$, such that:

$$\mathbb{X} : \Omega \times \mathcal{T} \longrightarrow \left(\mathbb{R}^d, \mathcal{B}^d\right), \tag{2.2}$$

$$(\omega, t) \longmapsto X_t(\omega). \tag{2.3}$$

A realization $X.(\omega) := (X_t(\omega))_{t\in\mathcal{T}} \in \mathbb{R}^{d\times\mathcal{T}}$ is called a path of the stochastic process.



Remark 2.1.3. From a given stochastic process on $\mathbb{R}^d$ one can easily derive a new stochastic process on $\mathbb{R}^m$ by linear transformation with a matrix $A \in \mathbb{R}^{m\times d}$. The spatially filtered version $\mathbb{Y}_A := A\mathbb{X} : (\omega, t) \mapsto AX_t(\omega)$ defines a family of m-dimensional random variables $(Y_t)_{t\in\mathcal{T}}$ and hence a stochastic process. In a similar manner the path-wise temporally filtered version $\mathbb{Y}_b := \mathbb{X} * b : \omega \mapsto X.(\omega) * b$ using an (in-)finite impulse response filter b gives rise to another stochastic process. Note that the stochastic processes $\mathbb{X}, \mathbb{Y}_A$, and $\mathbb{Y}_b$ are defined within a common probability space $(\Omega, \mathcal{F}, P)$ (Fig. 2.2).

2.2 Conditional expectations

In order to introduce the concept of conditional expectation, let us consider the example of two random variables X and Y, sharing a common probability space $(\Omega, \mathcal{F}, P)$.

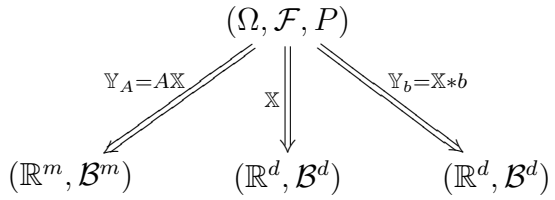


Figure 2.2: Schematic of three stochastic processes sharing a common probability space. The processes $\mathbb{Y}_A$ and $\mathbb{Y}_b$ are given as filtered versions of the process $\mathbb{X}$.

More precisely, let X be Ω'-valued and Y be an real valued integrable random variable $Y : \Omega \to \mathbb{R}$, see Fig. 2.3 for a schematic. Again, using the same formalism as before,

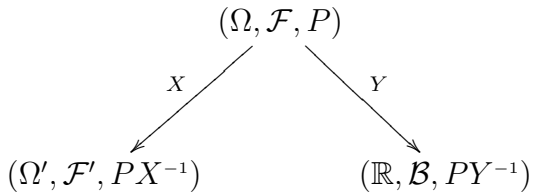


Figure 2.3: Schematic of two random variables X and Y, sharing a common probability space $(\Omega, \mathcal{F}, P)$.

PY^{-1} defines a probability measure on $\mathbb{R}$. According to the conventional definition, the expected value of Y is given as $\mathbb{E}[Y] := \int_\Omega Y(\omega)\, dP(\omega) = \int_{\mathbb{R}} y\, dPY^{-1}$.

Definition 2.2.1 (Conditional Expectation). Let $\mathcal{F}_X \subseteq \mathcal{F}$ denote the smallest σ-algebra, for which X is measurable, i.e., $\mathcal{F}_X$ contains all sets of the form $X^{-1}(F), F \in \mathcal{F}$. Then a version of the conditional expectation of Y given X is an integrable random variable $\mathbb{E}[Y\,|X] : \Omega \to \mathbb{R}$, such that

$$i) \qquad \mathbb{E}[Y\,|X] \text{ is an } \mathcal{F}_X - \mathcal{B} \text{ measurable function} \tag{2.4}$$

$$ii) \qquad \int_B \mathbb{E}[Y\,|X]\,(\omega)\, dP(\omega) = \int_B Y(\omega)\, dP(\omega), \quad \forall B \in \mathcal{F}_X. \tag{2.5}$$

Example 2.2.1. To quote a simple example, let us consider a fair die, i.e.,

$$\begin{aligned} \Omega &= \{1,2,3,4,5,6\} \\ \mathcal{F} &= \mathcal{P}(\Omega) = 2^\Omega \\ P(\omega) &= \frac{1}{6},\ \forall \omega \in \Omega. \end{aligned} \tag{2.6}$$

Then we define a random variable Y as the number on the die, i.e., $Y(\omega) = \omega$, and furthermore a random variable $X : \Omega \to \{0,1\}$, on which we will condition,

as the function that assigns 1 to all even ω and 0 to all odd ω. It follows, that $\mathcal{F}_X = \{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}$ and $P(F) = \frac{1}{2}$ for any non-trivial set F of $\mathcal{F}_X$. Putting everything together we finally obtain

$$\mathbb{E}[Y\,|X]\,(\omega) = \begin{cases} \frac{1}{\frac{1}{2}} \cdot \left(\frac{1}{6} + \frac{3}{6} + \frac{5}{6}\right) = \left(\frac{1}{3} + \frac{3}{3} + \frac{5}{3}\right) = 3 & \omega \text{ odd}, \\ \frac{1}{\frac{1}{2}} \cdot \left(\frac{2}{6} + \frac{4}{6} + \frac{6}{6}\right) = \left(\frac{2}{3} + \frac{4}{3} + \frac{6}{3}\right) = 4 & \omega \text{ even}. \end{cases} \tag{2.7}$$

Thus $\mathbb{E}[Y\,|X] : \Omega \to \mathbb{R}$ is a random variable, that takes the value "3" on all odd ω and the value "4" for even ω respectively. One can easily prove that the integrals of Y and $\mathbb{E}[Y\,|X]$ over any measurable set $F \in \mathcal{F}_X$ are equivalent. Notably, integrating $\mathbb{E}[Y\,|X]$ over the entire space Ω gives $\frac{3}{2} + \frac{4}{2} = 3.5$ that is equivalent to $\mathbb{E}[Y]$.

To get some intuition as to what conditional expectation is all about, you may interpret it as a random variable that resembles a "less random" version of Y, where randomness of Y has been "reduced" by the amount of information that is provided by the observation of the random variable X. Metaphorically speaking, given a certain knowledge about X, e.g., $X \in F$, we can restrict the feasible set in Ω to $X^{-1}(F) \subseteq \Omega$. Along with this additional information about ω our expectation of the random variable Y changes, and is finally expressed by the conditional expectation. In the context of the example 2.2.1 this corresponds to: given the partial observation $X(\omega) =$"odd", we conclude that $\omega \in \{1,3,5\}$, thus our expectation about Y changes "3".

Proposition 2.2.2 (Basic Properties). *Let X, Y and Z be random variables, defined on a common probability space $(\Omega, \mathcal{F}, P)$, with Y and Z integrable, i.e., the integral of Y and Z exist. Furthermore, let $\mathcal{F}_X \subseteq \mathcal{F}$ be the smallest σ-algebra, for which X is measurable. Then the following equations hold P almost surely:*

$$a) \quad \mathbb{E}[aY + bZ\,|X] = a\,\mathbb{E}[Y\,|X] + b\,\mathbb{E}[Z\,|X]\,, \text{ for } a, b \in \mathbb{R}, \tag{2.8}$$

$$b) \quad \mathbb{E}[\alpha Y\,|X] = \alpha\,\mathbb{E}[Y\,|X]\,, \text{ for any } \mathcal{F}_X\text{-measurable function } \alpha \tag{2.9}$$

$$c) \quad \mathbb{E}[\,\mathbb{E}[Y\,|X]] = \mathbb{E}[Y] \tag{2.10}$$

$$d) \quad \mathbb{E}[Y\,|X] = \mathbb{E}[Y]\,, \text{ for independent } X, Y. \tag{2.11}$$

Additional properties of the conditional expectation and detailed proofs, e.g., of (2.9) can be found in [27].

Remark 2.2.2. Property (2.8) just states that the conditional expectation is linear. The second property (2.9) extends the linearity to $\mathcal{F}_X-$measurable functions. Since

the integrals of Y and $\mathbb{E}[Y\,|X]$ are equal for all measurable sets, and the space Ω is measurable per default, property (2.10) is a direct corollary from the definition of the conditional expectation. Notably, property (2.11) says that conditioning on independent information does not allow the inference of further knowledge about the expectation of Y. Consequently the conditional expectation is a constant function, i.e., $\mathbb{E}[Y\,|X]\,(\omega) \equiv \mathbb{E}[Y]\,, \forall \omega \in \Omega$.

Proposition 2.2.3. *There is one and only one integrable function $\psi : \Omega' \to \mathbb{R}$, such that $\mathbb{E}[Y\,|X] = \psi \circ X$ and for any measurable subset $F' \in \mathcal{F}'$:*

$$\int_{F'} \psi \, dPX^{-1} = \int_{X^{-1}(F')} \psi \circ X \, dP = \int_{X^{-1}(F')} \mathbb{E}[Y\,|X] \; dP = \int_{X^{-1}(F')} Y(\omega)\, dP(\omega). \quad (2.12)$$

The proof of this proposition is a direct application of the factorization theorem (for a proof of 2.2.3 and the factorization theorem we refer to [27]). However, Fig. 2.4 summarizes the concept of conditional expectation in terms of a commutative diagram.

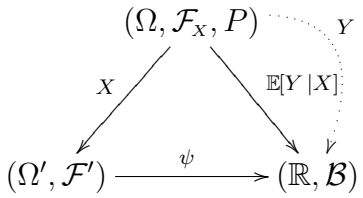


Figure 2.4: Schematic of the conditional expectation of the real-valued integrable random variable Y given the random variable X, defined on the same probability space $(\Omega, \mathcal{F}, P)$. The indicated equivalence between Y and $\mathbb{E}[Y\,|X]$ means that integrals over sets of the form $X^{-1}(F')$ for $F' \in \mathcal{F}'$ are equal.

Remark 2.2.3. Since a stochastic process just refers to a collection of random variables on a common probability space, the framework of conditional expectation can easily be extended from single random variables to stochastic processes. Therefore let $\mathbb{Y} = (Y_t)_{t\in\mathcal{T}}$ be a family of $d-$dimensional real valued random variables and X a single random variable. According to definition 2.2.1 $\mathbb{E}[Y_t\,|X]$ is a random variable. Thus $(\mathbb{E}[Y_t\,|X])_{t\in\mathcal{T}}$ defines a family of random variables, i.e, a stochastic process that we denote by $\mathbb{E}[\mathbb{Y}\,|X]$. Consequently, applying proposition 2.2.3 to each conditional expectation $\mathbb{E}[Y_t\,|X]$ separately results in a family of functions $(\psi_t)_{t\in\mathcal{T}}$.

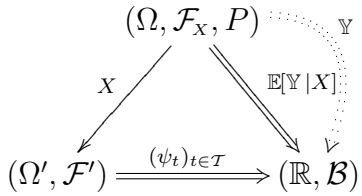
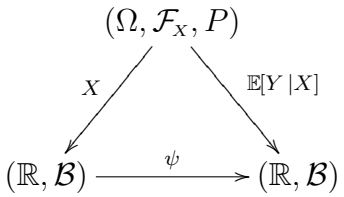


Figure 2.5: Schematic of the conditional expectation of the real-valued integrable random variable Y given the random variable X, defined on the same probability space $(\Omega, \mathcal{F}, P)$. The indicated equivalence between Y and $\mathbb{E}[Y\,|X]$ means that integrals over sets of the form $X^{-1}(F')$ for $F' \in \mathcal{F}'$ are equal.

2.3 Estimating conditional expectations

For the sake of simplicity, let us return to the case of two random variables and additionally consider X to be a real valued, univariate random variable on Ω and Y as previously. Then according to proposition 2.2.3 there exists a function $\psi : \mathbb{R} \to \mathbb{R}$ such that $\psi(x) = \mathbb{E}[Y\,|X = x]$.



In the following we derive a closed form solution for the integrable function ψ, given a set of independent samples of the joint random variables (X, Y). Using the notation of $f_{X,Y}$ for the joint probability density function (PDF), the marginal PDF of X is given as $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dy$. Based on this we define the conditional PDF of Y given X as

$$f_{Y\,|X}(y\,|x) := \frac{f_{X,Y}(x, y)}{f_X(x)}. \tag{2.13}$$

Accordingly, the conditional expectation $\mathbb{E}[Y\,|X = x]$ is expressed as

$$\psi(x) = \mathbb{E}[Y\,|X = x] = \int_R y\, f_{Y\,|X}(y\,|x)\, dy \tag{2.14}$$

$$= \frac{1}{f_X(x)} \int_R y\, f_{X,Y}(x, y)\, dy. \tag{2.15}$$

Remark 2.3.1. Compare the above formula also with (2.7) in example 2.2.1.

In order to apply formula (2.15) we have to estimate the involved PDFs based on the

observations of the random variables. For this purpose the next paragraph elaborates on kernel density estimation.

2.3.1 Kernel density estimation

Kernel density estimators are frequently used to infer PDFs from a given set of observations. In case of univariate data, a kernel density estimator is generally defined as follows.

Definition 2.3.1 (Univariate kernel density estimator). Given a set of independent observations $\mathcal{D} := \{x^k\}_{k=1}^N$, the univariate kernel density estimator of the PDF f_X is defined by

$$\hat{f}_X(x\,|\mathcal{D}) := = \frac{1}{Nh}\sum_{k=1}^{N}\Phi\left(\frac{x^k - x}{h}\right), \tag{2.16}$$

where h is called the bandwidth or smoothing parameter and Φ is called the kernel function that satisfies

$$\Phi(u) \geq 0,\ \forall u \in \mathbb{R} \tag{2.17}$$

$$\int_{\mathbb{R}} \Phi(u)\,du = 1 \tag{2.18}$$

$$\int_{\mathbb{R}} u\Phi(u)\,du = 0. \tag{2.19}$$

Commonly, there are further assumptions imposed on the bandwidth, e.g.,

$$h \longrightarrow 0,\ \text{as } N \to \infty \tag{2.20}$$

$$hN \longrightarrow \infty,\ \text{as } N \to \infty. \tag{2.21}$$

There are several popular kernel functions $\Phi(u)$, which satisfy the above requirements, for instance:

Triangular kernel : $(1 - |u|)\,\mathbb{1}_{\{|u|\leq 1\}}$,

Epanechnikov kernel : $\frac{3}{4}(1 - u^2)\,\mathbb{1}_{\{|u|\leq 1\}}$,

Gaussian kernel : $(2\pi)^{-\frac{1}{2}} \exp -\frac{u^2}{2}$.

In the following, especially for application purposes, we will only consider Gaussian kernels. In case of Gaussian kernels it has been shown that (see [89]) the optimal

choice of the bandwidth, with respect to the Asymptotic Mean Integrated Squared Error (AMISE), equals

$$h := \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma_X \, N^{-\frac{1}{5}} \approx 1.06 \, \hat{\sigma}_X \, N^{-\frac{1}{5}}. \tag{2.22}$$

Where σ_X is the standard deviation of the distribution of X, $\hat{\sigma}_X$ the corresponding empirical estimate and N the number of observations.

Based on the definition of a univariate kernel density estimator, we will now extend the density estimation to the bivariate case.

Definition 2.3.2 (Bivariate kernel density estimator). Given a set of independent observations $\mathcal{D} := \left\{(x^k, y^k)\right\}_{k=1}^{N}$, the bivariate kernel density estimator of a joint PDF $f_{X,Y}$ at point (x, y) is defined by

$$\hat{f}_{X,Y}(x, y \,|\mathcal{D}) = \frac{1}{Nh^2} \sum_{k=1}^{N} \Phi_2 \left(\frac{y^k - y}{h}, \frac{x^k - x}{h}\right). \tag{2.23}$$

Here Φ_2 is the product kernel defined by

$$\Phi_2(u, v) := \Phi(u) \, \Phi(v), \tag{2.24}$$

where $\Phi(.)$ is a univariate kernel function.

Remark 2.3.2. Integrating (2.23) over y yields an estimator of the marginal PDF, i.e.,

$$\begin{aligned}
\hat{f}_X(x \,|\mathcal{D}) &= \int_{\mathbb{R}} \frac{1}{Nh^2} \sum_{k=1}^{N} \Phi_2 \left(\frac{y^k - y}{h}, \frac{x^k - x}{h}\right) dy && (2.25)\\
&= \frac{1}{Nh^2} \sum_{k=1}^{N} \Phi \left(\frac{x^k - x}{h}\right) \int_{\mathbb{R}} \Phi \left(\frac{y^k - y}{h}\right) dy && (2.26)\\
&= \frac{1}{Nh} \sum_{k=1}^{N} \Phi \left(\frac{x^k - x}{h}\right). && (2.27)
\end{aligned}$$

which again is the standard univariate kernel density estimator, as defined in 2.3.1.

2.3.2 Nadaraya Watson estimator

Knowing how to obtain bivariate densities estimates, we can now approach the issue of estimating the conditional expectation $\mathbb{E}[Y \,|X]$ and $\psi(x) = \mathbb{E}[Y \,|X = x]$ respectively.

Merging (2.16) and (2.23) into (2.15) results in the following estimator of the conditional expectation

$$\begin{aligned}
\hat{\psi}(x\,|\mathcal{D}) &= \frac{(Nh^2)^{-1}}{\frac{1}{Nh}\sum_{l=1}^{N}\Phi\left(\frac{x^l-x}{h}\right)}\int_{\mathbb{R}} y\sum_{k=1}^{N}\Phi\left(\frac{x^k-x}{h}\right)\Phi\left(\frac{y^k-y}{h},\right)\,dy && (2.28)\\
&= \frac{(Nh^2)^{-1}}{\frac{1}{Nh}\sum_{l=1}^{N}\Phi\left(\frac{x^l-x}{h}\right)}\sum_{k=1}^{N}\Phi\left(\frac{x^k-x}{h}\right)\int_{\mathbb{R}} y\,\Phi\left(\frac{y^k-y}{h},\right)\,dy && (2.29)\\
&= \frac{(Nh)^{-1}}{\frac{1}{Nh}\sum_{l=1}^{N}\Phi\left(\frac{x^l-x}{h}\right)}\sum_{k=1}^{N}\Phi\left(\frac{x^k-x}{h}\right)y^k && (2.30)\\
&= \sum_{k=1}^{N} y^k\frac{\Phi\left(\frac{x^k-x}{h}\right)}{\sum_{l=1}^{N}\Phi\left(\frac{x^l-x}{h}\right)}. && (2.31)
\end{aligned}$$

The last equation is the well-known Nadaraya-Watson estimator [69, 101].

Definition 2.3.3 (Nadaraya Watson estimator). Let $\mathcal{D} := \left\{(x^k, y^k)\right\}_{k=1}^{N}$ be the set of independent identically distributed observations of the joint random variables (X, Y). Then the Nadaraya Watson estimator of the conditional expectation $\mathbb{E}[Y\,|X = x]$ is the weighted average of the observations $\{y^k\}_{k=1}^{N}$, i.e.,

$$\hat{\psi}(x\,|\mathcal{D}) = \sum_{k=1}^{N} y^k\, g_k(x), \tag{2.32}$$

where the weight for the k^{th} observation is given by

$$g_k(x) := \frac{\Phi\left(\frac{x^k-x}{h}\right)}{\sum_{l=1}^{N}\Phi\left(\frac{x^l-x}{h}\right)}. \tag{2.33}$$

The Nadaraya Watson estimator assigns the largest weight to the observations y^k, for which the corresponding x^k is closest to x. In this sense, the Nadaraya Watson estimator can be interpreted as a local convex combination of the observations $\{y^k\}_{k=1}^{N}$, where *local* refers to the explanatory variable X.

2.3.3 Examples

Before we transfer the former to the context of EEG analyses in the next chapter, we would like to give two more brief, sophisticated examples of conditional expectation.

The first example corresponds to the typical scenario of estimating a regression function $\psi : X \mapsto Y$, for two real valued continuous random variables X and Y under the assumption of additive Gaussian noise.

The second example is more complicated and illustrates the estimation of conditional expectation $\mathbb{E}[\mathbb{Y}\,|X = x]$ of a stochastic process $\mathbb{Y}$ given a real valued continuous random variable X. It reveals a clear underlying functional dependency of the dynamic system $\mathbb{Y}$ on an initial environmental state variable, represented by X.

Conditional expectation of a random variable

In contrast to example 2.2.1 we will now illustrate the concept of conditional expectation in terms of two continous random variables X and Y defined on a common probability space $(\Omega, \mathcal{F}, P)$. To this end, let X be uniformly distributed and Y a function $f(X)$ with additive, independent Gaussian noise. In particular we choose

$$Y = \operatorname{sinc}(X) + \eta, \tag{2.34}$$

$$X \sim U_{[-4,4]}, \tag{2.35}$$

$$\eta \sim \mathcal{N}\left(0, \frac{1}{4}\right). \tag{2.36}$$

Drawing 200 i.i.d. samples from Ω results in a sample set $\{\omega_1, \ldots, \omega_{200}\}$. Using the convenient notation of $x^k = X(\omega_k)$ and $y^k = Y(\omega_k)$ respectively, the set of observations corresponds to $\mathcal{D} = \{(x^k, y^k)\}_{k=1}^{200}$. Based on this set of observations we estimate

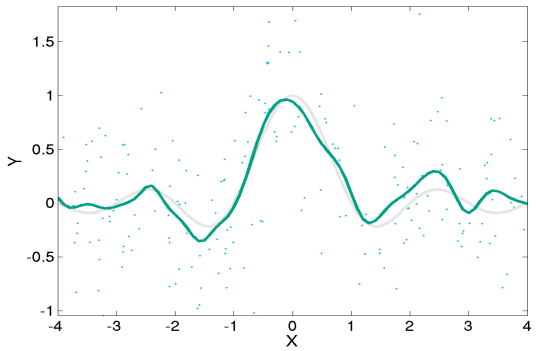


Figure 2.6: The estimated conditional expectation $\hat{\psi}(x\,|\mathcal{D})$ (green) along with the true functional relationship $\mathbb{E}[Y\,|X = x]$ (gray) for the two random variables X and Y, defined in (2.35)) and (2.34). The light blue points represent the set of 200 i.i.d. observations.

the conditional expectation $\mathbb{E}[Y\,|X = x]$ by means of a Nadaraya Watson estimator

$\hat{\psi}(x\,|\mathcal{D})$, (cf. definition 2.3.3). More precisely we use the Nadaraya Watson estimator with Gaussian kernel functions and set the bandwidth parameter to its default value (cf. (2.22)). Fig. 2.6 summarizes the results by presenting a scatter plot of the samples along with the estimated conditional expectation. Note that due to the symmetric additive Gaussian noise, the (mathematically) true functional relationship of the conditional expectation equals $\mathbb{E}[Y\,|X=x] = \mathrm{sinc}(x)$. Remarkably, despite the small sample size, the estimated conditional expectation $\hat{\psi}(x\,|\mathcal{D})$ resembles the truth quite accurately.

Conditional expectation of a stochastic process

The setting for the second example is a bit more complicated, since we have to define a stochastic process together with a random variable on a common probability space $(\Omega, \mathcal{F}, P)$. For this purpose let the random variable X be uniformly distributed on $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and let $\mathcal{T} = \{-2\pi + \frac{j-1}{199}(2\pi - 2)\}_{j=1}^{200} \subseteq [-2\pi, -2]$ be an index set, referring to the sample points in the temporal domain. For this set we define a stochastic process $\mathbb{Y}$ as the family of random variables $(Y_t)_{t\in\mathcal{T}}$, such that for a given $t \in \mathcal{T}$

$$Y_t := \mathrm{sinc}(t) + \eta_t \cdot \frac{2\pi + t}{2(2\pi - 2)} \cdot \sin\left(X + \frac{5}{2}t\right), \tag{2.37}$$

$$X \sim \mathrm{U}_{[-\frac{\pi}{2}, \frac{\pi}{2}]}, \tag{2.38}$$

$$\eta_t \sim \mathrm{U}_{[0,1]}. \tag{2.39}$$

Under the assumption of mutually independent X and η_t the stochastic process consists of a deterministic component, i.e., the sinc function and additive noise. However, this additive noise depends on the random variable X, and thus $\mathbb{Y}$. Remember that an elementary event $\omega \in \Omega$ simultaneously determines the entire process, yielding a path $Y_\cdot(\omega)$. Drawing 200 i.i.d. samples from Ω results in a sample set $\{\omega_1, \ldots, \omega_{200}\}$ and yields a set of two hundred single trials along with the corresponding realization of $\mathbb{Y}$ and X, i.e., $\mathcal{D} := \{(Y_\cdot^k, X^k)\}_{k=1}^{200}$. Based on this set of observations we estimate the conditional expectation $\mathbb{E}[\mathbb{Y}\,|X]$, as the family of functions $\hat{\psi}_t(x\,|\mathcal{D}) = (\,\mathbb{E}[Y_t\,|X=x])_{t\in\mathcal{T}}$. To this end we apply the Nadaraya-Watson estimator (cf. definition 2.3.3) with Gaussian basis functions and set the bandwidth according to (2.22). Fig. 2.7 and Fig. 2.8 visually represent the results in different ways. Fig. 2.7 presents only the observed single trials $\{Y_\cdot^k\}_{k=1}^{200}$ and highlights three particularly chosen traces of $\mathbb{E}[\mathbb{Y}\,|X=x]$, namely at $x \in \{-\frac{\pi}{3}, 0, \frac{\pi}{3}\}$. Due to the specific construction of $\mathbb{Y}$ all paths start at the same

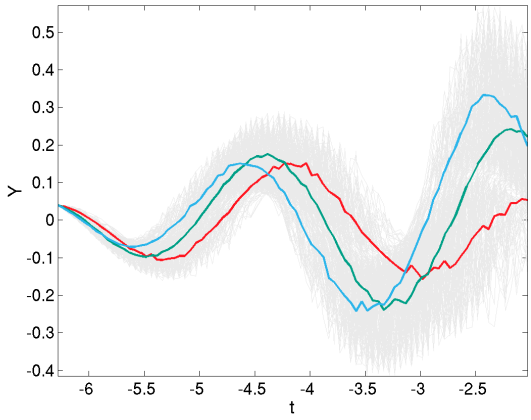


Figure 2.7: Toy example of the conditional expectation of a stochastic process. Two hundred realizations of the stochastic process $\mathbb{Y}$ are displayed as single traces (gray). The colored traces refer to the estimated conditional expectation $\mathbb{E}[\mathbb{Y}\,|X = x]$ at three different values of the initial state x, namely $\{-\frac{\pi}{3}, 0, \frac{\pi}{3}\}$ (red, green and blue).

position, but dependent on the initial state X, the conditional expectations diverge quickly for the three different values. In case of an univariate explanatory variable X, the dependency on the initial state can be demonstrated in even more detail, shown in Fig. 2.8. Here the vertical axis refers to the explanatory variable X, while the horizontal axis again indicates time. The particular value of $\hat{\psi}_t(x\,|\mathcal{D}) = \mathbb{E}[Y_t\,|X = x]$ is presented using a color coding scheme. In order to generate this drawing, we evaluated (2.32) for a

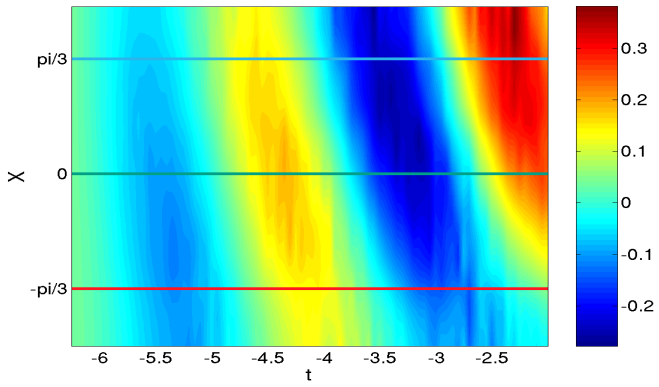


Figure 2.8: Conditional expectation $\mathbb{E}[\mathbb{Y}\,|X]$ of a stochastic process. The horizontal axis represents the time, while the vertical axis distinguishes different conditions X. Values of $\mathbb{E}[\mathbb{Y}\,|X]$ are visualized using a color coding scheme. The three highlighted horizontal lines reflect the conditional expectation at those values that were already emphasized in Fig. 2.7.

multitude of different values of the explanatory variable from the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. The results obtained at the 64 equally spaced different values are summarized in Fig. 2.8. In this X-t plane, each horizontal line corresponds to a certain value of the explanatory variable, e.g., the light blue line corresponds to $X = \frac{\pi}{3}$. In contrast a vertical line at time t corresponds to the regression function $\psi(x\,|\mathcal{D}) : X \mapsto Y_t$. From Fig. 2.7 and Fig. 2.8 the dependency of the stochastic process on the explanatory variable X in terms of different latencies and magnitudes of the extreme values can be clearly inferred. For example, at $X = \frac{\pi}{3}$ there is a clear deep minima around $t = -3.5$, which resides much later in time at $t = -3$ for $X = -\frac{\pi}{3}$, and is additionally flattened.

Summary

In this chapter, we introduced the mathematical concept of conditional expectations. We pointed out that the conditional expectation $\mathbb{E}[Y\,|X]$ by itself is a random variable. Nevertheless, there exists a mapping ψ, such that $\psi(x) = \mathbb{E}[Y\,|X = x]$. Furthermore, this mapping can be inferred from a given set of observations by means of the Nadaraya Watson estimator. We also showed how the framework of conditional expectation can be extended to stochastic processes.

Chapter 3

Conditional ERD

(Towards Single Trial ERD)

The variability of *single trial* responses has to be taken in to account and sufficiently explained, when describing human cortical neurophysiology. Such a framework should therefore explicitly incorporate dependencies of the stimulus processing on internal (mental) or external (environmental) conditions. Investigations of phenomena of this kind require both suitable data analytical concepts and elaborate experimental paradigms which enable a reliable discovery of functional relationships between explanatory variables and the single trial characteristics of brain responses. So far investigations of trial-to-trial variability were mostly limited to event-related potentials (ERPs). Here the impact of various external and internal factors on the latency and the magnitude of ERP components has been studied intensively [30, 35, 36, 41, 73]. On the opposite, the variability of ERD has been investigated only occasionally and mostly with respect to a few external factors [31, 72, 92]. Despite the general interest in the topic, there rarely exist studies on a relationship between ERD and internal factors (see, e.g., [19, 102]). From our point of view, this is basically due to the absence of an appropriate model for analyzing such dependencies. It is the ultimate ambition of this chapter to bridge this gap by introducing a novel framework to illustrate how such relationships can be analyzed. At this point we would like to stress that it is the aim of this chapter to provide the fundamentals of the data analytic concept rather than to pursue a comprehensive studies, which prove certain neurophysiological hypotheses. That's why the application section is designed to serve as a proof of concept. Here, subsequent to the analysis of

artificially generated state dependent ERD data we will highlight the potential of our framework by means of analyzing EEG data of a single subject. As a matter-of-fact this does not provide a comprehensive neurophysiological study, but it already envisions possible directions of future investigation.

Typically ERD is quantified as an averaged response over a set of single trials. This approach can hardly disclose any state dependencies of the ERD on internal or external factors. As opposed to the *averaged ERD* we will denote the novel approach by *conditional ERD*. Moreover, we will show that the use of a fixed reference interval, as it is incorporated in the conventional ERD model, hampers a reliable study of the variability of ERD responses and may lead to spurious observations of ERD/S. To overcome this issue, we generalize the conventional ERD framework with respect to the reference condition. To additionally distinguish between these opposite approaches we refer to them as the *conventional* and the *generalized ERD* framework, respectively.

The chapter is organized as follows: at the beginning, we formulate the *conventional ERD* framework as a stochastic model. To this end, we first interpret EEG single trials as realizations of a stochastic process and consecutively construct a stochastic model for the *conventional averaged ERD* framework. From this conventional model we derive a naive extension towards the analysis of *conditional ERD*. In a second line of argument, we first generalize the *conventional averaged ERD* framework, with respect to the definition of the referential baseline condition. This generalized framework is then expanded towards the analysis of *conditional ERD*. The concluding application section starts with a comparison of the *conventional* and the *generalized ERD* frameworks and their individual capabilities to recover different, known dependencies of ERD characteristic in a fully controlled, artificial environment of simulated single trial data. This comparison will basically expose the limitations of the conventional framework. The comparative analysis of artificially generated data is followed by an investigation of μ-rhythm ERD, evoked by median nerve stimulation. There we will present three case studies of analyzing the dependency of the ERD on three different internal factors, i.e., on initial states of activation of different cortical areas.

3.1 Preliminaries

3.1.1 Stochastic model for single trial data

Let $(\Omega, \mathcal{F}, P)$ be a probability space. On this probability space we will now consecutively construct the different stochastic ERD models. We start with a common temporal index set $\mathcal{T} \in \mathbb{Z}$, indicating the sampling time instances of each single trial. Then the single trial encephalogram data can be represented as a d-dimensional real valued stochastic process $\mathbb{X} = (\boldsymbol{X}_t)_{t\in\mathcal{T}}$, where d refers to the number of electrodes:

$$(\Omega, \mathcal{F}, P) \xRightarrow{\mathbb{X}} (\mathbb{R}^d, \mathcal{B}^d) \tag{3.1}$$

In this setting each elementary event $\omega \in \Omega$ corresponds to a particular realization $\boldsymbol{X}_\cdot(\omega)$ of the stochastic process and thereby represents one observed single trial. Consequently a set of K elementary events $\omega_1, \ldots, \omega_K$ drawn i.i.d. from Ω corresponds to a set of K independently observed single trials $\{\boldsymbol{X}_\cdot^k\}_{k=1}^K$. For notational convenience we will denoted the k^{th} single trial by $\boldsymbol{X}_\cdot^k := \boldsymbol{X}_\cdot(\omega_k)$.

3.2 Conventional ERD framework

The following section introduces the stochastic framework underlying the *conventional ERD* measures [40, 77]. At the beginning we will expand the stochastic model (3.1) of encephalogram recordings by additionally representing the rhythmic activity under consideration, e.g., a projection on the μ-rhythm of a single hemisphere. This first model will already be sufficient to enable the *conventional averaged ERD* analysis. Later on we will consider a naive extension of this model, to enable the conventional framework to analyze *conditional ERD*. To this end, we add an explanatory variable, whose influence on the characteristic of the ERD is to be analyzed. Each introduced stochastic model is always accompanied by a derived ERD measure and an associated empirical estimator.

3.2.1 Averaged ERD

The analysis of ERD is always concerned with power modulations of a certain rhythmic activity. Consequently the model (3.1) is now equipped with an additional

stochastic process reflecting the power envelope of the rhythmic activity under consideration.

Stochastic model

Remember that according to remark 2.1.3, $\mathbb{Y} := A\mathbb{X} * b$ defines a stochastic process on the common probability space Ω for an arbitrary matrix A and an FIR filter b. Hence, let $w \in \mathbb{R}^d$ be an appropriate spatial filter, focussing on the cortical region of interest and b denote a suitable FIR bandpass filter for the corresponding frequency band, e.g., a complex Morlet wavelet [96]. Then, $\mathbb{Y} := \left|w^\top \mathbb{X} * b\right|^2$ defines an univariate stochastic process that represents the instantaneous power of the rhythmic activity under investigation.

$$\begin{array}{ccc} (\Omega, \mathcal{F}, P) & \xrightarrow{\quad\mathbb{X}\quad} & (\mathbb{R}^d, \mathcal{B}^d) \\ & \searrow^{\mathbb{Y}=\left|w^\top \mathbb{X} * b\right|^2} & \\ & & (\mathbb{R}, \mathcal{B}) \end{array} \tag{3.2}$$

Remember that the quantification of ERD requires a reference and an event-related condition. Consequently we suppose the common index set to be composed of two disjoint subsets, i.e., $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1$, where $\mathcal{T}_0$ serves as reference interval, while $\mathcal{T}_1$ indicates the disjoint period, when the ERD effect is to be quantified. Typically the stochastic process at rest is assumed to be stationary, i.e., it has identical distributions for all $t \in \mathcal{T}_0$. This implies that the statistical moments are independent of the time index, i.e., for $s, t \in \mathcal{T}_0$ it follows that $\mathbb{E}[Y_t] = \mathbb{E}[Y_s]$. Thus for reasons of robustness the reference power is usually defined as average across the reference interval

$$Y_{\text{REF}} := \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} Y_t. \tag{3.3}$$

Conventionally the ERD is expressed as the relative deviation of the power from this reference and hence it follows

$$\text{ERD}_{\text{CONV}}[t] = \frac{\mathbb{E}[Y_t]}{\mathbb{E}[Y_{\text{REF}}]} - 1, \quad t \in \mathcal{T}. \tag{3.4}$$

Remark 3.2.1. Note that (3.4) describes the ERD measure according to the power method [77]. With minor modifications the intertrial variance method [40] (cf. also

Fig. 1.3) can be represented as well. To this end, the stochastic process $\mathbb{Y}$ is substituted by $\mathbb{Y} := w^\top \mathbb{X} * b$ and the expectation values in (3.4) change accordingly such that they reflect the intertrial variance, i.e., $\mathbb{E}\Big[|Y_t - \mathbb{E}[Y_t]|^2\Big]$. In this context, $\mathbb{E}[Y_t]$ corresponds to the phase locked components of the spatially and bandpass filtered signal. Since the intertrial variance method can be easily derived from the power method, but requires a slightly more complicated notation, we will subsequently restrict ourselves solely to the power method.

Empirical estimator

In case of the stochastic model (3.2) a finite set of K observations corresponds to $\mathcal{D} = \{Y_\cdot^k\}_{k=1}^K$. Note that we suppressed the contextual information about the stochastic process $\mathbb{X}$, as $\mathbb{Y}$ already contains the relevant information. Using the index set $\mathcal{T}_0 \subset \mathcal{T}$ as reference condition, the empirical estimates of the expected reference band power and the expected power at $t \in \mathcal{T}$ are obtained as:

$$\hat{P}_{\text{REF}} = \frac{1}{K \cdot |\mathcal{T}_0|} \sum_{k=1}^{K} \sum_{t \in \mathcal{T}_0} Y_t^k \tag{3.5}$$

$$\hat{P}(t) = \frac{1}{K} \sum_{k=1}^{K} Y_t^k, \qquad t \in \mathcal{T}, \tag{3.6}$$

such that finally the *averaged ERD* at $t \in \mathcal{T}$ is estimated by

$$\text{ERD}_{\text{CONV}}[t\,|\mathcal{D}] = \frac{\hat{P}(t)}{\hat{P}_{\text{REF}}} - 1. \tag{3.7}$$

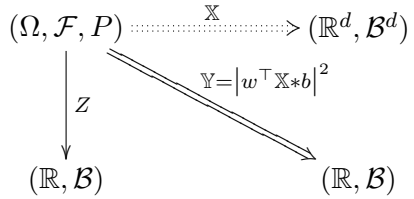
We refer to (3.7) as the empirical estimator of the *conventional averaged ERD*.

3.2.2 Conditional ERD

In order to enable the analysis of dependencies of the ERD characteristic on external or internal states, we naively extend the conventional stochastic ERD model, introduced in the previous paragraph. However, later on in the application section 3.4, we will show that this naive approach has the drawback of possibly giving rise to spurious observations of ERD. As we will point out, the reason for the failure originates from the definition of the reference condition. Hence in section 3.3 we will generalize the *conventional ERD* framework with respect to the reference condition. There we will also derive an improved *conditional ERD* estimator.

Stochastic Model

Let Z denote an explanatory variable, whose influence on the characteristic of the ERD is to be analyzed. Typically Z is a measurable function of the process $\mathbb{X}$, for instance the initial power of a certain cortical rhythm at the event onset. But Z may also reflect an external condition, such as the intensity of the stimulation, the interstimulus interval or any other parameter whose impact on the ERD characteristic shall be investigated. However, we incorporate Z without further specification as an additional random variable into the common probability space Ω:



(3.8)

This enables us to extend (3.4) towards the quantification of *conditional ERD* by simply replacing all expected values by their corresponding conditional expectations:

$$\mathrm{ERD}_{\text{CONV}}[t\,|Z] \quad = \quad \frac{\mathbb{E}[Y_t\,|Z]}{\mathbb{E}[Y_{\text{REF}}\,|Z]} - 1. \tag{3.9}$$

Here the conditional expectation of the reference power is given by the averaged conditional expectations of the individual random variables

$$\mathbb{E}[Y_{\text{REF}}\,|Z] \quad = \quad \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0}\mathbb{E}[Y_t\,|Z]\,. \tag{3.10}$$

Remember that conditional expectations are random variables and so is the *conditional ERD*. Fortunately, according to proposition 2.2.3 for any $t \in \mathcal{T}$ there exists a mapping $\psi_t : \mathbb{R} \to \mathbb{R}$ such that

$$\psi_t(z) \quad = \quad \mathbb{E}[Y_t\,|Z = z]\,. \tag{3.11}$$

Substituting the conditional expectations in (3.9) accordingly, enables us to express the *conditional ERD* at an arbitrary instance $Z = z$ as

$$\mathrm{ERD}_{\text{CONV}}\,[t\,|Z = z] = \frac{\psi_t(z)}{\psi_{\text{REF}}(z)} - 1, \tag{3.12}$$

where $\psi_{\text{REF}}(z)$ is defined according to (3.10). This measure quantifies the *conditional ERD* given the observation of state $Z = z$ and thereby establishes a functional relationship between the state variable Z and the ERD.

Empirical estimator

In terms of the stochastic model (3.8) a finite set of K observations comprises the realizations of the explanatory variable Z along with the observed paths of the stochastic process $\mathbb{Y}$. Thus we denote the set of observed samples by $\mathcal{D} = \{(Y_\cdot^k, z^k)\}_{k=1}^K$.

In order to derive an empirical estimator for *conditional ERD* measure (3.12), we replace $\psi_t(z) = \mathbb{E}[Y_t \,|Z = z]$ with the corresponding Nadaraya Watson estimator (cf. definition 2.3.3) that is given by

$$\hat{\psi}_t(z\,|\mathcal{D}) = \sum_{k=1}^{K} Y_t^k \, g_k(z), \qquad t \in \mathcal{T}. \tag{3.13}$$

Here the weight $g_k(z)$ of the k^{th} observation depends on the choice of the kernel function Φ and is given by

$$g_k(z) = \Phi\left(\frac{z^k - z}{h}\right)\left(\sum_{l=1}^{K} \Phi\left(\frac{z^l - z}{h}\right)\right)^{-1}. \tag{3.14}$$

Moreover, using $\hat{\psi}_{\text{REF}}(z\,|\mathcal{D}) = \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0}\hat{\psi}_t(z\,|\mathcal{D})$ as reference, the empirical estimator for *conditional ERD* is finally determined by

$$\text{ERD}_{\text{CONV}}\,[t\,|Z = z, \mathcal{D}] = \frac{\hat{\psi}_t(z\,|\mathcal{D})}{\hat{\psi}_{\text{REF}}(z\,|\mathcal{D})} - 1. \tag{3.15}$$

We refer to (3.15) as the empirical estimator for the *conventional conditional ERD*.

3.3 Generalized ERD framework

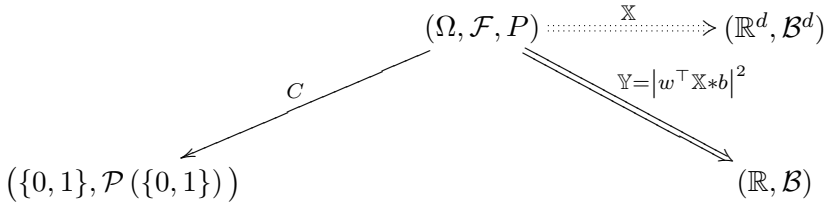
In the following we derive a novel, alternative measure for the quantification of the *conditional ERD*. To this end, we generalize the ERD framework with respect to the reference condition. Instead of using a fixed reference value, the novel measure contrasts the evolution (time course) of the instantaneous power (envelopes) between a rest and an event-related condition. In particular we consider the (conditional) expectation of the envelope of the ongoing dynamics as reference and define the generalized (conditional) ERD as the relative deviation of the (conditionally) expected power envelope of the event-related dynamics.

3.3.1 Generalized averaged ERD

Before elaborating on the analysis of *conditional ERD*, we introduce the generalized stochastic framework with respect to the calculation of *averaged ERD*. In particular we extend the stochastic model (3.2) further such that it comprises the observation of both dynamics, i.e., perturbed (event-related) and unperturbed (ongoing) activity, respectively.

Stochastic model

We additionally equip the common probability space Ω with an auxiliary binary random variable C that distinguishes between perturbed and unperturbed activity.



(3.16)

By convention, $C = 0$ refers to the rest condition, while $C = 1$ corresponds to event-related single trials. Based on the above stochastic model, we define the *generalized averaged ERD* as follows:

$$\mathrm{ERD}_{\text{GEN}}[t] := \frac{\mathbb{E}[Y_t \,|C = 1]}{\mathbb{E}[Y_t \,|C = 0]} - 1, \qquad t \in \mathcal{T}. \tag{3.17}$$

Consequently for each $t \in \mathcal{T}$ we use a separate reference that is given by $\mathbb{E}[Y_t \,|C = 0]$, and determine the ERD at time t as the relative deviation of the event-related expectation $\mathbb{E}[Y_t \,|C = 1]$ from the reference.

Remark 3.3.1. Under the assumption of a stationary dynamics at rest ($C = 0$), the *conventional* (3.4) and the *generalized* (3.17) ERD measurements are equivalent. This directly follows from the following consideration: For a stationary stochastic process $\mathbb{Y}$ the distributions of Y_t are identical for all $t \in \mathcal{T}$. This implies that all statistical moments of the stochastic process representing the rest condition are constant across time, i.e., particularly $\mathbb{E}[Y_t \,|C = 0] = \mathbb{E}[Y_s \,|C = 0]\ \forall s, t \in \mathcal{T}$. Hence the formulas (3.4) and (3.17) of the *conventional* and the *generalized ERD* are equivalent and lead to an identical measurement of the *averaged ERD*. However, as we will see in the first artificial example, if the process at rest is non-stationary the measurements give rise

to differently estimated ERD. A typical scenario, where the stationarity assumption is expected to be violated, is given in the context of repeated stimulations at a short inter-stimulus interval (ISI). For instance if the stimuli are randomly delivered ($P(C=1)=p$, $P(C=0)=1-p$) at the time instance, when an ERS response of the preceding stimulus is likely to show up, then the expectation of the dynamics at rest, i.e., $\mathbb{E}[Y_t\,|C=0]$ is expected to exhibit a negative trend and consequently the stochastic process is non-stationary. However, typically the interval between two consecutive events is chosen sufficiently large such that the resulting process at rest can be considered stationary.

Empirical estimator

Suppose a finite set of $2K$ observations drawn i.i.d. from Ω. In the context of the generalized stochastic model (3.16) this corresponds to $\mathcal{D} := \left\{\left(Y_\cdot^k, c^k\right)\right\}_{k=1}^{2K}$. Moreover, as the random variable C is typically under the control of the investigator, we further assume an equal number of K trials observed at both conditions (rest and excitation). The subsets of trials corresponding to each condition are denoted by $\mathcal{C}^0 = \{k : C(\omega^k) = 0\}$ and $\mathcal{C}^1 = \{k : C(\omega^k) = 1\}$, respectively. The empirical estimators at time $t \in \mathcal{T}$ of the expected reference power and the expected event-related power are obtained straightforwardly as:

$$\hat{P}_{\text{REF}}(t) = \frac{1}{K}\sum_{k\in\mathcal{C}^0} Y_t^k, \qquad \text{and} \qquad \hat{P}(t) = \frac{1}{K}\sum_{k\in\mathcal{C}^1} Y_t^k. \tag{3.18}$$

Thus given a set of observations the *averaged ERD* at $t \in \mathcal{T}$ is estimated by

$$\text{ERD}_{\text{GEN}}[t\,|\mathcal{D}] = \frac{\hat{P}(t)}{\hat{P}_{\text{REF}}(t)} - 1. \tag{3.19}$$

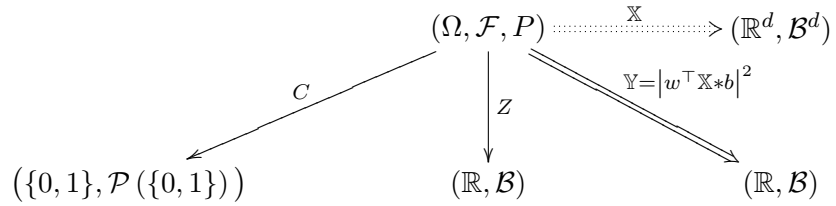
We refer to (3.19) as the empirical estimator of the *generalized averaged ERD*.

3.3.2 Generalized conditional ERD

Similar to what was described in section 3.2.2 we expand the *generalized ERD* framework towards the analysis of dependencies of the ERD characteristic on external or internal states.

Stochastic Model

We equip the stochastic model (3.16) with an additional explanatory variable Z, whose influence on the characteristic of the ERD is to be investigated:



(3.20)

Furthermore we suppose the random variable Z and C to be mutually independent, i.e., $P(Z, C) = P(Z)P(C)$. Basically, this assumption ensures that Z has identical conditional distributions at the rest and at the event-related condition. Using this stochastic model, we quantify the conditional ERD given the variable Z in terms of conditional expectations. Along the lines of the power method [77] we define the *generalized conditional ERD* at $t \in \mathcal{T}$ as follows:

$$\mathrm{ERD}_{\text{GEN}}\left[t \,|Z\right] := \frac{\mathbb{E}[Y_t \,|Z, C=1]}{\mathbb{E}[Y_t \,|Z, C=0]} - 1. \tag{3.21}$$

Remember, according to proposition 2.2.3 for $t \in \mathcal{T}$ there exist functions $\psi_t^0, \psi_t^1 : \mathbb{R} \to \mathbb{R}$ such that

$$\psi_t^0(z) = \mathbb{E}[Y_t \,|Z = z, C=0] \qquad \text{and} \qquad \psi_t^1(z) = \mathbb{E}[Y_t \,|Z = z, C=1]\,. \tag{3.22}$$

Substituting the conditional expectations in (3.21) accordingly, we obtain the *conditional ERD* given the state $Z = z$ as

$$\mathrm{ERD}_{\text{GEN}}\left[t \,|Z = z\right] = \frac{\psi_t^1(z)}{\psi_t^0(z)} - 1. \tag{3.23}$$

Verbalized, given the observation $Z = z$, we have a certain expectation of the time course of the unperturbed dynamics, namely $\left(\psi_t^0(z)\right)_{t\in\mathcal{T}}$, which serves as reference dynamics. The *conditional ERD* is then defined as the relative deviation of the expected event-related dynamics $\left(\psi_t^1(z)\right)_{t\in\mathcal{T}}$, from this reference.

Remark 3.3.2. This *conditional ERD* measure can be expanded from single instances to a subset of the state space. For this purpose let $\mathcal{Z} \in \mathcal{B}$ denote a set of the Borel σ-algebra. Then for $Z \in \mathcal{Z}$ the *conditional ERD* is calculated as

$$\mathrm{ERD}_{\text{GEN}}\left[t \,|Z \in \mathcal{Z}\right] = \frac{\int_{\mathcal{Z}} \psi_t^1(z)\, dPZ^{-1}}{\int_{\mathcal{Z}} \psi_t^0(z)\, dPZ^{-1}} - 1. \tag{3.24}$$

Moreover, setting $\mathcal{Z} = Z(\Omega)$ (the image of Ω under the mapping Z) and using the property $\mathbb{E}[\,\mathbb{E}[Y_t\,|Z]] = \mathbb{E}[Y_t]$ of the conditional expectation (cf. proposition 2.2.2), the *conditional ERD* becomes equivalent to the *generalized averaged ERD* measure, i.e.,

$$\mathrm{ERD}_{\text{GEN}}\,[t\,|Z \in Z(\Omega)] \;=\; \frac{\mathbb{E}[Y_t\,|C=1]}{\mathbb{E}[Y_t\,|C=0]} - 1. \tag{3.25}$$

This equivalence is quite intuitive, as the condition $Z \in Z(\Omega)$ is uninformative, i.e., provides no further information of the whereabouts of ω. Consequently it should yield the identical measure as the *averaged ERD*.

Empirical estimator

Let us suppose a finite set of $2K$ observations drawn i.i.d. from Ω. In terms of the stochastic model (3.20) this corresponds to $\mathcal{D} = \left\{\left(Y_\cdot^k, c^k, z^k\right)\right\}_{k=1}^{2K}$. Again, we assume an equal number of K observed trials at both conditions and denote the corresponding subsets of trials by $\mathcal{C}^0 = \{k : C(\omega^k) = 0\}$ and $\mathcal{C}^1 = \{k : C(\omega^k) = 1\}$, respectively.

In order to derive an empirical estimator for (3.25), we replace $\psi_t^0(z)$ and $\psi_t^1(z)$ by the corresponding Nadaraya Watson estimators (cf. definition 2.3.3), i.e., $\hat{\psi}_t^0(z\,|\mathcal{D})$ is estimated as

$$\hat{\psi}_t^0(z\,|\mathcal{D}) = \sum_{k\in\mathcal{C}^0} (Y_t^k)\, g_k^0(z), \tag{3.26}$$

where the weight $g_k^0(z)$ of the k^{th} observation is given as

$$g_k^0(z) = \Phi\left(\frac{z^k - z}{h}\right)\left(\sum_{l\in\mathcal{C}^0} \Phi\left(\frac{z^l - z}{h}\right)\right)^{-1}. \tag{3.27}$$

Accordingly, we define the empirical estimator $\hat{\psi}_t^1(z\,|\mathcal{D})$ by just exchanging the set $\mathcal{C}^0$ with $\mathcal{C}^1$. Note that the local smoothing depends on the choice of the kernel function Φ as well as on its bandwidth h.

Finally, the empirical estimator for *conditional ERD* is determined by

$$\mathrm{ERD}_{\text{GEN}}\,[t\,|Z=z,\mathcal{D}] = \frac{\hat{\psi}_t^1(z\,|\mathcal{D})}{\hat{\psi}_t^0(z\,|\mathcal{D})} - 1. \tag{3.28}$$

We refer to (3.28) as the empirical estimator of the *generalized conditional ERD*.

3.4 Application

The following applications will serve as a proof of concept of the proposed framework. We will point up the potential of the proposed *generalized ERD* framework for the analysis of *averaged* and *conditional ERD* and expose the limitations of the conventional methods. In order to obtain a first reliable evaluation, we apply both concepts to artificially generated data. The application in a controlled, artificial set-up will primarily reveal that the *conventional ERD* may give rise to observations of spurious ERD. Afterwards we investigate the state dependencies of μ-rhythm ERD evoked by median nerve stimulation. More precisely, we will study its dependency on the magnitude of the local pre-stimulus μ-activity, on the magnitude of the occipital pre-stimulus α-activity and finally on the magnitude of the ERS response of the preceding stimulus. However, the aim of this concluding investigation is rather to emphasize the potential that is offered by the *conditional ERD* framework, than to thoroughly prove any neurophysiological hypothesis about the dependency of the ERD characteristic, and is therefore restricted to data of a single subject.

3.4.1 Artificial data

In order to compare the capabilities of the *generalized* and the *conventional conditional ERD* framework properly, we generate three sets of surrogate ERD data that exhibit different kinds of dependency on an explanatory variable Z. The first data set will solely comprise a dependency of the event-related power envelope on the explanatory variable, i.e., only the stochastic process during stimulation exhibits this dependency, while the ongoing dynamics as such is independent, but has an inherent non-stationary characteristic. In contrast the second data set will solely comprise a dependency of the ongoing dynamics on the explanatory variable, while the ERD characteristic is unaffected. Finally, the last artificial data set will be most complicated and comprises both kinds of dependencies, i.e., the stochastic process at rest as well as the ERD characteristic exhibit separate dependencies on the explanatory variable Z.

Data generation

In order to allow for a first comparison of both approaches in relation to the differently accentuated dependencies on the explanatory variable, we define the three artificial

data sets based on a common setup. To this end, let us suppose $\mathcal{T} \subset [-\pi, \pi]$ as index set, comprising 100 equidistant sample points. Moreover, let us define the conditional stochastic processes $\mathbb{Y}$ of the instantaneous power at rest and at the event-related condition separately such that for $t \in \mathcal{T}$

$$Y_t(\omega) \quad := \quad \begin{cases} f_t(\omega) & C(\omega) = 0, \ i.e., \ at \ rest, \\ \alpha_t(\omega) \cdot f_t(\omega) & C(\omega) = 1, \ i.e., \ event\text{-}related. \end{cases} \tag{3.29}$$

Here $f_t \geq 0$ represents the power envelope of the unperturbed ongoing activity, while the ERD is emulated by a multiplicative dampening. In particular we use the following parameterized version of f_t and α_t:

$$f_t(\theta, \beta) \quad = \quad \frac{3}{2} + \sin(t + \theta) + \beta t, \quad t \in \mathcal{T}, \ \theta \in [0, 2\pi], \ \beta \in \left\{ -\frac{1}{3\pi}, 0 \right\} \tag{3.30}$$

$$\alpha_t(s) \quad = \quad \left[1 + \frac{(3-s)}{4} \left((t-s)^2 - 1 \right) \mathbb{1}_{|t-s| \leq 1} \right], \quad t \in \mathcal{T}, s \in [0, 1]. \tag{3.31}$$

The parameters θ and β influence the shape of the power envelope, where θ determines the initial phase of the power envelope, while β controls the presence of a negative linear trend. The multiplicative dampening function $\alpha_t(s)$ offers the parameter s to control the latency and the magnitude of the maximum attenuation. Here the extreme values are taken at $t = s$, with a magnitude of $\frac{1-s}{4}$. Different realizations of the dampening and the power envelope at several parameter values are exemplified in Fig. 3.1.

Based on the common architecture (3.29)–(3.31), we derive three artificial examples by defining random variables $\theta(\omega)$ and $s(\omega)$, while β is used in a strict deterministic fashion. Then the different interdependencies of the explanatory variable Z and the random variables θ and s determine the individual characteristic of each data set.

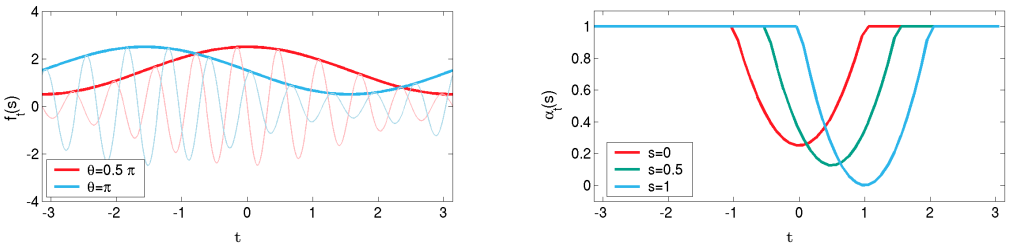


Figure 3.1: The left panel exemplifies two different realizations of the parameterized function $f_t(\theta, 0)$, simulating the power envelopes of rhythmic ongoing activity. To establish a better understanding, we also depicted two corresponding arbitrary oscillations (red and blue thin lines) beside the power envelopes. Right panel: The artificially generated multiplicative dampening factor $\alpha_t(s)$ at three different parameter values. The parameter s determines the latency and the magnitude of the dampening.

For all three data sets we assume the explanatory variable Z to be uniformly distributed on the interval $[0, 1]$, i.e.,

$$Z \sim \mathrm{U}_{[0,1]}. \tag{3.32}$$

DATA SET I The first data set solely comprises a dependency of the ERD characteristic on the explanatory variable, i.e., only the dampening process $\alpha_t(s)$ is affected, while the ongoing dynamics $f_t(\theta, \beta)$ varies independently, but exhibits a deterministic negative trend. In particular we define the parameters θ, β and s as follows

$$\theta \sim \mathrm{U}_{[0,2\pi]}, \qquad \beta = -\frac{1}{3\pi}, \qquad s = Z. \tag{3.33}$$

Additionally we assume mutually independence of θ and Z.

DATA SET II In the second example of surrogate ERD data, we solely implement a dependency of the initial phasing of the power envelope on the explanatory variable, while the dampening is deterministic. Consequently we remove the negative trend and let

$$\theta = 2\pi Z, \qquad \beta = 0, \qquad s = 0. \tag{3.34}$$

DATA SET III The third and last example of surrogate ERD data comprises, in the absence of a linear trend, dependencies of s and θ on the explanatory variable Z. In particular we set

$$\theta = 2\pi Z, \qquad \beta = 0, \qquad s = Z. \tag{3.35}$$

Consequently both the phasing of the power envelope of the unperturbed dynamics and the attenuation function are dependent on the explanatory variable.

Needless to say, the estimation of *conditional ERD* will require more data than those of the *averaged*. Consequently to allow for a proper evaluation of the *conditional ERD* estimators, we sampled 1000 independent single trials from the stochastic process $\mathbb{Y}$, according to the particular settings of each data set. More precisely we generated an equal number of single trials per condition, yielding 500 trials at rest and 500 event-related trials, respectively. Thus the observed sample set comprises

$$\mathcal{D} = \left\{ \left(Y_{\cdot}^{k}, c^k, z^k \right) \right\}_{k=1}^{1000}. \tag{3.36}$$

Results

Before comparing the results of the *conditional ERD* estimators, we first study the estimations of *averaged ERD*. To this end, let us start with some preliminary theoretical considerations about the expected outcome of each method.

AVERAGED ERD Based on the construction of the event-related power attenuation, using a multiplicative dampening term $\alpha_t(s) \in [0,1]$, we define the true averaged ERD at time t as

$$\mathrm{ERD}_{\mathrm{true}}[t] := \int \alpha_t(s)\, dPs^{-1} - 1. \tag{3.37}$$

Using the single time instance $t_{\mathrm{REF}} = -\pi$ as reference the *conventional ERD* measure yields:

$$\mathrm{ERD}_{\mathrm{CONV}}[t] \;=\; \frac{\int \alpha_t(s) f_t(\theta,\beta)\, dP(\theta,s)^{-1}}{\int f_{t_{\mathrm{REF}}}(\theta,\beta)\, dP\theta^{-1}} - 1. \tag{3.38}$$

Instead of using the fixed reference condition in denominator at each time instance, the generalized measure contrasts the time courses of the perturbed and unperturbed envelopes:

$$\mathrm{ERD}_{\mathrm{GEN}}[t] \;=\; \frac{\int \alpha_t(s) f_t(\theta,\beta)\, dP(\theta,s)^{-1}}{\int f_t(\theta,\beta)\, dP\theta^{-1}} - 1. \tag{3.39}$$

Considering the particular settings of the different data sets, e.g., with respect to the independence of random variables, yields the following analytic solutions:

dataset	$\mathrm{ERD}_{\mathrm{true}}[t]$	$\mathrm{ERD}_{\mathrm{CONV}}[t]$	$\mathrm{ERD}_{\mathrm{GEN}}[t]$
I	$\int_0^1 \alpha_t(z)\, dz - 1$	$\left(\frac{9}{11} - \frac{2t}{11\pi}\right) \int_0^1 \alpha_t(z)\, dz - 1$	$\int_0^1 \alpha_t(z)\, dz - 1$
II	$\alpha_t(0) - 1$	$\alpha_t(0) - 1$	$\alpha_t(0) - 1$
III	$\int_0^1 \alpha_t(z)\, dz - 1$	$\frac{2}{3} \int_0^1 f_t(2\pi z, 0)\alpha_t(z)\, dz - 1$	$\frac{2}{3} \int_0^1 f_t(2\pi z, 0)\alpha_t(z)\, dz - 1$

Consequently for the first data set we expect the conventional estimator to underestimate the ERD curve, i.e., to estimate an ERD that is enhanced in magnitude. On the other hand the generalized estimator shall obtain an estimate close to the true ERD. In the case of the second data set we expect both methods to perform equally well

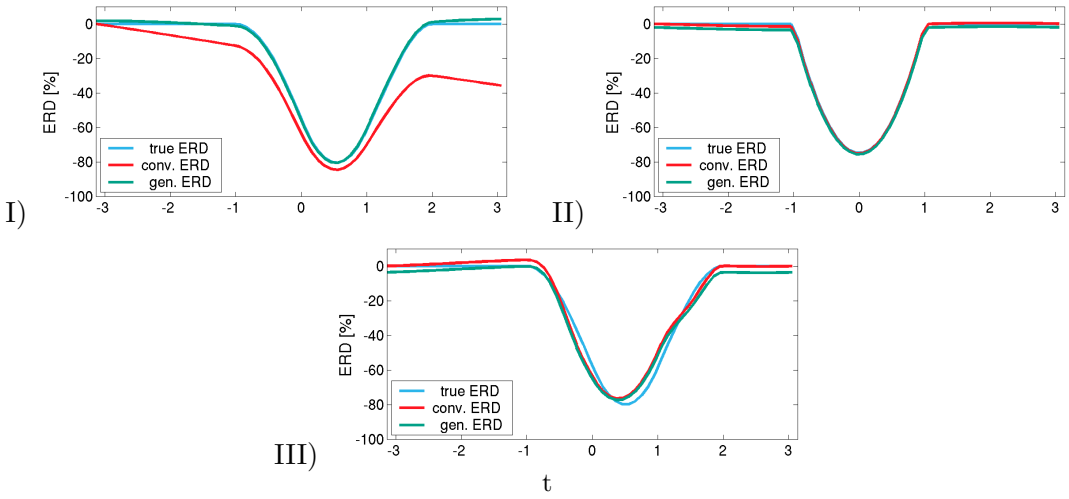


Figure 3.2: Each panel depicts the true averaged ERD (blue) and the estimates of the conventional (red) and the generalized method (green) for one of the three different toy examples I–III.

in estimating the averaged ERD. Finally, the analytic solutions for the third data set suggest a relatively small systematic error that is identical for both methods [1].

In order to verify these predictions, we apply the conventional and the novel generalized estimator, according to (3.7) and (3.19) to investigate the three data sets. The reference condition for the conventional estimator was set to the single time instance $\mathcal{T}_0 = \{-\pi\}$. The results, depicted in Fig. 3.2, empirically confirm the findings above.

CONDITIONAL ERD As for the *averaged ERD* we start with an analytic study of the different *conditional ERD* measures:

$$\mathrm{ERD}_{\mathrm{true}}[t\,|Z=z] = \int \alpha_t(s)\,dP(s\,|Z=z)^{-1} - 1, \tag{3.40}$$

$$\mathrm{ERD}_{\mathrm{CONV}}[t\,|Z=z] = \frac{\int \alpha_t(s) f_t(\theta,\beta)\,dP(s,\theta\,|Z=z)^{-1}}{\int f_{t_{\mathrm{REF}}}(\theta,\beta)\,dP(\theta\,|Z=z)^{-1}} - 1, \tag{3.41}$$

$$\mathrm{ERD}_{\mathrm{GEN}}[t\,|Z=z] = \frac{\int \alpha_t(s) f_t(\theta,\beta)\,dP(s,\theta\,|Z=z)^{-1}}{\int f_t(\theta,\beta)\,dP(\theta\,|Z=z)^{-1}} - 1. \tag{3.42}$$

Again using $t_{\mathrm{REF}} = -\pi$ and considering the particular settings of the different data sets yields the following analytic solutions:

[1] This systematic error is due to the dependence of the random variables θ and s. In the case of dependent variables it generally holds $\mathbb{E}[\alpha(s)f(\theta)] \neq \mathbb{E}[\alpha(s)]\,\mathbb{E}[f(\theta)]$. Consequently both ERD measures are doomed to a systematic false estimation of the averaged attenuation $\mathbb{E}[\alpha(s)]$.

dataset	$\mathrm{ERD}_{\mathrm{true}}[t \,\vert Z = z]$	$\mathrm{ERD}_{\mathrm{CONV}}[t \,\vert Z = z]$	$\mathrm{ERD}_{\mathrm{GEN}}[t \,\vert Z = z]$
I	$\alpha_t(z) - 1$	$\left(\frac{9}{11} - \frac{2t}{11\pi}\right) \alpha_t(z) - 1$	$\alpha_t(z) - 1$
II	$\alpha_t(0) - 1$	$\frac{f_t(z,0)}{f_{t_{\mathrm{REF}}}(z,0)} \alpha_t(0) - 1$	$\alpha_t(0) - 1$
III	$\alpha_t(z) - 1$	$\frac{f_t(z,0)}{f_{t_{\mathrm{REF}}}(z,0)} \alpha_t(z) - 1$	$\alpha_t(z) - 1$

So, based on those calculations, we expect the conventional estimator to fail for all three data sets, while the generalized estimator should be capable of revealing the underlying functional relationship between the explanatory variable and the ERD characteristic. Fig. 3.3 shows the true conditional ERD and contrasts the different results of the two competing methods. Comparing the empirical estimates it clearly reveals that the *generalized ERD* is capable of recovering the underlying relationship of the ERD characteristic on the explanatory variable Z, while the conventional estimator fails completely.

3.4.2 Median nerve stimulation data

In the following we will investigate the EEG recordings of a single subject, acquired during a median nerve stimulation (MNS) paradigm. Once again we would like to stress that it is not the purpose of the following investigation to thoroughly prove any neurophysiologically hypothesis about the dependency of the ERD characteristic. Instead we will point out the potential of the proposed *generalized ERD* framework and envision possible directions of future neurophysiological investigations.

The processing of MNS mainly causes responses in the somatosensory cortex, including modulations of the μ-rhythm. To this end, our analysis will primarily focus on the μ-rhythm ERD. In particular we will investigate its dependency on the magnitude of the local pre-stimulus μ-activity, on the magnitude of the occipital pre-stimulus α-activity and finally on the magnitude of the ERS response of the preceding stimulus.

Before we describe the details of the experimental design, let us start with the following considerations. The *generalized ERD* framework requires the observation of single trial data at rest and in state of excitation. In principle this requirement can be incorporated into any arbitrary experimental design, e.g., by using a simple randomized strategy such that with a certain probability a stimulus is either delivered or withheld. Alternatively, single trials at rest may also be acquired from data of a solely resting

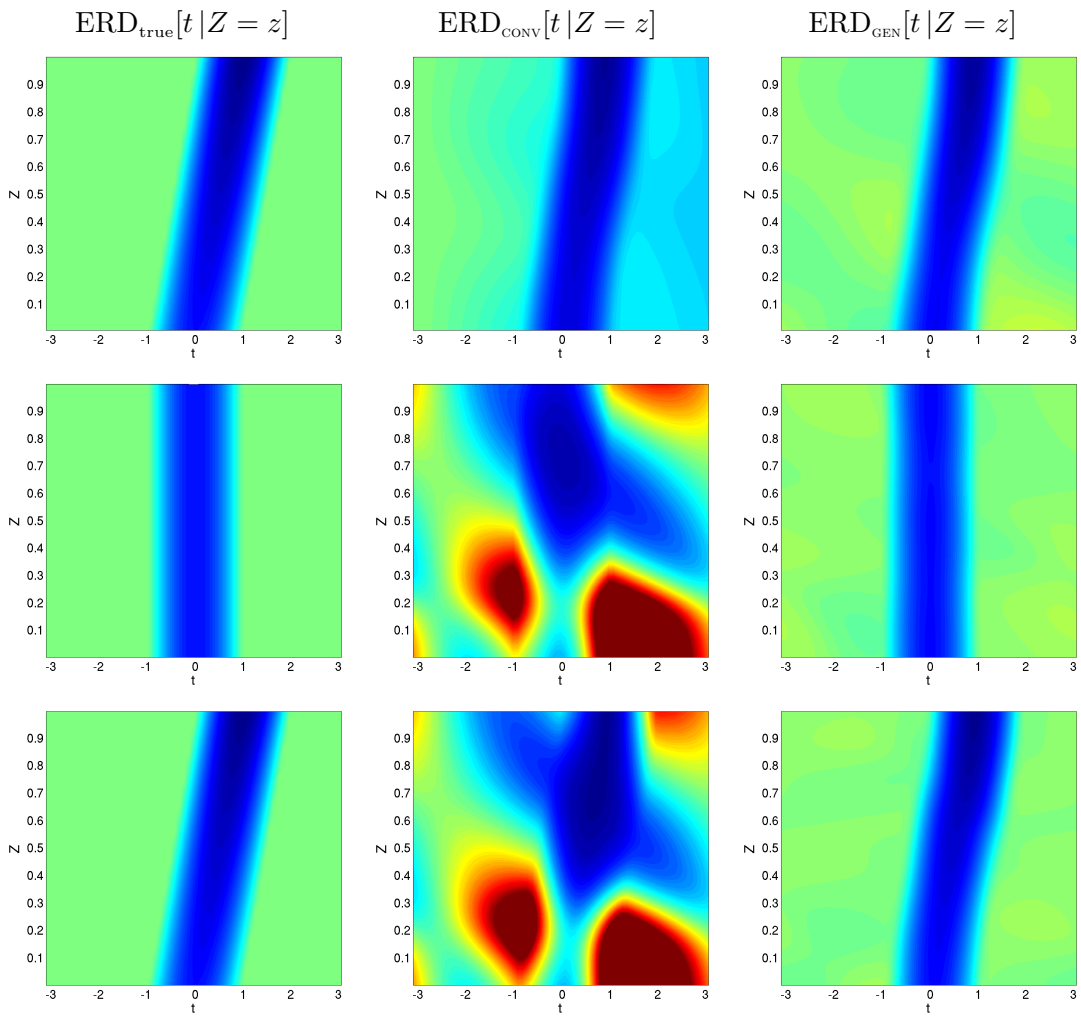


Figure 3.3: The figure contrasts the true conditional ERD (left column), the estimated conventional conditional ERD (central column) and the estimated generalized conditional ERD (right column). Each row corresponds to a particular artificial data set (I-III, top to bottom). The panel share an identical linear color coding scheme, where blue refers to -100% (ERD) and red indicates 100%, i.e., to an ERS.

paradigm. Generating virtual event triggers and defining the single trials accordingly, also provides single trial data for the requested reference condition. However, the latter procedure is not recommended for the following reason: Typically subjects are exposed to a sequence of repetitive identical stimuli, at a certain interstimulus interval (ISI). Thus, depending on the length of the ISI, the dynamics of cortical rhythmic activity under investigation might still be affected by the preceding stimulus. Thus the actual cortical activity at stimulus onset may differ from those during rest. Consequently gen-

erating virtual single trials from a pure resting paradigm or taking single trials of a randomized stimulation paradigm may give rise to differently estimated reference dynamics and thereby to differently estimated *conditional ERDs*. Nevertheless, there are also caveats concerning the randomized stimulation paradigm. For instance any randomized sequence of excitation and rest contains just by chance segments of consecutive trials at rest. The length of such periods has a direct influence on the ISI between the two neighboring excitations. Thus random sequences provide stimulations with an irregular ISI and consequently with different initial conditions. In order to circumvent this issue, we suggest composing each single trial as a pair of two stimuli. Where the

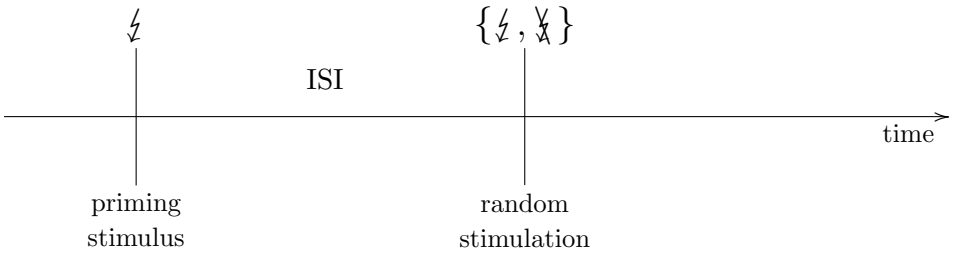


Figure 3.4: Schematic layout of a single trial. After a first priming stimulus a second stimulus is delivered randomly at a predefined ISI. The responses to the second stimulus are then used for the analysis of conditional ERD.

first excitation takes place surely and serves as an initial or priming stimulus, which is followed at a predefined ISI by a second, randomly delivered stimulus (Fig. 3.4 depicts the schematic of this procedure). The analysis of ERD is then confined to the analysis of the responses to the second stimulus.

Experimental design

The brain activity of a healthy subject was recorded using a 64-channel EEG system and at a sampling rate of 1000 Hz. During the experiment the subject was sitting relaxed in a comfortable chair and staring at a fixation cross. According to the single trial layout, depicted in Fig. 3.4, the right median nerve was electrically stimulated at the wrist. The ISI between the initial priming stimulus and the second randomly delivered stimulus was set to 2.5 seconds. The intertrial interval (the period between two consecutive initial stimuli) was set to approximately 5 seconds. The intensity of both stimuli was identically set to 10 mA, which was slightly below the motor threshold, i.e., the stimuli were not sufficient to evoke a thumb twitch. The probability of the second stimulus to

be delivered was $\frac{1}{2}$. Using a pseudo-random sequence, we recorded a sufficiently large number of single trials, which allows for the analysis of *conditional ERD*. In particular we acquired a total of 1200 single trials, i.e., 600 per condition. Since we are basically interested in the ERD response to the second, randomly delivered stimulus, we use that trigger signal as zero point on the time scale. Accordingly the initial stimulus is delivered at -2500 ms and positive time indices refer to the ERD/ERS period of the stimulus under investigation.

Data preprocessing

Generally MEG/EEG signals obtained at individual sensors are often composed as a (linear) superposition of several distinct signals. Consequently the amount of non-μ activity recorded at sensors over the somatosensory region is increased by these additional signals. As we have already seen in 1.3, the accuracy of the ERD measure highly depends on the ratio between background and μ-activity. Thus we use spatial projection techniques, in order to extract the signals of interest, while simultaneously suppressing interferences. Furthermore we use a wavelet transformation to achieve a high-resolution representation of the signals in the time-frequency domain.

Spatial Projection For our particular analysis of *conditional ERD* we need spatial filters for the left-hemispheric μ-rhythm and the occipital α-rhythm. To this end, we use the Common Spatial Pattern (CSP) algorithm on the bandpass filtered signal to project onto the signal originating from the contra-lateral somatosensory cortex, while Independent Component Analysis (ICA) is applied on the broadband signals for the extraction of occipital α-sources.

CSP is a technique known from statistical pattern recognition [29] and was suggested by [42] for the spatial analysis of EEG signals and by [84] to find spatial structures of ERD and ERS. The CSP analysis solves the task of finding a linear subspace, i.e., linear combination of channels, for which the variance of the signal is maximized for one condition while the variance of another condition is minimized. A more detailed mathematical introduction to the method of CSP will be found in chapter 4, where we present an extension of the original CSP algorithm. However, in order to apply this technique we define two virtual conditions based on the observed *averaged ERD* from an arbitrary sensor over the left somatosensory cortex. The first condition is defined as the

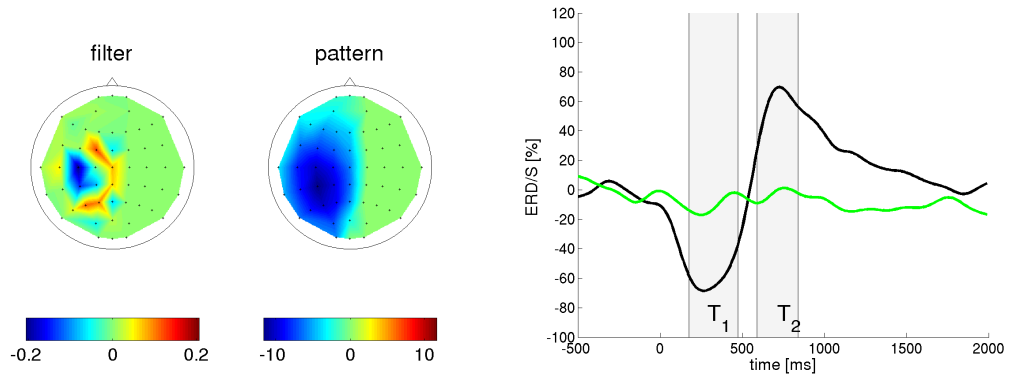


Figure 3.5: Using the virtually defined condition T_2 and T_1 (gray shaded regions in the right panel), the CSP method is applied. The panels at the left depict the estimated spatial filter and corresponding spatial pattern. The right panel shows the conventionally estimated averaged ERD response to the second stimulus separately for the stimulation condition (black) and the reference condition (green).

ERS period, while the opposite condition is set to the ERD period. Then applying the CSP algorithm to the band pass filtered signal, determines the optimal spatial filter such that the variance (the power) is maximized during the ERS period while it is minimized for the ERD phase. Intuitively, the obtained spatial filter reflects the best linear spatial projection on the modulated rhythmic μ-activity. Fig. 3.5 depicts the estimated spatial filter along with the corresponding common spatial pattern and the *averaged ERD* of the projected source. It also indicates the two virtual conditions.

In order to extract occipital α-activity, we apply an ICA algorithm to the broadband signals. For an introduction to ICA and its application to EEG data please refer to chapter 5, where we will present a method that incorporates prior knowledge into the ICA framework for an improved extraction of evoked responses. However, for the present purpose it is sufficient to know that ICA finds a common set of linear spatial filters such that the projected sources become statistically independent. For the particular application here we use the Temporal Decorrelation SEParation (TDSEP) algorithm [110], which exclusively uses second-order statistics in the form of temporally delayed covariance matrices. Fig. 3.6 depicts the estimated spatial filter along with the spatial pattern of an extracted occipital α-source.

TIME FREQUENCY REPRESENTATION. Elaborated analysis of the temporal evolution of evoked spectral perturbations requires a high-resolution representation of the data in

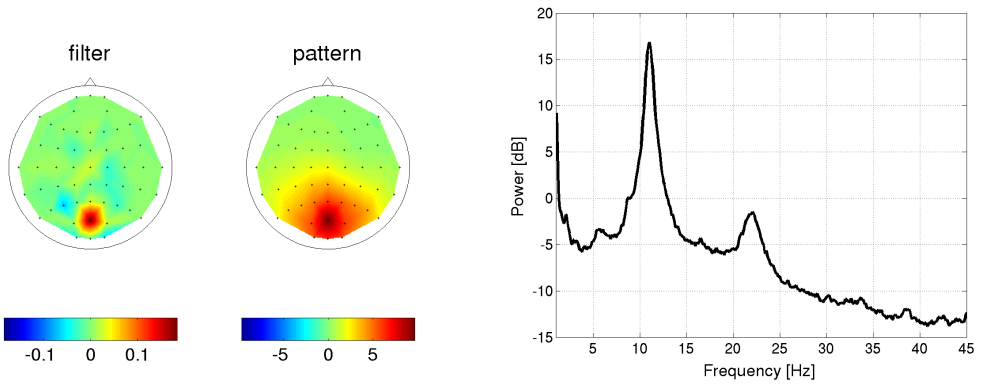


Figure 3.6: Left: The spatial filter and the corresponding spatial pattern of one extracted occipital α generator found by the TDSEP algorithm. The right panel shows the power spectrum of the projected source, with a distinct peak at 11 Hz.

the time-frequency domain. In classical ERD approaches [40, 77] the signals are first filtered in a narrow frequency band, squared, lowpass filtered and averaged over trials (see Fig. 1.3). Unfortunately this procedure has several drawbacks. Basically it yields a low resolution and low selectivity in the time-frequency domain. Various authors, e.g., [26, 72], therefore suggested the use of wavelet transformation, in order to achieve an appropriate representation of the dynamics. Following this suggestion, we use Morlet wavelets, which are known to achieve the best ratio between the resolution in the time and in the frequency domain. Moreover, Morlet wavelets are complex filters, which give rise to analytic signals and thereby enable access to the instantaneous phase and the instantaneous amplitude of rhythmic activity. For an easy introduction to wavelet decomposition, with a particular emphasis on Morlet wavelets we refer to [96]. In order to bandpass filter the EEG signals, we first determined the individual spectral peak in the 8–13 Hz domain at sensors covering the somatosensory and the occipital region. Secondly, we applied a Morlet wavelet centered at the spectral peak at congruently 11 Hz.

Referring to the obtained spatial and spectral filters intuitivly as w_{CSP}, w_{ICA} and $b_{11\text{Hz}}$, respectively, we define the stochastic processes of contra-lateral μ-rhythm and occipital α-power as:

$$\mathbb{Y} := \left| w_{\text{CSP}}^{\top} \cdot \mathbb{X} * b_{11\text{Hz}} \right|^2 \qquad \text{and} \qquad \mathbb{O} := \left| w_{\text{ICA}}^{\top} \cdot \mathbb{X} * b_{11\text{Hz}} \right|^2, \tag{3.43}$$

where $\mathbb{X}$ represents the stochastic process of single trial EEG, and $\mathcal{T}$ =[-2500,2000]ms

denotes the common index set [2]. With respect to the experimental design, the period [-2500,0]ms contains the responses to the initial stimuli, while the interval [0,2000]ms is used for the analysis of *conditional ERD*. By defining the explanatory random variable Z differently, we are going to analyze the dependency of the contra-lateral μ-rhythm ERD on the magnitude of

1.) the pre-stimulus contra-lateral μ activity itself, i.e.,

$$Z_\mu := \log \sum_{t \in \mathcal{T}_{\text{PRE}}} Y_t - \log |\mathcal{T}_{\text{PRE}}|, \qquad \mathcal{T}_{\text{PRE}} := [\text{-500,-100}]\text{ms}, \tag{3.44}$$

2.) the pre-stimulus occipital α-activity, i.e.,

$$Z_\alpha := \log \sum_{t \in \mathcal{T}_{\text{PRE}}} O_t - \log |\mathcal{T}_{\text{PRE}}|, \qquad \mathcal{T}_{\text{PRE}} := [\text{-500,-100}]\text{ms}, \tag{3.45}$$

3.) the ERS response to the preceding priming stimulus, i.e.,

$$Z_{\text{ERS}} := \log \sum_{t \in \mathcal{T}_{\text{ERS}}} Y_t - \log |\mathcal{T}_{\text{ERS}}|, \qquad \mathcal{T}_{\text{ERS}} := [\text{-1950,-1700}]\text{ms}, . \tag{3.46}$$

Note that the right boundary of the pre-stimulus intervals are commonly set to -100 ms in order to prevent stimulus related artifacts entering the calculations. Furthermore $\mathcal{T}_{\text{ERS}}$, as defined in (3.46), corresponds to the interval [550,800]ms relative to the priming stimulus and hence to its ERS period.

Remark 3.4.1. The use of the logarithm in the definition of the explanatory variables (3.44)–(3.46) is motivated by the fact that the distribution of bandpower is typically similar to a log-normal distribution. Thus taking the logarithm of the averaged bandpower yields a distribution similar to that of a Gaussian. In particular, using a constant kernel width, as for the Nadaraya-Watson-estimator, is more appropriate in case of a homogeneous distribution. However, since the logarithm is a monotonic transformation it preserves the locality property of the data and thus does not affect any monotonic relationship between the explanatory variable and the ERD characteristic.

[2] More precisely, as we are going to apply the intertrial variance method, we are using $\mathbb{Y} = \left| w_{\text{CSP}}^\top \cdot \mathbb{X} * b_{11\text{Hz}} - \mathbb{E}\left[w_{\text{CSP}}^\top \cdot \mathbb{X} * b_{11\text{Hz}} \right] \right|^2$.

Results

For each of the three differently defined explanatory variables (3.44)–(3.46), we compare the *averaged* and the *conditional* μ-rhythm ERD found by the *generalized* and the *conventional* framework. As previously stated, it is not within the scope of this thesis to thoroughly prove any neurophysiologically hypothesis about the dependency of the ERD characteristic. Moreover, there exists no genuine truth about such relationship. Consequently we will compare both methods in a descriptive rather than in a qualitative manner. In order to apply the conventional estimator we set the reference interval to $\mathcal{T}_{\text{REF}} := [\text{-}500,\text{-}100\,]\text{ms}$, which is the same interval, as it is used to determine the pre-stimulus activities.

AVERAGED ERD. As before we start with the comparison of the *averaged ERD* estimates, which are contrasted in Fig. 3.7. Here the results of the generalized and the conventional estimator are slightly different. In particular the *generalized averaged ERD* exhibit a more pronounced, longer lasting ERS than the classical estimate. In order to trace back that difference to its origin, Fig. 3.8 depicts the averaged event-related power envelopes along with the corresponding references of each method, i.e., a constant reference in the case of the conventional estimator (left panel) and the reference dynamics for the generalized measure (right panel). Obviously, the generalized reference exhibits a distinct negative linear trend. This negative trend clearly indicates that the cortical μ-generator is still at an excited state and the dynamics at rest must therefore be considered non-stationary. In this context, please recall the results of the first artificially generated data set (Fig. 3.2 panel I) and the related discussion in remark 3.3.1.

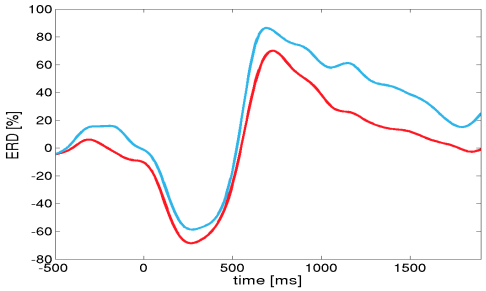


Figure 3.7: The resulting averaged ERD obtained by the two methods. The red curve indicates the conventional averaged ERD, while blue corresponds to the generalized averaged ERD.

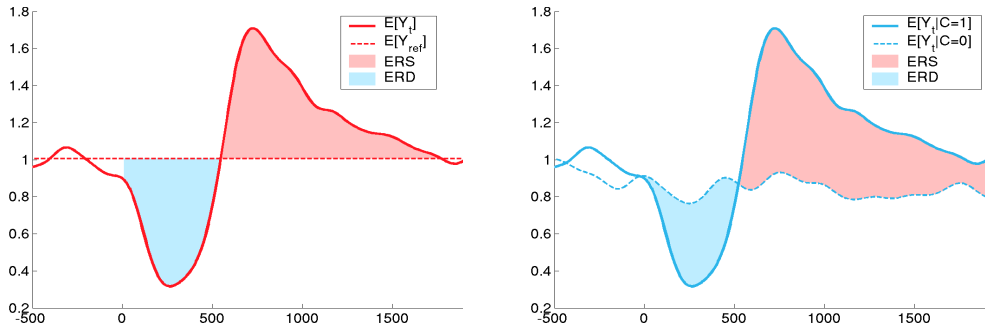


Figure 3.8: Each panel contrasts the estimated averaged event-related dynamics (solid lines) with the corresponding references (dashed lines). The left panel refers to the conventional ERD measure, using a constant reference that is estimated in from the interval [-500,-100]ms. The right panel corresponds to the generalized model. The bluish and reddish areas of each panel indicate the periods of ERD and ERS, respectively.

Conditional ERD. For the estimation of the *conditional ERD* we set the bandwidth of the Gaussian kernel to the default value (cf. (2.22)). The resulting *conditional ERD* estimates of both methods are contrasted in Fig. 3.9. At a first glance the ERD/ERS complex exhibits less variability across the domain of the explanatory variable in the case of the generalized estimator, i.e., the ERD and ERS are similar in magnitude and latency at different levels of Z. Moreover, in accordance with the findings for the *averaged ERDs*, the generalized framework consistently estimates a longer lasting ERS response than the conventional estimator.

Going into detail, the most conspicuous discrepancy is visible for the μ-rhythm ERD conditioned on its own pre-stimulus activity. In order to trace these differences back to their origin, Fig. 3.10 contrasts the conditional expectation of the μ-rhythm band power at rest versus the stimulation condition at three particularly chosen values of pre-stimulus activity. The reference dynamics ($C = 0$) at lowest pre-stimulus activity exhibits a distinct positive linear trend (cf. left panel of Fig. 3.10). Consequently taking the pre-stimulus band power as constant reference, as it is done in the *conventional ERD* framework, results in a spurious interpretation of ERS even for the reference dynamics itself. The opposite behavior is observed in the case of increased pre-stimulus activity (right panel of Fig. 3.10). Here the reference dynamics exhibits a prominent negative linear trend. Since this pre-stimulus level exceeds even the hyper-synchronization level after actual stimulation at approximately 800 ms, the *conventional ERD* measure gives rise to a spurious interpretation of solely ERD, without any indication of an ERS.

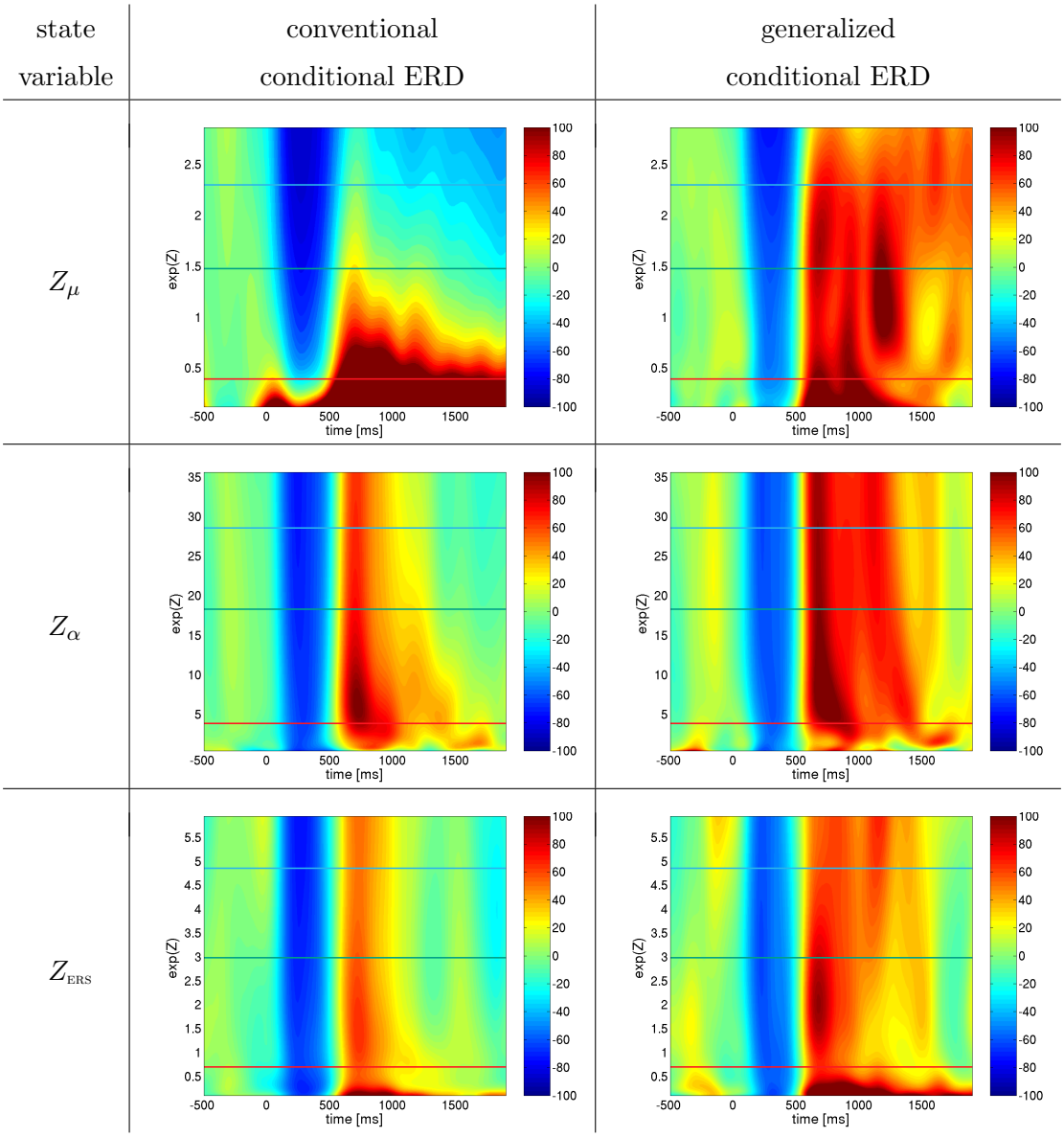


Figure 3.9: Each panel depicts the estimated conditional μ-rhythm ERD using a color-coding scheme. The individual rows contrast the results of the two methods for a particular explanatory variable, i.e., magnitude of the pre-stimulus μ-activity, pre-stimulus occipital α-activity, and preceding ERS response (top to bottom). The additionally highlighted horizontal lines of each panel indicate selected values of the state variable, used for a further investigations. Note that the vertical axes indicate the exponential of the explanatory variables and thus correspond to band power values.

Fig. 3.11 contrasts the corresponding *conditional ERDs* obtained from both methods. Here the conventional method exposes the forecasted misinterpretation, while the ERD estimates of the generalized method indicates a slight, directly correlated dependence

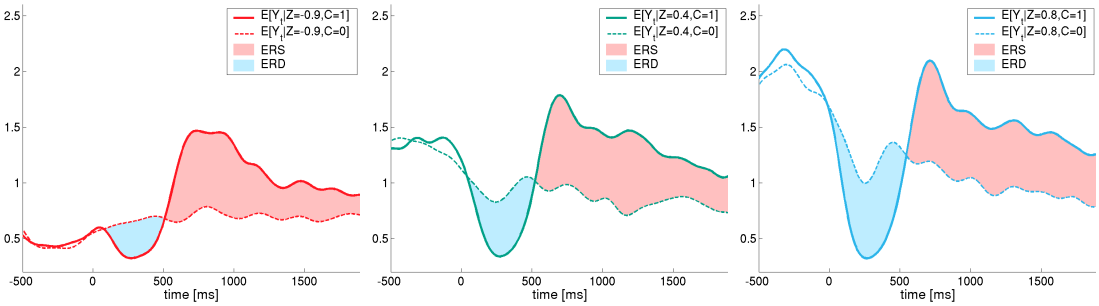


Figure 3.10: The different panels display the estimated conditional dynamics of the μ-power given its pre-stimulus activity at three different levels. The particularly chosen state value of each panel corresponds in color and magnitude to those highlighted in Fig. 3.9. The dashed lines corresponds to the reference dynamics, as they are used in the generalized framework, while the solid lines indicate the dynamics after an actual stimulation. The colored areas indicate the periods of ERD (blue) and ERS (red) identified by the generalized framework.

of the ERD magnitude on the initial μ-power. However, the ERS characteristic across the three different pre-stimulus activity levels appears rather similar.

The discrepancies revealed by the ERD analysis conditioned on the occipital α-power are less apparent. As before we start to study the conditional expectations of the μ-rhythm power envelopes. Here Fig. 3.12 contrasts the event-related and the reference dynamics of the μ-rhythm band power at three particularly chosen values of the explanatory variable. Comparing the three panels the only detectable difference here comprises the observation, that a low α-power (left panel) coincides with a lower μ pre-stimulus activity. Thus the static reference condition of the conventional estimator is lower compared to the two other α-activity levels. The difference in the baseline

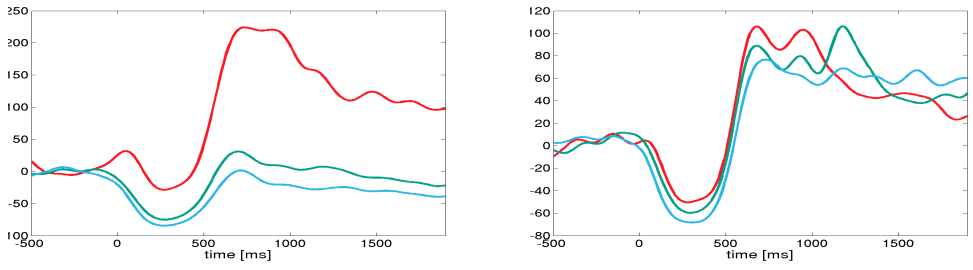


Figure 3.11: The different panels display the estimated conditional ERD of the μ-rhythm given its pre-stimulus activity at three different levels. The left panel refers to the conventional measure, while the right panel depicts the solution of the generalized estimator. Again, the particularly chosen state values correspond in color and magnitude to those highlighted in Fig. 3.9.

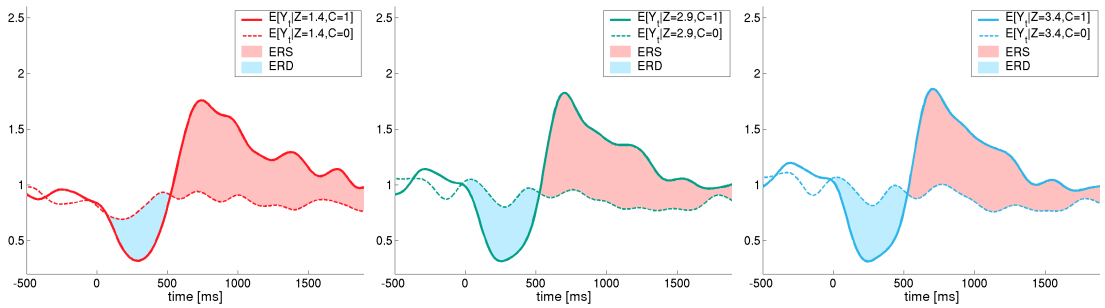


Figure 3.12: The different panels display the estimated conditional dynamics of the μ-power given the occipital pre-stimulus α-activity at three different levels. The particularly chosen state values correspond in color and magnitude to those highlighted in Fig. 3.9. The dashed lines correspond to the reference dynamics, while the solid lines indicate the dynamics after actual stimulation. The colored areas denote the periods where the event-related dynamics falls below the reference dynamics (bluish area) and where it is exceeded (reddish area).

causes the upward shift of the corresponding *conventional conditional ERD* estimate (cf. left panel of Fig. 3.13). However, according to the *generalized conditional ERD* analysis the occipital α-power seems to have no effect on the characteristics of the ERD (cf. right panel Fig. 3.13).

Next, we are going to compare both methods with respect to the estimated ERD, conditioned on the magnitude of the ERS response to the preceding initial stimulus. Again we begin with an examination of the conditional expectations of the μ-rhythm power envelopes. Here Fig. 3.14 contrasts the event-related and the reference dynamics of the μ-rhythm band power at three particularly chosen values of the appropriate explanatory variable. The most prominent finding here corresponds to the observation

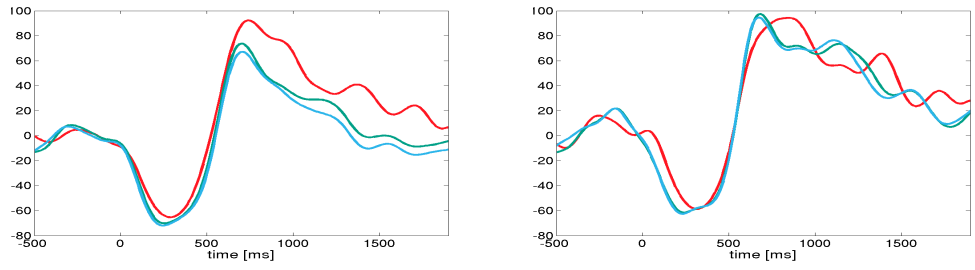


Figure 3.13: The panels display the estimated conditional ERD of the μ-rhythm given the occipital pre-stimulus α-activity at three different levels. The left panel refers to the conventional measure, while the right panel depicts the solution of the generalized estimator. Again, the particularly chosen value of each panel corresponds in color and magnitude to those highlighted in Fig. 3.9.

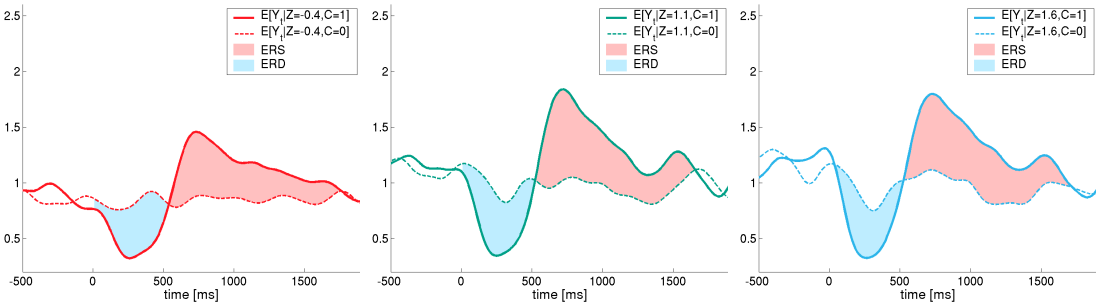


Figure 3.14: The different panels display the estimated conditional dynamics of the μ-power given the magnitude of the ERS response to the preceding initial stimulus at three different levels. The particularly chosen state values correspond in color and magnitude to those highlighted in Fig. 3.9. The dashed lines correspond to the reference dynamics, while the solid lines indicate the dynamics after actual stimulation. The colored areas denote the periods where the event-related dynamics falls below the reference dynamics (bluish area) and where it is exceeded (reddish area).

that in case of a weak ERS response (left panel) the reference dynamics appears to be stationary and at a lower power level than those of the two other selected values. Even more interesting is the fact that the left panel suggests that a low-power ERS response to the initial stimulus is correlated with a low-power ERS response to the subsequent stimulus. However, comparing the resulting *conditional ERD* estimates in Fig. 3.15 reveals no significant difference between the three states. A comparison of both methods again discloses a longer lasting ERS response in case of the generalized estimator, but exposes no further differences.

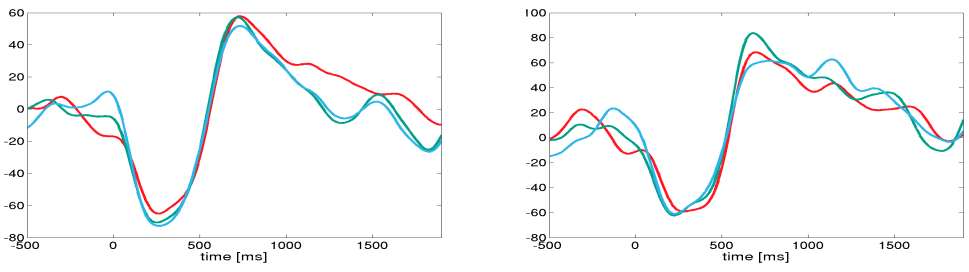


Figure 3.15: The panels display the estimated conditional ERD of the μ-rhythm given the magnitude of the ERS response to the preceding initial stimulus at three different levels. The left panel refers to the conventional measure, while the right panel depicts the solution of the generalized estimator. Again, the particularly chosen value of each panel corresponds in color and magnitude to those highlighted in Fig. 3.9.

Conclusion

We presented the novel framework of *conditional ERD* that allows for the analysis of dependencies of the ERD characteristic on external or internal explanatory variables. To this end, we first extended the *conventional averaged ERD* measure towards the analysis of *conditional ERD* by means of conditional expectations. Here it turned out that this naive extension is not capable of reliably disclosing functional relationships between the ERD characteristic and arbitrary explanatory variables. We then identified the fixed reference condition as the main cause for this inability. Consequently we generalized the *conventional ERD* framework with respect to the reference. In particular we substituted the static reference value by a reference dynamics. Based on this generalized framework, we derived novel measures for the quantification of *averaged* and *conditional ERD*, by defining ERD/ERS as the relative deviation of the event-related dynamics from this reference dynamics. In this context we also discussed how the acquisition of such a reference dynamics can be incorporated into an experimental design.

In order to provide a proof of concept, we compared the individual capabilities of the conventional and the generalized framework. To this end we applied the corresponding *averaged* and *conditional ERD* estimators to simulated and real ERD data. Here the analysis of artificially generated data in a controlled scenario revealed the limitations of the *conventional ERD* framework. Unlike the conventional method, which failed completely, the novel generalized measures performed well at retrieving the underlying functional relationship of the ERD characteristic on an explanatory variable from the surrogate data. Finally we envisioned the potential of the proposed novel framework for neurophysiological investigations. Here the analysis of ERD data from a median nerve stimulation paradigm demonstrated the capabilities of the proposed framework. In particular we applied the novel estimator of *generalized conditional ERD* in order to analyze the contra-lateral μ-rhythm ERD and its dependency on:

- the magnitude of its own pre-stimulus activity,
- the magnitude of the occipital α-activity,
- and the magnitude of the preceding ERS response.

De facto, these three examples represent just the tip of the iceberg. Possible future investigations shall include comprehensive studies on the impact of various external

factors such as the inter-stimulus interval or the simultaneous processing of multiple stimuli, but also the influence of internal factors such as the activation of adjacent cortical areas, the phase and magnitude of various brain rhythms, et cetera. Moreover, applications such as brain-computer interfacing can also benefit from this *generalized conditional ERD* framework. Here advanced classifiers which consider state dependent behavior of ERD are expected to provide an improved accuracy of discrimination between different mental states. However, there are definitely more exciting issues that are now ready for being approached on the basis of this novel ERD concept.

Chapter 4

Spatio-spectral filters

Electroencephalogram (EEG) data in general is very noisy, non-stationary and contaminated with artifacts that can deteriorate its analysis, such as the online detection of mental states in the context of a brain computer interface (BCI). Thus the major goal of preprocessing or feature extraction is to improve the signal-to-noise ratio of the data significantly, for instance by spatial projections or spectral filtering. Here Common Spatial Pattern (CSP) [29, 42, 68, 84] is an efficient method of obtaining optimally discriminative spatial projections for bandpass filtered data. However, the frequency band has to be specified appropriately in advance and is kept fixed during the subsequent optimization. In the following chapter we extend the CSP algorithm such that it additionally optimizes the spectral filter. To this end, we suggest expanding CSP to the state space by means of the method of time delay embedding. As we will show, this allows for individually tuned frequency filters for each channel and hence yields an improved and more robust feature extraction. The advantages of the proposed method over the original CSP method are verified in the context of single trial classification of recordings from a set of BCI experiments of imagined limb movements. Here we show the efficiency of the proposed method in terms of an improved information transfer rate (bits per trial).

The chapter is organized as follows: We will start elaborating on the mathematical background of CSP. Subsequently we will introduce two methods of obtaining spatio-spectral filters. Here we will first sketch the Common Sparse Spectral Spatial Pattern (CSSSP) [23], before the Common Spatio-Spectral Pattern (CSSP) [46] algorithm will be presented in detail. Finally the performances of the three different methods will be

compared on the basis of a comprehensive set of BCI-experiments in section 4.3.

4.1 Common spatial pattern

Since the EEG data is noisy, consists of a superposition of simultaneously active brain sources, is typically distorted by artifacts and often exposes non-stationary behavior, automated feature extraction becomes challenging. Moreover, outliers and artifacts can distort the analysis dramatically, e.g., yielding classifiers with bad generalization ability [65], i.e., the performance on previously unseen data, can become arbitrarily worse. So it is important to strive for robust machine learning and signal processing methods that are invariant against such distortions (e.g. [63, 43, 66, 90]). The common spatial pattern (CSP) algorithm [29] is highly successful in calculating spatial filters for detecting ERD/ERS effects [42] and for ERD-based BCIs, see [84] and has been extended to multi-class problems in [21]. Given two class distributions in a high-dimensional space, the CSP algorithm finds directions (spatial filters) that maximize the variance for one class, while simultaneously minimizing the variance for the opposite class. In order to see how this fits into the context of ERD and ERS, let us consider the issue of discriminating between left hand and right hand imaginary movements. It is known that motor imagery causes a contra-lateral ERD. Thus a spatial filter, that focuses on signals originating from the left motor area, yields a signal that is characterized by a present motor rhythm during the imagination of right hand movements (left hand is in idle state), and by an attenuated motor rhythm in case of a left hand movement. But this exactly corresponds to the optimization criterion of the CSP algorithm: maximizing variance for the class of right hand trials, while simultaneously minimizing the variance for left hand trials. Moreover, the CSP algorithm calculates the dual filter that will focus on the area of the right hand (and it will even calculate several filters for both optimizations by considering orthogonal subspaces).

To be more precise, let $\mathbb{X} = (\boldsymbol{X}_t)_{t\in\mathcal{T}}$ denote the stochastic process of the (potentially bandpass filtered) multi-channel EEG over the temporal index set $\mathcal{T} \subset \mathbb{Z}$, representing the individual sampling points of each single trial. Note, $\boldsymbol{X}_t$ represents a multivariate random variable. Thus, whenever we would like to refer to a single channel $c \in 1, \ldots, C$ explicitly, we use $X_{c,t}$ for notational convenience. Furthermore let $\mathcal{D} = \left\{\left(\boldsymbol{X}_{\cdot}^{k}, Y^{k}\right)\right\}_{k=1}^{K}$ denotes the set of K labelled realizations of $\mathbb{X}$, where

$Y^k \in \{1;2\}$ represents the class label of the k^{th} trial. Using the short hand notation of $\mathcal{Y}_1 := \{k : Y^k = 1\}$ and $\mathcal{Y}_2$ for the set of class '2' trials, respectively, the estimator of the two class conditional covariance matrices reads as,

$$\Sigma_1 = \frac{1}{|\mathcal{Y}_1||\mathcal{T}|} \sum_{k\in\mathcal{Y}_1} \sum_{t\in\mathcal{T}} \boldsymbol{X}_t^k \boldsymbol{X}_t^{k\top} \quad \text{and} \quad \Sigma_2 = \frac{1}{|\mathcal{Y}_2||\mathcal{T}|} \sum_{k\in\mathcal{Y}_2} \sum_{t\in\mathcal{T}} \boldsymbol{X}_t^k \boldsymbol{X}_t^{k\top}. \tag{4.1}$$

Given these two class conditional covariance matrices, CSP is formulated as the following optimization problem:

$$\max_w w^\top \Sigma_1 w, \qquad \text{s.t.} \;\; w^\top (\Sigma_1 + \Sigma_2) w = 1. \tag{4.2}$$

This optimization problem can be solved by formulating its dual and calculating a matrix W and diagonal matrix D with elements in $[0, 1]$ such that

$$W\Sigma_1 W^\top = D \qquad \text{and} \qquad W\Sigma_2 W^\top = I - D. \tag{4.3}$$

Here the row corresponding to the largest diagonal element of D yields the solution w for the optimization problem (4.2). However a solution for (4.3) can be obtained by: First *whiten* the matrix $\Sigma_1 + \Sigma_2$, i.e., determine a matrix P such that

$$P(\Sigma_1 + \Sigma_2)P^\top = I. \tag{4.4}$$

This decomposition can always be found due to positive definiteness of $\Sigma_1 + \Sigma_2$. Secondly define $S_1 := P\Sigma_1 P^\top$ and $S_2 := P\Sigma_2 P^\top$ respectively and calculate an orthogonal matrix R and a diagonal matrix D by spectral theory such that

$${S_1}^\top = RDR^\top. \tag{4.5}$$

From $S_1 + S_2 = I$ directly follows ${S_2}^\top = R(I - D)R^\top$. Note that the projection given by the p^{th} row of matrix R has a relative variance of d_p (p^{th} element of D) for trials of class 1 and relative variance $1 - d_p$ for trials of class 2. If d_p is close to 1 the spatial filter given by the p^{th} row of R maximizes variance for trials of class 1. Conversely, since $1 - d_p$ is close to 0 it minimizes the variance for trials of class 2. The final solution of (4.3) is given by

$$W := R^\top P. \tag{4.6}$$

Based on artificial data, Fig. 4.1 illustrates the two main steps involved in solving the optimization problem (4.3), i.e., the *whitening* and the final rotation.

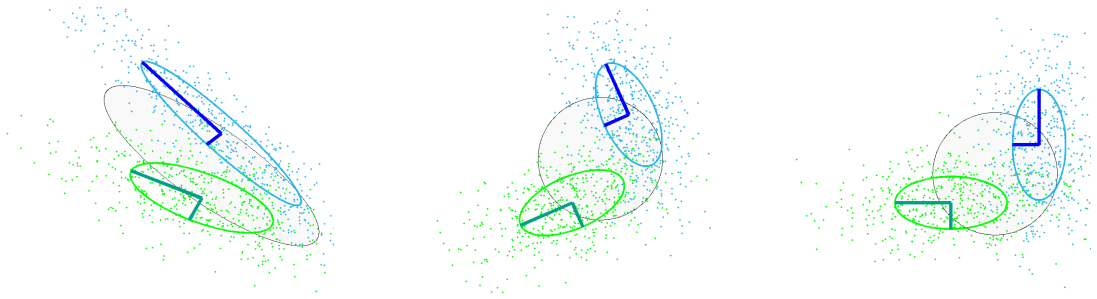


Figure 4.1: The two essential steps of the CSP optimization. The blue and green ellipsoids refer to the two class conditional covariance matrices, along with the indicated principal axes. The gray ellipsoid represents the overall covariance matrix. The left panel depicts the raw data, the central panel shows the corresponding distributions after the *whitening* step, cf. (4.4). The final rotation (cf. (4.5)) in the right panel aligns the principle axes with the coordinate axes, such that the variance along the horizontal direction is maximal for the *green* class, while it is minimal for the *blue* class and vice versa along the vertical direction.

Applying the decomposition matrix W, the EEG recordings $\boldsymbol{X}^k$ are projected to discriminative source signals

$$\boldsymbol{Z}^k = W\boldsymbol{X}^k. \tag{4.7}$$

The interpretation of W is two-fold, the rows of W are the spatial filters, whereas the columns of W^{-1} can be seen as the *common spatial patterns*, i.e., the time-invariant coupling coefficients of each source with the different scalp electrodes.

4.1.1 Single trial features

Finally, the source signals used for the classification are obtained by decomposing the EEG according to (4.7). Typically one would retain only a small number $2m$ of projections that contain most of the discriminative information between the two classes. These projections are given by the rows of W that correspond to the m largest and m smallest eigenvalues $\{d_{p_i}\}_{i=1}^{2m}$. The final feature vector of each single trial is represented by the vector of the log-transformed signal variances (across time) of each projected single trial $\{Z^k_{p_i,\cdot}\}_{i=1}^{2m}$, i.e.,

$$f_i^k = \log\left(\mathrm{Var}\left[Z^k_{p_i,\cdot}\right]\right),\ k = 1,\ldots,K,\ i = 1,\ldots,2m. \tag{4.8}$$

The $2m$-dimensional feature vectors $\{\boldsymbol{f}^k\}_{k=1}^K$ serve as the final input to the corresponding classifier.

4.2 Spatio-spectral methods

The performance of the CSP algorithm directly depends on the particular choice of the passband which is used for spectrally filtering the EEG signals in advance. Although [68] found evidence that a broad band filter is the best general choice, a subject-specific choice mostly improves the result. Inspired by the Common Spatio-Spectral Pattern algorithm [46] the Common Sparse Spectral Spatial Pattern algorithm was introduced in [23]. The basic idea behind both approaches is to derive a discriminative spatio-temporal filter in the spirit of CSP. To this end, the CSSP and CSSSP algorithms alleviate the problem of manually fine-tuning the frequency band by simultaneously optimizing a temporal and a spatial filter, i.e., the methods not only return optimized spatial filters, but also Finite Impulse Response (FIR) filters, which act along with the spatial filter in order to improve the discriminability of different brain states.

4.2.1 Sparse spectral spatial pattern

In order to tackle this issue, the CSSSP method considers the following optimization problem

$$\begin{aligned} \max_{\boldsymbol{b},b(1)=1} \max_{\boldsymbol{w}} \quad & \boldsymbol{w}^\top \left(\sum_{\tau=0,\ldots,T-1} \left(\sum_{j=1,\ldots,T-\tau} b(j)b(j+\tau) \right) \Sigma_1^\tau \right) \boldsymbol{w} - C/T \left\| \boldsymbol{b} \right\|_1 , \\ s.t. \quad & \boldsymbol{w}^\top \left(\sum_{\tau=0,\ldots,T-1} \left(\sum_{j=1,\ldots,T-\tau} b(j)b(j+\tau) \right) (\Sigma_1^\tau + \Sigma_2^\tau) \right) \boldsymbol{w} = 1. \end{aligned} \tag{4.9}$$

Here $\boldsymbol{w}$ represents the spatial filter, while $\boldsymbol{b}$ is an FIR filter of length T. Note that the inner maximization step (for a fixed $\boldsymbol{b}$) is similar to the standard CSP-method, but uses temporally delayed auto-covariance matrices Σ_1^τ and Σ_2^τ instead. In order to avoid overfitting, the complexity of the sparse spectral filter is limited, i.e., a penalty term for its complexity is added. The particular choice of the L_1-norm of the FIR filter ensures the sparsity of the obtained solution. Moreover, the degree of sparsity is controlled by the non-negative regularization factor C. However a solution to (4.9) can be found using optimization techniques such as gradient descent or line-search methods. Hence the CSSSP algorithm finally finds global spatio-spectral filters. Note that the obtained sparse spectral filter $\boldsymbol{b}$ is global, i.e., is applied to all channels.

In the next section we introduce the CSSP algorithm, which solves the problem of simultaneously optimizing spectral and spatial filters by extending the standard CSP method to the state space. As we will explain in detail, solving the CSP problem in state space yields spectral filters that are individually tuned for each sensor. However, this additional degree of freedom comes at the expense of a stronger limitation to the feasible set of FIR filters. Nevertheless, our proposed model is even simpler than the CSSSP algorithm and can be solved quite efficiently, and as we will show in 4.3 performs competitively.

4.2.2 Common spatio-spectral pattern

Let us start with a brief introduction to the concept of state space, followed by the extension of the CSP algorithms to the state space. Subsequently we discuss its consequences for the optimization problem and propose an intuitive representation of the results.

Introduction to state space

Very few natural systems have actually been found to be low-dimensional deterministic in the strict sense of theory. Nevertheless the concept of deterministic low-dimensional chaos has proven to be fruitful in the understanding of many complex phenomena. Also a number of attempts have been made to analyze various aspects of EEG time series in the context of nonlinear deterministic dynamic systems.

Determinism in a strict mathematical sense means that there exists an autonomous dynamic system, defined typically by a first order differential equation $\dot{\mathbf{y}} = f(\mathbf{y})$ in a state space $\Gamma \subseteq R^d$, which is assumed to be observed through a single measurable quantity $s = h(\mathbf{y})$. The system thus possesses d natural variables, but the measurement is usually a nonlinear projection onto a scalar value. In order to recover the deterministic properties of such a system, we have to reconstruct an equivalent of the state space Γ. The time delay embedding method is one way to do so. From a sequence of scalar observations $s_1, s_2, \ldots, s_N$ overlapping vectors $\mathbf{s}_n = (s_n, s_{n-\tau}, \ldots, s_{n-(m-1)\tau})$ are formed, with τ as the delay time. Then according to Takens Theorem [93] under certain conditions, such that for mathematically perfect, noise free observations s_n and m sufficiently large, there exists a one-to-one relation between $\mathbf{s}_n$ and the unobserved

vectors $\mathbf{y}_n$. Thus the attractor of any non-linear dynamics can be reconstructed in the state space using an appropriate delay coordinate function.

CSP in state space

Since it is not our target to reconstruct the entire dynamics of the EEG-signal, but rather to extract *robust* (invariant) features, we extend (4.7) just by one delayed coordinate, i.e.,

$$\boldsymbol{Z}_t^k = W^{(0)}\boldsymbol{X}_t^k + W^{(\tau)}\delta^\tau \boldsymbol{X}_t^k = \left(W^{(0)}, W^{(\tau)}\right)\begin{pmatrix} \boldsymbol{X}_t^k \\ \delta^\tau \boldsymbol{X}_t^k \end{pmatrix}. \tag{4.10}$$

Here, for notational convenience, δ^τ denotes the delay operator, i.e,

$$\delta^\tau \boldsymbol{X}_t = \boldsymbol{X}_{t-\tau}. \tag{4.11}$$

Once again, the optimization criterion is to find projections $W^{(0)}$ and $W^{(\tau)}$ such that signal variance of different Z_p discriminates two given classes best, i.e., maximizing the variance for one class while minimizing it for the opposite class. In order to use the identical mathematical concepts, as introduced in section 4.1, we just append the delayed vectors $\delta^\tau \boldsymbol{X}^k$ as additional channels to $\boldsymbol{X}^k$, i.e.,

$$\hat{\boldsymbol{X}}^k = \begin{pmatrix} \boldsymbol{X}^k \\ \delta^\tau \boldsymbol{X}^k \end{pmatrix}. \tag{4.12}$$

Then the optimization criterion can be formulated equivalent to (4.2) and (4.3) respectively, using the class conditional covariance matrices $\hat{\Sigma}_l, l \in \{1, 2\}$ obtained from $\hat{\boldsymbol{X}}^k$. Following the steps of (4.4)–(4.6) yields a solution $\hat{W}$ of the modified optimization problem. This filter matrix $\hat{W}$ can be decomposed into two submatrices $\hat{W}^{(0)}$ and $\hat{W}^{(\tau)}$, such that $\hat{W}^{(0)}$ applies to $\boldsymbol{X}^k$ and $\hat{W}^{(\tau)}$ applies to the delayed channels $\delta^\tau \boldsymbol{X}^k$, i.e., $\hat{W}\hat{\boldsymbol{X}}^k = \left(\hat{W}^{(0)}\,\hat{W}^{(\tau)}\right)\hat{\boldsymbol{X}}^k$.

Spatio-spectral filters

Based on this, we will now explore the implications of this decomposition further. Especially we will derive an interpretation into a spatial and a spectral filter. To this end, let $\boldsymbol{w}$ denote the p^{th} row of the decomposition matrix $\hat{W}$, then the projected signal

$Z_p^k = \boldsymbol{w}\,\hat{\boldsymbol{X}}^k$ can be written as

$$\begin{aligned} Z_{p,t}^k &= \boldsymbol{w}^{(0)}\boldsymbol{X}_t^k + \boldsymbol{w}^{(\tau)}\delta^\tau \boldsymbol{X}_t^k && (4.13)\\ &= \sum_{c=1}^{C} w_c^{(0)} X_{c,t}^k + w_c^{(\tau)}\delta^\tau X_{c,t}^k && (4.14)\\ &= \sum_{c=1}^{C} \gamma_c \left(\frac{w_c^{(0)}}{\gamma_c} X_{c,t}^k + \frac{w_c^{(\tau)}}{\gamma_c} X_{c,t-\tau}^k \right),\ t \in \mathcal{T}, && (4.15) \end{aligned}$$

where $(\gamma_c)_{c=1,\ldots,C}$ is a pure spatial filter and $(\frac{w_c^{(0)}}{\gamma_c}, \overbrace{0,\ldots,0}^{\tau-1}, \frac{w_c^{(\tau)}}{\gamma_c})$ defines a FIR filter separately for each channel c. This decomposition into a spatial and a FIR filter is not unique, but there exists a natural partitioning, i.e.,

$$\gamma_c := \frac{\sqrt{{w_c^{(0)}}^2 + {w_c^{(\tau)}}^2}}{\widetilde{\text{sign}}\left(w_c^{(0)}\right)}, \tag{4.16}$$

where

$$\widetilde{\text{sign}}(w) = \begin{cases} -1, & w < 0 \\ +1, & w \geq 0 \end{cases}. \tag{4.17}$$

Using the signed norm in the definition of the spatial filters $\boldsymbol{\gamma}$, maps the non-zero coefficients of the corresponding FIR filter onto one half of the two dimensional unit-sphere. Consequently we can easily parameterize the FIR filters by the angle $\phi_c^{(\tau)}$, defined as

$$\phi_c^{(\tau)} := \text{atan}\left(\frac{w_c^{(0)}}{w_c^{(\tau)}}\right) \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \tag{4.18}$$

Fig. 4.2 illustrate these FIR filters by means of the resulting magnitude responses curves for various values of τ and at different angles $\phi^{(\tau)}$. Note that for each electrode there is an individual FIR filter. Consequently each recording site can be tuned individually to focus on spectral components of interest and thus allows for an adaptation to the spectral peaks.

Similar to the original CSP algorithm, the spatial origin of each source signal can be examined using the inverse of the matrix $\Gamma = (\gamma)_{p,c}$. Here each column of the inverse corresponds to the coupling strength of one particular source with the scalp electrodes.

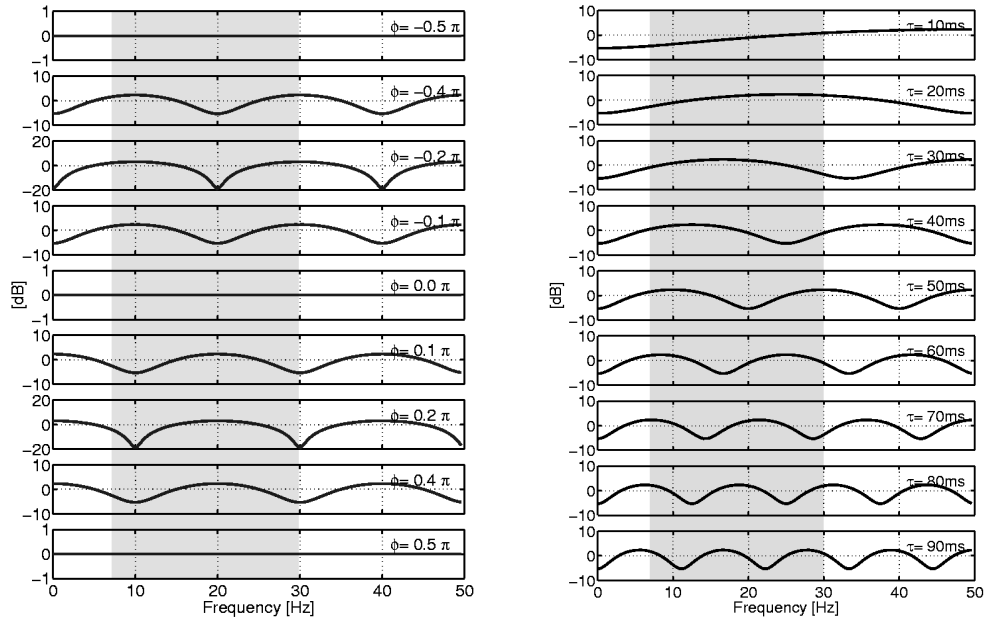


Figure 4.2: Magnitude responses of the FIR filters at different values of ϕ and τ. The shaded region denotes the frequency range (7–30 Hz) of interest. Left panel: Varying $\phi^{(\tau)}$ at single fixed delay τ =50 ms. Increasing $\phi^{(\tau)}$ from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ keeps the position of the extreme values in the frequency domain fixed, but turns maxima into minima. Since a minimum corresponds to a suppression of the spectral information at this frequency, the FIR filter for $\phi^{(\tau)} = -\frac{1}{5}\pi$ focuses mainly on $\{10, 30, 50\}$ Hz. Conversely, the filter given by $\phi^{(\tau)} = \frac{1}{5}\pi$ has the contrary effect, i.e., it cancels these frequencies. Right panel: Varying τ, while keeping ϕ fixed at $-\frac{2}{5}\pi$. Increasing τ changes the position and increases the number of the extreme values. Thus the various FIR filters focus on different sub-bands in the frequency spectrum.

4.2.3 Online applicability

A major concern in online applications is to implement an algorithm which is as efficient as possible.

CSP

In case of a CSP based classification the involved operations such as bandpass filtering and spatial projections define the bottleneck for the speed of the processing. Especially the spectral filtering of each EEG channel (64-128) is quite time consuming and dramatically effects the overall processing speed. But fortunately the preprocessing steps of bandpass filtering, that is typically realized by convolution with a FIR filter b, i.e., $(b \star \boldsymbol{X})$, and the spatial filtering $(W\boldsymbol{X})$ are strictly linear. Consequently both

operations are commutative, i.e.,

$$W\left(b \star \boldsymbol{X}\right) = b \star \left(W\boldsymbol{X}\right). \tag{4.19}$$

This implies that we are allowed to first apply the spatial projections and have to filter only the few selected source signals in the desired bandpass, such that the CSP algorithm becomes easily applicable in real time.

Spatio-spectral filter

In case of CSSP and due to the non-linear embedding operation δ^τ an arbitrary order of the spatial and spectral filtering will not yield identical results. But we can easily work around this and are allowed to exchange the operations in the following manner:

$$\hat{W}\begin{pmatrix} b \star \boldsymbol{X} \\ \delta^\tau(b \star \boldsymbol{X}) \end{pmatrix} = \begin{pmatrix} I_C \\ I_C \end{pmatrix}^\top \begin{pmatrix} b \star \hat{W}^{(0)}\boldsymbol{X} \\ \delta^\tau(b \star (\hat{W}^{(\tau)}\boldsymbol{X})) \end{pmatrix}. \tag{4.20}$$

Where I_C denotes the C-dimensional identity-matrices and basically adds up corresponding source signals. This enables us to first filter the signals spatially and to select appropriate sources. Afterwards we can apply the general bandpass filter, e.g., 7–30 Hz and apply the delay operator, before we finally sum up the corresponding source signals. Consequently the computational demands are just doubled compared to the original CSP. Hence the proposed extended CSSSP model is applicable in real time as well.

4.3 Classification of imaginary movements

In this section we compare the CSP, CSSP and CSSSP algorithm on data from 60 EEG experiments of imaginary limb movements performed by 22 different subjects in terms of the achieved classification performances. The investigated mental tasks were imagined movements of the left hand (l), the right hand (r), and one foot (f). In this study we investigate all resulting two-class classification problems, i.e., all possible combination of two classes (l-r,l-f and r-f). Note that in a few experiments only two mental tasks were used.

4.3.1 Experimental design

During the experiment the subjects were sitting in a comfortable chair with arms lying relaxed on the armrests. Each experiment started with a calibration session in which the subjects performed mental motor imagery tasks in response to a visual cue. In such a way (labeled) examples of single trial brain activity during the different mental tasks were obtained. In the calibration session every 4.5–6 seconds one of 3 different visual stimuli indicated for 3–3.5 seconds which mental task the subject should accomplish during that period. The brain activity was recorded from the scalp at a sampling rate of 100 Hz with multi-channel EEG amplifiers using either 32, 64 or 128 channels. In addition to the EEG channels, we recorded the electromyogram (EMG) from both forearms and the right leg as well as horizontal and vertical electrooculogram (EOG) from the eyes (cf. Fig. 1.1 for a schematic of the montage). The EMG and EOG channels were exclusively used to ensure that the subjects performed no real limb or eye movements correlated with the mental tasks that could directly (artifacts) or indirectly (afferent signals from muscles and joint receptors) be reflected in the EEG channels and thus be detected by the classifier, which should solely operate on the brain activity. For each involved mental task we obtained between 120 and 200 single trials of recorded brain activity. In the original experiment, these recorded single trials were then used to train a classifier which was in a second session applied online to produce a feedback signal for (unlabeled) continuous brain activity.

Here in this off-line study we will solely use data from the first (calibration) session to evaluate the performance of the three algorithms. This is basically due to the issue, that if feedback is provided to a subject, he/she will adapt to this particular feedback (output of the classifier). Hence the data acquired during a feedback session is biased towards the specifically used classifier that produced the feedback. Since the feedback in the original experiment was based on CSP features, we consequently decided to exclude the data of the feedback session for the evaluation process.

4.3.2 Classification and validation

After choosing all channels except the EOG, the EMG and a few outermost channels of the cap that are supposed to have non-stationary signal quality, we applied a causal bandpass filter from 7–30 Hz to the data, which encompasses both the μ- and the β-

rhythm. The single trials were extracted from the temporal frame 500–3500 ms after the presentation of the visual stimulus. During this period discriminative brain patterns are present in most of the subjects. On these preprocessed single trials we perform the feature extraction by the CSP, CSSSP and the proposed CSSP method separately. For each method we project the data to the three most informative directions of each class, yielding a 6-dimensional subspace. For these six dimensions we calculate the logarithms of the variances as feature vectors, according to (4.8). Finally we apply a linear discriminant analysis (LDA) to the feature vectors, to find separating hyperplanes. Note that in contrast to [8, 10] we omitted the regularization of the linear classifier itself (RLDA [28]), since the dimensionality of the features is rather small compared the number of training examples. Furthermore, the introduced delay τ (in case of CSSP) and the regularization constant C (in CSSSP) appear as underlying hyper-parameters in the overall optimization scheme. Consequently both have to be subjected to a model selection procedure in order to find the optimal values for each specific classification task.

Remark 4.3.1. Using $\tau = 0$ in the set of feasible hyper-parameters incorporates the original CSP algorithm into the model selection procedure. Note that CSSSP becomes asymptotically equivalent to CSP if the regularization parameter C tends to infinity [23].

In order to compare the results of the three methods (CSP vs. CSSSP vs. CSSP) we split the data set in two. On the (chronological) first half we performed the training of the classifier, i.e., the feature extraction, model selection and the LDA. The performance of the estimated models were then evaluated on the second half of the data, to which the algorithms have had no access before. In the following we refer to these halves of the data set as "training data" and "test data".

To select the best CSSP and CSSSP model for each binary classification problem, i.e., find the optimal τ and C respectively, we estimate the performance of these algorithms by means of a 2×5-fold cross validation (CV) on the corresponding training data. Especially we run a model selection procedure over $\tau = 0, \ldots, 15$ (corresponding to $0, 10, \ldots, 150$ ms) in case of CSSP and over $C \in \{0, 0.01, 0.1, 0.2, 0.5, 1, 2.5\}$ for the CSSSP method respectively. In this 2×5-fold CV scheme the training data is split randomly into 5 disjoint subsets of nearly equal size. Now the feature extraction and

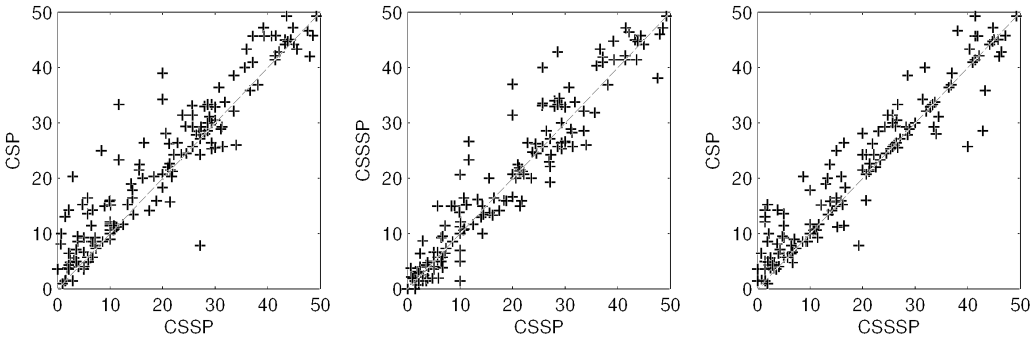


Figure 4.3: Each panel compares the obtained test errors of two algorithms over all datasets. Crosses above the diagonal implies that the algorithm associated with the x-axis outperforms the other and vice versa. Left panel: the proposed CSSP clearly outperforms the CSP. Right panel: CSSSP is superior to CSP. The central panel compares the two variants of CSP, i.e., CSSP (x-axis) vs. CSSSP (y-axis). However, there is no precise predominance of one algorithm (see also Fig. 4.4).

the classifier training are performed on 4 subsets and are applied to the excluded subset. This is repeated for all 5 subsets for 2 different splittings, such that we finally get 10 estimates of the classification error values. The mean of these errors for each hyper-parameter (τ and C resp.) were then used to select the corresponding hyper-parameter of each method.

Finally, after the model selection the corresponding CSSP, CSSSP and the CSP based models for each classification problem were trained on the entire training data and afterwards applied to the corresponding test data (the previously unseen second half of the calibration data) of each experiment.

4.3.3 Results

The resulting test errors for all datasets are compared in Fig. 4.3. The superiority of both CSP extensions to the standard CSP algorithm is clearly observable. However, in Fig. 4.4 the results of the individual methods are further summarized in terms of boxplots stating the median-value, the minimum and maximum, the 25% and 75% percentiles of the test error. The median classification rate for CSP is 23.3%, for CSSSP 20.7% and for CSSP 21.0%. Comparing the two CSP variants directly does not reveal any predominance of one particular algorithm. Thus the CSSP result is competitive with CSSSP, even though CSSP solves a much easier optimization problem. However, both methods optimize over different sets of feasible filters. Here CSSP allows for individually

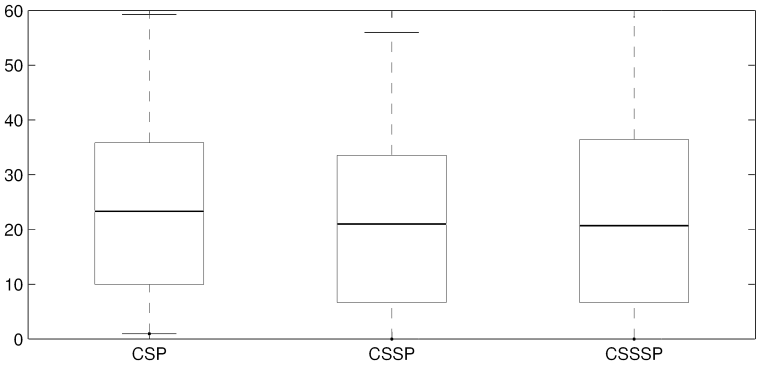


Figure 4.4: The boxplots summarizing the individual results for the three different methods in terms of the median-value, the minimum and maximum and 25% and 75% percentiles of the test error.

tuned but very simple FIR filters for each channel. On the contrary CSSSP restricts to a single, but more complex FIR filter simultaneously applied to all channels.

For further illustrations of the properties of CSSP, we study one specific dataset in more detail. In particular we focus on one particular classification task of imaginary *foot* and *right hand* movement. For this selected dataset we oppose the spatial filters found by the CSSP method to those of the CSP-method and discuss the impact of the additional spectral filters.

According to the model selection procedure, described in section 4.3.2, the model with the lowest CV-error on the training data ($\tau = 70\,\text{ms}$) has been chosen for the final application. The spatial and spectral filters of the selected model for each class are visualized in Fig. 4.5 and Fig. 4.6 respectively. These figures also oppose the filters found by CSP method. The first two spectral filters for class *foot* supply insight into how additional spectral information is exploited. Here the corresponding spatial filters found by the CSSP method are almost identical and focus on the central region, while the spectral filters have opposite signs in this area. This indicates that information from the same spatial location, but different frequencies is used for the discrimination. Moreover, a closer examination of the FIR filter corresponding to $\tau = 70\,\text{ms}$ in Fig. 4.2-right reveals that in the relevant frequency range (7-30 Hz) the maxima and minima are roughly at 14,21,28 Hz. Remember that for opposite signs of $\Phi^{(\tau)}$ the maxima and the minima are exchanged. Combining these facts, the first spectral filter focuses basically on the β band, whereas the second spectral filter has its focus close to the upper α band (11–13 Hz). So instead of having a spatial projection onto a broad band (7–30 Hz) signal

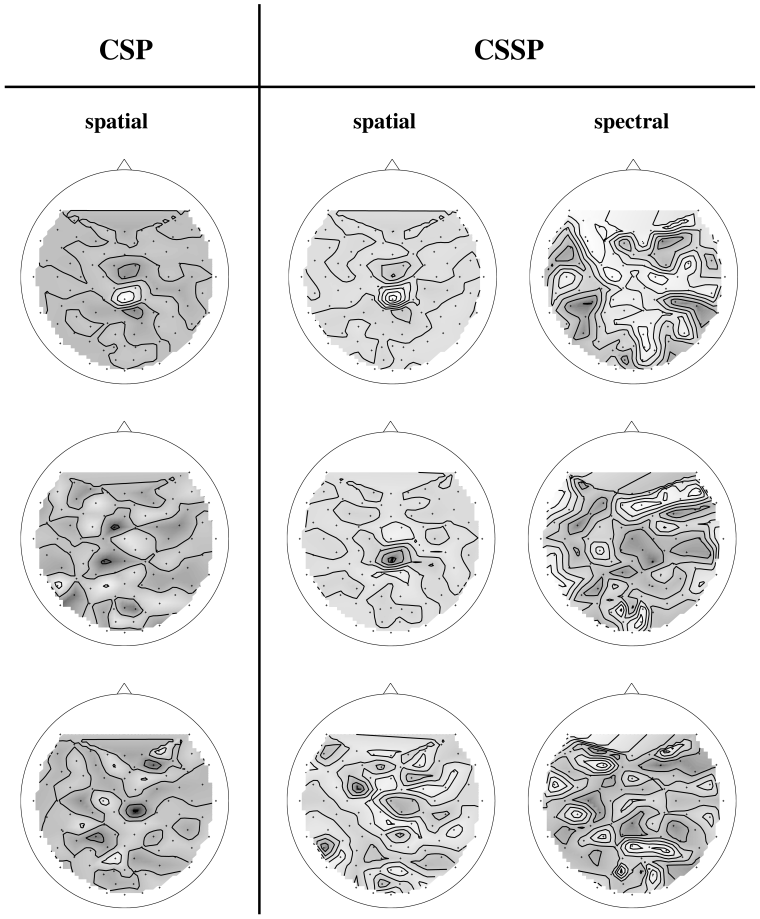


Figure 4.5: The scalp maps of the three spatial and spectral filters for the class *foot* in descending order of the eigenvalues for both the CSP and the CSSP method. The spectral filters are gray-scale-coded in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$. The first spatial filters are almost identical for CSP and CSSP, but already those for the second largest eigenvalue diverges. Here the spatial filter found by CSP exhibit no clear structure, while the corresponding CSSP filter resembles the first spatial filter. The main difference in the projection occurs only in the spectral filter, where these filters have opposite signs in the central region, indicating that different spectral information is exploited from the same location.

as the solution given by the standard CSP, CSSP splits the information by projecting onto two signals of the same local origin, but stemming from different sub-bands, such that each projection fulfills the optimization criterion of maximizing the variance for one class, while having minimal variance for the other class.

In this sense the CSSP algorithm is not only able to automatically adapt to the spectral EEG characteristics of a subject, but also to treat different spectral information, originating from closely adjoint (or identical) focal areas independently. As a result this

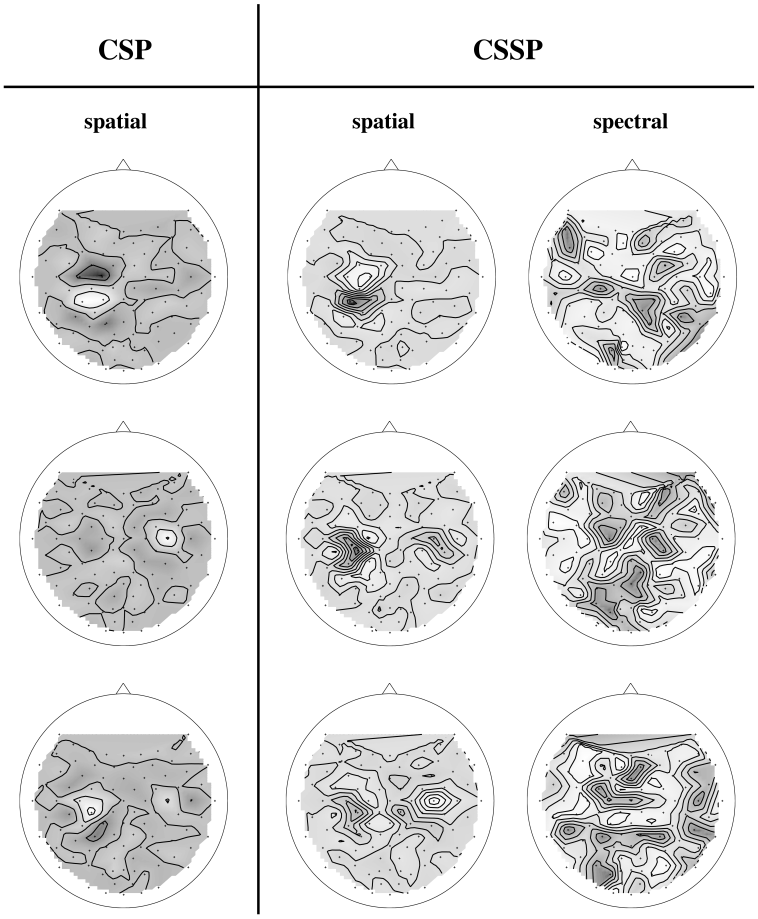


Figure 4.6: The scalp maps correspond to the spatial and spectral filters of the three leading eigenvalues for class *right hand* of the CSP and the CSSP method in descending order. The first spatial filters are almost identical for CSP and CSSP (except for the sign). For the second largest eigenvalue the CSSP filters work at the same location, whereas the CSP filter exhibit no clear focal point at the area of the left motor cortex.

yields an improved spatio-spectral resolution of the discriminative signals and hence improves the robustness and accuracy of the final classification.

Summary

In this chapter we used the method of delay embedding in order to extend the CSP-algorithm to the state space. The advantages of the proposed method were proved by its application to the classification of imaginary limb movements on a broad set of BCI experiments. We found that the CSSP algorithm, introduced here, outperforms

the standard CSP algorithm in terms of classification accuracy as well as generalization ability. Furthermore we compared the obtained results to those of the CSSSP method that also tackles the problem of simultaneously optimizing spatial and spectral filters. Here we reliably showed that CSSP performs competitively, although the underlying optimization procedure is much simpler and more efficient to solve.

It is worth mentioning, that in principle it is possible to further extend the suggested model by incorporating more than just one temporal delay. But this will come at the expense of a quadratically increasing number of parameters for the estimation of the covariance matrices, while the number of single trials for training remain the same. Hence, consistent with our observations, this approach will tend to over-fit the training data, i.e., the simultaneous diagonalization of the class conditional covariance matrices finds directions that explain the training data best, but might have poor generalization ability. The overfitting will primarily be present in the spectral domain, here asymptotically (with the number of delayed coordinates) the solution will converge to FIR filters that are responsive to only a single frequency, i.e., the coefficients of the FIR will resemble a sine or cosine function.

Recently, another spatio spectral method has been published [95]. In this work the authors carry out an extensive comparison of CSP, CSSP, CSSSP and the proposed method, finding all extensions of CSP performing competitive, but superior to the original CSP method.

Chapter 5

Improving the signal-to-noise ratio of ERPs

Recordings of encephalogram data typically consist of a (linear) superposition of several distinct signals. Hence the application of Blind Source Separation (BSS) techniques, such as Independent Component Analysis (ICA) is a quite common approach in order to disentangle the true underlying source signals. However, incorporating physiological prior knowledge into these decomposition techniques can help in extracting physiologically meaningful sources.

In this chapter we develop a method that uses prior information about the phase-locking property of event-related potential (ERP) in a regularization framework, which shifts the focus of the BSS methods towards the extraction of single-trial ERP. The proposed method is specifically tailored to trade off between single-trial decomposition and the separation of the averaged responses. In particular we propose an additional preprocessing of the data, i.e., we introduce a linear, temporal transformation, that increases the signal-to-noise ratio (SNR) in the subspace spanned by the event-related phase-locked components, prior to the application of ICA. We show that the assumed linear mixture model of the data is invariant under the proposed transformation. Hence, the estimated spatial filters can be directly used for the decomposition of the single trial EEG.

The chapter is organized as follow: After a brief review of several approaches for improving the SNR of single ERPs, we give an introduction into ICA and its application

to encephalographic data. In section 5.2 the linear transformation is described, along with the proofs of its basic properties. In the experimental section we illustrate the benefit of the proposed approach in terms of an increased SNR of extracted evoked components. Here the method is first evaluated in a controlled scenario using artificially generated data, before we finally present results of recovering somatosensory evoked potentials from multichannel EEG-recordings.

5.1 Introduction

The analysis of *single trial* data is an important research issue because variable behavior could potentially be traced back to variable brain states. Single trial analysis, however, suffers from the superposition of task-relevant signals by task-unrelated brain activities, resulting in a low SNR of the observed single trial responses. Here in the context of the SNR we refer to the ERPs as the *signals* and to all non-phase-locked neural activity as well as to non-neural artifacts as interfering *noise*. Accordingly, the major goal of data processing prior to the analysis of single trial ERPs is to enhance their SNR significantly, in other words isolating the phase-locked ERP *signal* from the interfering *noise*.

To this end, the analysis of ERP is mostly focussed on averaged responses to repeated identical stimuli. This procedure takes advantage of the fact that the phase locked ERPs persist under averaging over trials, whereas components of arbitrary phase, such as non-neural artifacts and ongoing "background" activity, cancel out. Consequently, averaging across trials increases the SNR for phase-locked ERPs, but has the drawback of masking single-trial variability of the task-related responses, e.g., in amplitude or latency. A more advanced averaging technique, called periodic stacking [91], aims to overcome this problem by simultaneously extracting averaged and differential responses. Since the method implicitly relies on trial-averaging the analysis of possible interactions between the single trial responses and the ongoing activity remains a challenge.

Other techniques suggested to improve the SNR for single trial analysis are based on temporal or spatial filtering. Commonly used are bandpass, notch or Laplace filters as well as principle component analysis (PCA) or more sophisticated techniques such as wavelet denoising [83], or BSS techniques.

5.1.1 Independent component analysis

Due to the fact that the electric fields of different bioelectric current sources superimpose linearly, the measured EEG can be modeled as a linear combination of component vectors

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t), \tag{5.1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)^\top$, $\boldsymbol{s} = (s_1, \ldots, s_M)$ and $A \in \mathbb{R}^{N \times M}$. Moreover, in case of independent component analysis it is further assumed that the observed signals $\boldsymbol{x}(t)$ are a linear mixture of $M \leq N$ mutually independent sources $\boldsymbol{s}(t)$, i.e., their joint probability density function factorizes.

Under these assumptions ICA decomposes the observed data $\boldsymbol{x}(t)$ into independent components $\boldsymbol{y}(t)$ by estimating the inverse decomposition matrix, such that $\boldsymbol{y}(t) = W\boldsymbol{x}(t)$. However, this recovers the original sources $\boldsymbol{s}(t)$ except for scaling and permutation. In general, as both the mixing process and the sources are unknown, this technique belongs to the so-called blind source separation methods [18].

FastICA

Most of the research conducted in this field uses higher-order statistics for the estimation of the independent components [18]. For instance Hyvärinen and Oja [34] maximized the kurtosis of the output signals. They developed a general fix-point iteration algorithm termed FastICA, that optimizes a contrast function measuring the distance of the source probability distributions from a Gaussian distribution [18, 34]. By whitening the data in the preprocessing step first, the complexity of the problem reduces, i.e., afterwards the estimation of an orthogonal matrix W remains [14]. For the estimation of the orthogonal matrix from whitened data $\boldsymbol{x}$ the iterative update of FastICA takes in matrix notation the form

$$W^+ = W - \Gamma^{-1} \left(\mathbb{E}\left[g(\boldsymbol{y})\boldsymbol{y}^\top \right] - \operatorname{diag}(\beta_i) \right) W, \tag{5.2}$$

here $\boldsymbol{y} = W\boldsymbol{x}$, $\beta_i = \mathbb{E}[y_i g(y_i)]$ and $\Gamma = \operatorname{diag}(\mathbb{E}[g'(y_i)] - \beta_i)$, where $g(\boldsymbol{y})$ is a non-linear contrast function. Here $g(y_i) = y_i^3$ is commonly used and corresponds to an optimization with respect to the kurtosis. Alternatively people apply hyperbolic tangent as contrast function.

TDSEP

The Temporal Decorrelation SEParation (TDSEP) algorithm [110], which is equivalent to Second-Order Blind Identification (SOBI) [5], relies on distinct spectral/temporal characteristics of the sources and exclusively uses second-order statistics in the form of temporally delayed covariance matrices $R_{\tau,\boldsymbol{x}} := = \mathbb{E}\left[(\delta^\tau \boldsymbol{x})\boldsymbol{x}^\top\right]$. Here δ^τ denotes the temporal delay operator, that shifts the signal $\boldsymbol{x}(t)$ by τ time instances, i.e., $\delta^\tau \boldsymbol{x}(t) := \boldsymbol{x}(t-\tau)$. TDSEP estimates the unknown mixing matrix A by simultaneous diagonalization of a set of correlation matrices $\{R_{\tau,\boldsymbol{x}} | \tau = \tau_1, \ldots, \tau_K\}$. Here TDSEP explicitly exploits the property that the cross-correlation between independent signals are zero. In general, the issue of simultaneous diagonalization of more than 2 matrices can be solved only approximatively, and there exist several optimization schemes solving this problem [15, 107, 109].

5.1.2 Application to EEG data

The application of BSS to neurophysiological signals, especially the decomposition of ERPs in human scalp EEG, is a challenging task because of the multitude of active brain sources contrasting with the relative paucity of sensors. Furthermore, non-stationarity is a general issue for EEG data analysis and can strongly affect the solution of BSS. However, the practical use of ICA for decomposing brain signals was first introduced in [52, 98].

Averaged data

Commonly the separation of EEG into neurophysiological sources is approached by decomposing averaged data. This procedure takes advantage of an increased SNR along the spatial directions of the phase-locked brain responses; however, the analysis of their single-trial latency or amplitude variability is hampered. It is important to note that spatial projections, that are estimated on averaged data, are not suitable to study the underlying single trials. This is because such filters are "blind" for the interfering single-trial noise that has canceled out under averaging. Consequently, applying these filters to single trial data yield a suboptimal decomposition of the data. Moreover, as trial-averaging ideally cancels out non-phase-locked sources leaving (a few) phase-locked event-related sources, the intrinsic dimensionality of the data is reduced and so

overfitting becomes an issue, when applying ICA. This is usually counterbalanced by projecting onto a lower dimensional subspace prior to the application of ICA (cf. [60, 67, 99, 100]).

Single trial data

The alternative approach, i.e., applying ICA to event-related single-trial EEG, is rarely studied [54, 39, 56] and suffers from the non-stationarity of EEG as well as from a low SNR of single-trial ERPs, embedded in ongoing EEG. Additionally, one fundamental question for the application of BSS poses itself by: *'How many sources?'*. Since the answer to this question directly addresses the issue whether ICA has to solve an under- or an over-determined system, this question is rather a fundamental data analytical issue than just a philosophical one. In the following we will assume that the number of sources exceeds the number of sensors. Hence we are facing the problem of under-determined BSS. As pointed out in [53], under these circumstances standard ICA techniques tend to extract mainly prominent signal sources, i.e., non-neural artifacts and non-phase-locked background brain activity, which are often much larger in amplitude than the ERPs one is interested in. This is basically due to the fact that the statistical optimization criteria (contrast functions) of ICA, such as kurtosis, negentropy, time lagged covariances, are dominated rather by the noise sources than by the weak ERP components. In addition, cortical sources of ERP are usually active only for a brief period of time. Consequently they are reflected to a minor extent in the statistical properties that are typically estimated as averages across time. For these reasons the application of ICA to single-trial EEG data is restricted to the extraction of the dominant sources instead of the weak and short lasting ERP components. Consequently ICA has been mainly used as a tool for removing artifacts such as eye blinks, power line noise or muscle movements from ongoing physiological recordings [38, 98, 99, 100, 111] and only occasionally for the separation of single-trial data into functionally independent sources [54, 56, 55, 106].

Solving an under-determined system (less sensors than sources) usually requires additional assumptions about the underlying sources, such as sparsity or super-Gaussian distributions [13, 59, 108]. In this chapter we tackle this issue by exploiting prior knowledge about the sources, concentrating on utilizing the phase-locked characteristics of ERP signals to improve on their extraction.

5.2 Incorporating physiological prior knowledge

Before we introduce the incorporation of a spatial prior into the ICA framework, we will briefly outline the mathematical foundations, and specify the performance measure, which is later on used for evaluation.

5.2.1 Mathematical preliminaries

Let $\mathbb{X} = (\boldsymbol{X}_t)_{t\in\mathcal{T}}$ denote the stochastic process of the multi-channel EEG over a temporal index set $\mathcal{T} \subset \mathbb{Z}$, representing the individual sampling points of each single trial. Here $\boldsymbol{X}_t$ represents a multivariate random variable. Furthermore we consider $\mathbb{X}$ to be generated as a stationary linear mixture of $M > N$ independent sources $\mathbb{S}$, i.e., $\mathbb{X} = A \cdot \mathbb{S}$. The set of K realizations (single trials) of $\mathbb{X}$ is given by $\mathcal{D} = \left\{\boldsymbol{X}_{\cdot}^{k}\right\}_{k=1}^{K}$. Whenever we would like to refer to an individual channel $c \in 1, \ldots, C$ or source i, we explicitly use $X_{c,t}^{k}$ and $S_{i,t}^{k}$ respectively. Without a loss of generality, we assume the event-related (phase-locked) sources to be embedded in an M_e-dimensional subspace, spanned by the first $M_e < N$ independent sources. The remaining $M - M_e$ dimensions are characterized by artifacts and non-phase-locked background brain sources. Thus averaging across trials yields:

$$\mathbb{E}[S_i] \equiv 0, \quad \forall\, i > M_e \tag{5.3}$$

$$\mathbb{E}[S_i] \not\equiv 0, \quad \forall\, i \leq M_e. \tag{5.4}$$

Note that (5.4) is not restricted to identical single-trial responses for the event-related sources. It only assumes the existence of stimulus locked components that will not vanish asymptotically under trial-averaging. Consequently (5.4) also covers ERPs that undergo single trial variability either in amplitude or in latency.

Consequently, for the EEG signal X_c acquired by the c^{th} electrode it follows

$$\mathbb{E}[X_c] = \mathbb{E}\left[\sum_{i=1}^{M} A_{c,i} \cdot S_i\right] = \sum_{i=1}^{M_e} A_{c,i} \cdot \mathbb{E}[S_i]\,. \tag{5.5}$$

Thus averaging over trial asymptotically maintains only information about the phase-locked components and converges to the M_e-dimensional subspace spanned by the phase-locked sources. Implicitly we assume stationarity of the spatial coupling of the ERP sources with the electrodes, represented by the columns $A_{\cdot,i}$, $i < M_e$ of the mixing matrix.

Performance measure

In order to access the quality of an extracted phase-locked source, we measure its SNR. The SNR is commonly defined as the ratio between the variance (power) of the averaged ERP (*signal*) and the averaged variance of the single trial residuals (*noise*):

$$\mathrm{SNR} := \frac{\mathrm{Power}\,(\mathrm{ERP})}{\mathbb{E}\,[\mathrm{Power}\,(\mathrm{trial} - \mathrm{ERP})]}. \tag{5.6}$$

Note that the variance (power) is computed pathwise, i.e., for each realization of the stochastic process (single trial) across time. We will explicitly denote this by $\mathrm{Var}_t[\cdot]$. Conversely the expectation $\mathbb{E}[\cdot]$ corresponds to averaging over trials. Given an estimate of the decomposition matrix W and using the same notation for the recovered sources $\mathbb{Y} = W \cdot \mathbb{X}$ as previously introduced in the context of $\mathbb{X}$, the SNR of a source signal Y_i reads as

$$\mathrm{SNR}(Y_i) \quad = \quad \frac{\mathrm{Var}_t[\,\mathbb{E}[Y_i]]}{\mathbb{E}[\,\mathrm{Var}_t[Y_i - \,\mathbb{E}[Y_i]]]}. \tag{5.7}$$

Remark 5.2.1. It is easy to prove that the above definition of SNR is invariant with respect to rescaling the signals. This is important, as ICA recovers the independent sources uniquely except for scaling.

5.2.2 Temporal transformation

As carried out in [53] in an under-determined environment (more linearly mixed sources than sensors), BSS techniques tend to extract sources that are most prominent with respect to the statistical optimization criteria, such as kurtosis, negentropy or time lagged covariance. In order to redirect the focus of ICA on the event-related, phase-locked signal subspace, we utilize property (5.5) and define a filter $L(\mathbb{X})$ of the mixed process, that enhances the signal along the direction of the event-related components, while dampening all noise directions. In order to control the degree of the signal amplification a regularization parameter $\lambda \in [0, 1)$ is included:

$$L_\lambda(\mathbb{X}) : \boldsymbol{X}_t \longmapsto (1-\lambda)\boldsymbol{X}_t + \lambda\,\mathbb{E}[\boldsymbol{X}_t]\,,\ \forall t \in \mathcal{T}. \tag{5.8}$$

For the set of realizations (single trials) $\mathcal{D} = \left\{\boldsymbol{X}^k_\cdot\right\}_{k=1}^K$ this translates to

$$L_\lambda(\mathcal{D}) : \boldsymbol{X}^k_\cdot \longmapsto (1-\lambda)\boldsymbol{X}^k_\cdot + \lambda\bar{\boldsymbol{X}}_\cdot,\ k = 1, \ldots, K, \tag{5.9}$$

where $\bar{\boldsymbol{X}}_{\cdot}$ represents the empirical expectation (average across trials). Each single trial $\boldsymbol{X}_{\cdot}^{k}$ is replaced by the weighted average, more precisely a convex combination, of itself and the average response. Raising the parameter λ from zero towards one increasingly replaces the single-trial responses by the averaged responses. Simultaneously the noise contained in the single trial is monotonically suppressed. Consequently the distribution of the data becomes increasingly concentrated onto the subspace spanned by the phase-locked ERP components. This property is illustrated on artificially generated data in Fig. 5.2.

Obviously the temporal transformation $L_\lambda(\mathbb{X})$ in (5.8) is linear. Based on this observation we can easily derive the following properties.

Lemma 5.2.1 (Properties of $L_\lambda(\mathbb{X})$). *Let $\lambda \in [0,1)$ and L_λ be defined according (5.8), then for $\mathbb{X} = A \cdot \mathbb{S}$ the following holds:*

i) L_λ is invertible, i.e., L_λ^{-1} exists

ii) the linear mixture model is invariant under L_λ, i.e., $L_\lambda(\mathbb{X}) = AL_\lambda(\mathbb{S})$.

Proof. i) In order to prove the existence of the inverse of L_λ, we will construct it explicitly. To this end, let $\lambda \in [0,1)$ be fixed and $\mathbb{Y} = L_\lambda(\mathbb{X})$, i.e., $\boldsymbol{Y}_t = (1-\lambda)\boldsymbol{X}_t + \lambda\,\mathbb{E}[\boldsymbol{X}_t]\,,\ \forall t \in \mathcal{T}$. Obviously the following equivalence holds $\mathbb{E}[\boldsymbol{Y}_t] = (1-\lambda)\,\mathbb{E}[\boldsymbol{X}_t] + \lambda\,\mathbb{E}[\,\mathbb{E}[\boldsymbol{X}_t]] = \mathbb{E}[\boldsymbol{X}_t]$, which implies $\boldsymbol{Y}_t - \lambda\,\mathbb{E}[\boldsymbol{Y}_t] = (1-\lambda)\boldsymbol{X}_t$. Dividing both sides of this equation by $(1-\lambda)$ recovers $\boldsymbol{X}_t$. Thus the inverse of L_λ for $\lambda \in [0,1)$ is given by:

$$L_\lambda^{-1}(\mathbb{Y}) : \boldsymbol{Y}_t \longmapsto \frac{\boldsymbol{Y}_t - \lambda\,\mathbb{E}[\boldsymbol{Y}_t]}{1-\lambda},\ \forall t, \tag{5.10}$$

ii) In order to prove the invariance of the linear mixture model under the transformation L_λ it is sufficient to show that L_λ is linear. To this end, let $\mathbb{Y} = L_\lambda(\mathbb{X})$ and $\boldsymbol{X}_t = A\boldsymbol{S}_t$, then

$$\begin{aligned} \boldsymbol{Y}_t &= (1-\lambda)\boldsymbol{X}_t + \lambda\,\mathbb{E}[\boldsymbol{X}_t] && (5.11)\\ &= (1-\lambda)A\boldsymbol{S}_t + \lambda\,\mathbb{E}[A\boldsymbol{S}_t] && (5.12)\\ &= A\left((1-\lambda)\boldsymbol{S}_t + \lambda\,\mathbb{E}[\boldsymbol{S}_t]\right), && (5.13) \end{aligned}$$

and finally $L_\lambda(\mathbb{X}) = A \cdot L_\lambda(\mathbb{S}),\ \forall \lambda \in [0,1)$.

□

Note, as the transformation is invertible for all $\lambda \in [0, 1)$, the spatial and spectral information about the noise processes is preserved under the transformation.

Remark 5.2.2. Along with L_λ its inverse L_λ^{-1} is also linear, thus the following holds for all $\lambda \in [0, 1)$

$$L_\lambda^{-1}\left(W \cdot L_\lambda(\mathbb{X})\right) = W \cdot L_\lambda^{-1} L_\lambda(\mathbb{X}) = W\mathbb{X}. \tag{5.14}$$

This ensures that a decomposition matrix W, obtained on the basis of the transformed data $L_\lambda(\mathcal{D})$, is directly applicable to the raw single-trial $\boldsymbol{X}^k$. Note that λ equal zero corresponds to raw single-trial data, while $\lambda \to 1$ applies to the trial-averaged data.[1] Consequently, by virtue of the transformation (5.8) we are able to trade off between applying ICA on single trial data in noisy environments and the decomposition of the averaged responses.This particular processing of the single trial data, prior to the application of any ICA algorithm, enables us to redirect the focus of the separation onto the event-related signal subspace, while simultaneously sustaining the information about the spatial and spectral structure of the single-trial noise.

In order to obtain improved ERP-components we simply apply an ICA method at several degrees of regularization $\lambda \in [0, 1)$ and decompose the raw data using the correspondingly estimated demixing matrices $W(\lambda)$. This yields different estimations of the underlying independent sources, i.e., $\boldsymbol{Y}(\lambda) = W(\lambda)\boldsymbol{X}$. At each degree λ we then identify the extracted phase-locked component and evaluate its signal quality. Finally we simply take the decomposition of the data that extracts the ERP-component best.

Remark 5.2.3. If the feasible set of regularization parameters contains $\lambda = 0$, then the solutions of the non-regularized vanilla ICA are considered in the model selection. Consequently the signal quality of the ERPs extracted by the proposed method is trivially greater than or equal to the achieved SNR of the vanilla approach.

5.3 Experiments

In this section we will demonstrate the advantage of the proposed method for the extraction of single-trial ERPs. For this purpose, we will first study its application in a controlled environment of artificially generated data. In this artificial setting we will

[1] In the case of $\lambda = 1$ (decomposition of the averaged data) the transformation L_λ is not invertible, thus the estimated filters W_λ are not meaningfully applicable in order to decompose the single trial data.

embed one single simulated ERP-component in a three dimensional noisy environment and will compare the benefit of our method in relation to a standard ICA approach. The application to artificial data will be followed by real world examples of improved extraction of somatosensory evoked potentials from multichannel EEG-recordings.

For the decomposition of the multivariate data into independent sources we will apply the TDSEP-algorithm [5, 110].

5.3.1 Artificial data

The verification of novel approaches on the basis of artificially generated data is quite advantageous, as it allows us to examine the response of the system with respect to changes of certain environmental parameters. Thus, for our pupose we generate low dimensional surrogate EEG data.

Data generation

To meet the assumption of under-determined BSS while keeping things simple, we simulate three EEG channels as a linear mixture of four independent artificial sources (simulated ERP, 10Hz narrow band source, white Gaussian and $\frac{1}{f}$ noise; see Fig. 5.1). In order to validate our approach at different initial SNR of the ERP, we generated different data sets by scaling the amplitude of the normalized ERP-component with a factor $\sigma \in [0.01, 10]$, while keeping the non-phase-locked sources normalized and the mixing matrix A fixed. In particular A was chosen as:

$$A = \begin{pmatrix} 0.5 & 1 & 1 & -0.1 \\ 1 & 0.1 & 1 & 1 \\ 0.5 & 1 & 1 & 1 \end{pmatrix} \cdot D,$$

where D is the diagonal matrix, such that the Euclidean norm of the columns of A is normalized to unity. Each of the simulated data sets consists of 100 single trials. The task for the ICA algorithm is to recover the ERP component from the simulated single trial EEG.

Note that any invertible decomposition matrix of the data corresponds to a basis of $\mathbb{R}^3$. Consequently there will exist at least one independent component that contains parts of the phase-locked ERP signal. Moreover, from the particular definition of the

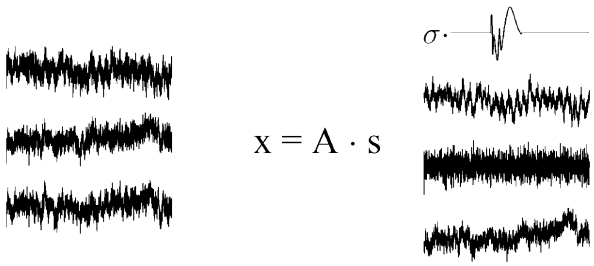


Figure 5.1: Three simulated EEG channels, given as a stationary linear mixture of four artificially generated sources, i.e., ERP, 10Hz, white Gaussian noise and $\frac{1}{f}$ noise. The scaling factor σ, here figuratively indicated as a multiplier in front of the ERP signal, is used to generate several dataset with a different initial SNR ($\{\sigma_i \in [0.0110]\}$).

mixing process it follows immediately that the ERP signal can not be recovered perfectly, i.e, there exists no direction with an infinite SNR. Hence it is the goal of the regularization approach to find a separation of the data, such that the ERP signal is recovered by a *single* independent component and at a high SNR.

Results and discussion

For each data set, indexed by $\sigma \in [0.01, 10]$ we transform the single trials according to the transformation (5.8) at different values of λ from a fixed, selected set $\Lambda \subseteq [0, 1)$. Fig. 5.2 shows the 3D-scatter plots of the transformed data for different values λ. As it was expected, the data gets increasingly concentrated along the direction of the ERP signal. The transformed data was subsequently processed by the TDSEP algorithm. For each data set this yields a collection of decomposition matrices $\{W(\lambda) : \lambda \in \Lambda\}$. According to the basic properties of the transformation, we can apply $W(\lambda)$ directly to the raw data, yielding differently recovered sources $\boldsymbol{Y}(\lambda) = W(\lambda)\boldsymbol{X}$. At each degree of regularization we determine the ERP source as the independent component with the largest SNR. The optimal degree of regularization λ^* is then individually defined for each data set, as the regularization value $\lambda \in \Lambda$, that yields the maximum SNR.

For each data set we refer to the ratio between the SNR at the optimal level of regularization and the SNR of the standard ICA ($\lambda = 0$) as the relative gain in SNR, i.e.,

$$\text{relative gain}(\lambda) = \frac{\text{SNR}(Y_i(\lambda))}{\text{SNR}(Y_i(0))}. \tag{5.15}$$

In Fig. 5.3-left we depict the relative gains for all data sets, indexed by $\sigma \in [0.01, 10]$.

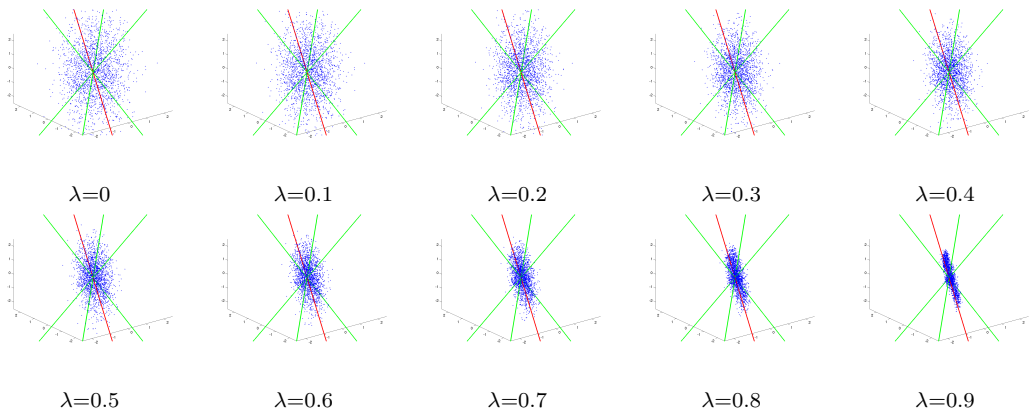


Figure 5.2: Each panel depicts a 3D-scatter plot of the transformed data at certain degrees of regularization. The green lines indicate the spatial directions (columns of the mixing matrix A) of the three noise signals, while the red line corresponds to the spatial direction of the ERP signal. Along with an increasing degree of regularization the data gets stronger concentrated along the subspace spanned by the ERP.

At very low initial SNR of the ERP, i.e., $\sigma \leq 0.1$ there is no improvement, which is to be expected since the ERP signal is buried under a strong noise floor, even on average due to the small amount of just one hundred single trials. Thus the spatial direction of the average across trials is still dominated by the dominant non-phase-locked components rather than by the ERP sources and consequently the regularization can not provide a strong enough bias toward the ERP subspace. This changes drastically as the strength of the raw ERP signal increases with a peak performance at $\sigma = 0.8$. When the ERP becomes more strongly pronounced in the raw data ($\sigma > 0.8$), the relative gain – although above one – starts to decay. This coincides with the level, at which the ERP source in the mixture becomes stronger pronounced, such that even the vanilla ICA starts extracting the ERP. It is worth mentioning that even in this situation, when ICA starts to extract the ERP signal by itself, the regularization approach improves slightly. To give an impression of the strength of the provided bias, the averaged ERP of the simulated EEG channel with the best SNR (second channel) is shown as inserted plots in the left panel of Fig. 5.3 at three different levels of ERP amplitude, i.e., $\sigma = \{0.1, 0.8, 3\}$.

The right panel of Fig. 5.3 provides information about the particularly selected optimal degree of regularization, λ^* for each data set. As discussed before, for data sets with an ERP amplitude $\sigma \leq 0.1$, even regularization does not help to extract the ERP source of marginal signal strength. For these data sets the SNR remains unchanged

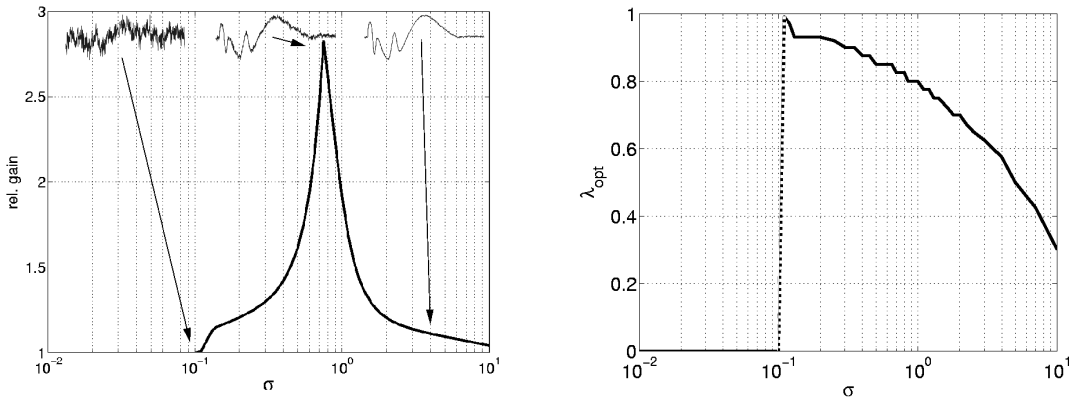


Figure 5.3: The left panel compares the optimal solutions found by the proposed method and by non-regularized, standard ICA. More precisely it presents the relative gain (cf. (5.15)) in the SNR of the extracted phase-locked component for the different data sets (indexed by $\sigma \in [0.01, 10]$). For very weak signals there is an almost no improvement, due to the fact that even the average shows no clear ERP-signal (illustrated by the inserted plots). The right panel gives the corresponding optimal degree of regularization, at which the SNR is maximized for each data set. Note in cases where all degrees of regularization yield the same SNR, we conservatively prefer lower degrees. Especially for the data sets with ERP amplitude $\sigma < 0.1$ the SNR is equal at each degree of regularization, consequently the optimum refers to the standard ICA solution at $\lambda = 0$.

at each degree of regularization. In such a case of equal SNR we conservatively prefer smaller values for the optimal degree of regularization, hence for these data sets λ^* equals to zero. If the ERP becomes slightly more pronounced in the raw data, but stays at a low level, it requires a high degree of regularization ($\lambda^* = 0.99$) in order to extract the weak ERP source from the over-complete mixture. For data sets with an even larger ERP amplitude, the degree of regularization that is needed for recovering the ERP signal best, reduces monotonically.

In order to further elaborate on the properties of the regularization scheme, we study one exemplarily chosen data set at an ERP amplitude level of $\sigma = 0.8$. For that particular data set Fig. 5.4 shows the evolution of the SNR of the three decomposed sources as functions of the regularization parameter λ. At $\lambda = 0$ the SNR refers to the solution of standard ICA (no regularization). Obviously the non-regularized, standard ICA does not focus on the extraction of the ERP signal, which is evident from basically two observations: the low SNR and the simultaneous presence of two components with a non-zero SNR. Increasing the degree of regularization forces the separation process to focus on the extraction of the ERP into one independent component and increases the

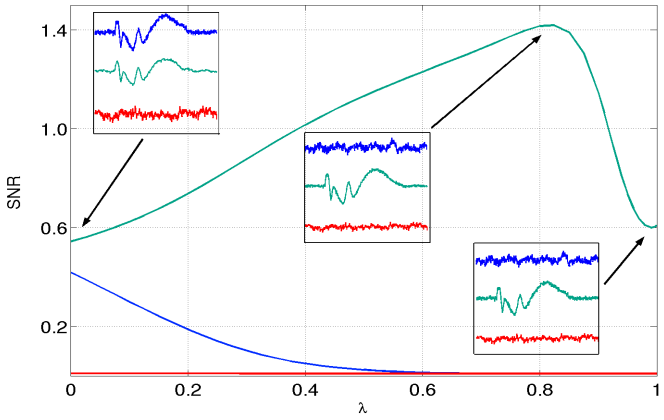


Figure 5.4: The SNR of the three recovered sources as a function of the regularization parameter λ for a specifically chosen artificial data set ($\sigma = 0.8$). The green line reflects the SNR of the recovered source that is associated with the ERP. Due to the regularization the SNR of that particular source at $\lambda^* = 0.85$ is almost thrice the SNR of the vanilla ICA solution at $\lambda = 0$. The inserted plots shows the averaged signal (100 single trials) for the three recovered sources at three exposed degrees of regularization, i.e., $\lambda = \{0, 0.85, 0.95\}$. At $\lambda = 0$ the ERP signal is present in two components, at $\lambda = 0.85$ the SNR of the ERP source has a clear maximum and the ERP is captured by a single source. However, at higher degrees of regularization, e.g., at $\lambda = 0.95$, the ERP is still represented in one component, but with a lower SNR, mainly due an insufficient representation of the noise.

SNR of the extracted phase-locked component. For that particular data set the maximum in the SNR suggests an optimal degree of regularization at $\lambda^* = 0.85$. Increasing the regularization parameter further, the SNR of the extracted ERP component starts to decay. However, in strict contrast to the non-regularized decomposition, the ERP component is captured by a single component. The observed decrease in the SNR can be best explained by a phenomena termed over-regularization. Here the system is at a state where the averaged signals prevail and the extracted component becomes "less invariant" against the single trial noise, which is reflected in the decrease of the SNR. For this specific data set the optimal SNR yields an improvement by roughly a factor of 2.8 compared to the SNR of the solution obtained from the vanilla decomposition of the raw single trial data.

5.3.2 Somatosensory evoked potentials

In order to illustrate its usefulness on real data, we apply the proposed method to five data sets of single trial EEG recordings of Somatosensory Evoked Potentials (SEPs).

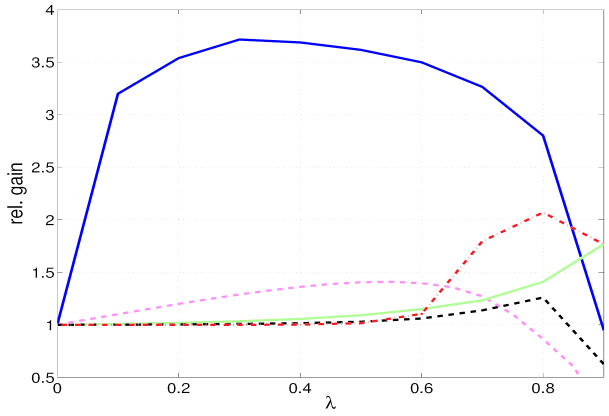


Figure 5.5: The figure presents the relative gain in the SNR of the extracted SEP component for five different subjects. The achieved relative gain is given as a function over the regularization parameter λ. The relative improvement strongly varies for the different data sets, ranging from a factor of 1.2 (black dashed) to a factor of 3.5 (blue solid).

SEPs excited by median nerve stimulations are well studied and various cortical responses with different timing and amplitudes are known, e.g., the earliest responses are at the contralateral primary somatosensory cortex (SI) [1, 2, 57] that is activated at $18 - 150$ ms, while later responses at ipsilateral SI [2, 57] and bilateral activations with similar timing in the secondary somatosensory cortex (SII) [57] have been reported.

Data acquisition

In the present study we will examine EEG recordings from five healthy subjects. The SEPs were excited by weak median nerve stimulation (MNS) delivered at the right wrist at an intensity of 25% above the individual sensory threshold, but well below the individual motor threshold. The intensities of the delivered stimuli for the different subjects range from 1.9 – 2.8 mA at a constant impulse-width of 0.2 ms. Each data set consists of 100 single trials of weak MNS. The used inter-stimulus interval was about 3 s with an additive uniformly distributed jitter ([0–250] ms). The EEG was recorded from 56 electrodes, placed on a subset of the 10-10 system [16]. The referential recordings (against nose reference) were sampled at 1 kHz. Prior to the analysis, a bandpass filter in the range of $[0.1, 80]$ Hz was applied to the data.

Results and discussion

For each data set we transformed the single trials according (5.8) at different values $\lambda \in \Lambda \subseteq [0, 1)$, and finally applied the TDSEP algorithm to the transformed data separately. This yields separate collections of decomposition matrices $\{W(\lambda) : \lambda \in \Lambda\}$ and correspondingly differently recovered source signals. At all degrees of regularization, the estimated components could either be distinguished by its phase-locked or non-phase-locked property. Throughout all degrees of regularization we identified one component in each data set that was persistently extracted and could clearly be identified by means of similarity of the spatial distribution and similar time courses of the averaged signal. In particular these components enable us to directly quantify the improvement in the SNR depending on the degree of regularization. As in section 5.3.1 we will refer to the ratio of the SNR of an ERP source, recovered at a specific degree of regularization, and the SNR of an ERP source obtained by the standard ICA solution ($\lambda = 0$) as the relative gain (cf. (5.15)). Fig. 5.5 shows the relative gain as a function of the regularization parameter λ for the extracted SEP component of the five different data sets. For each data set the SNR of the extracted SEP source monotonically ascends with an increasing degree of regularization up to a clear maximum. The achieved peak performance in the relative gain for the different data sets ranges from 25% to 270%, which corresponds to an improvement of the SNR by a factor $1.25 - 3.7$, relative to the solution of vanilla ICA. The different performance gains resemble the observed differences of the obtained achievement on the artificial data, see Fig. 5.3. Again these differences are probably due to the different statistical confidence about the provided spatial bias, and therefore are directly linked with the ratio between the signal strength of the ERP source and the interfering noise, i.e., the SNR of the raw data.

For one exemplary chosen data set the SNR as a function of the regularization constant λ is shown in Fig. 5.6. At three different degrees of regularization, $\lambda \in \{0, 0.55, 0.9\}$, the spatial distribution at the scalp and the averaged evoked response of the recovered ERP component are inserted, emphasizing the similarity of the extracted sources. With λ, starting at zero (the vanilla decomposition of the raw single trial data), the SNR increases monotonically up to a clear maximum at $\lambda^* = 0.55$. This maximum in SNR can be interpreted as the best separation into the signal and the noise space, with respect to this component. Increasing λ further drives the system into

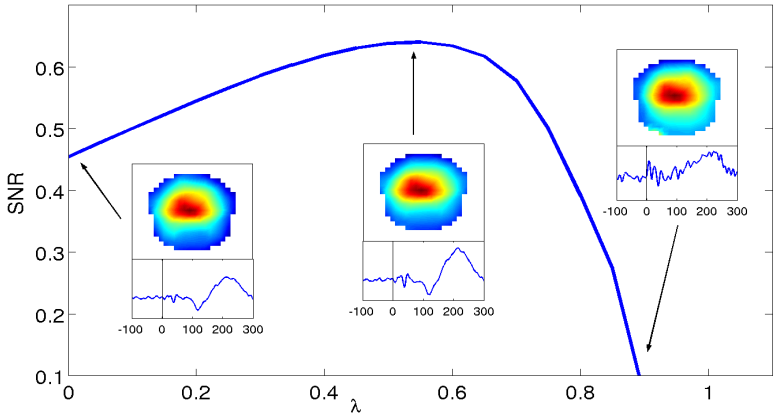


Figure 5.6: The estimated SNR of the decomposed phase-locked source as a function of the regularization parameter λ for one particular data set (magenta dashed line in Fig. 5.5). Examples of the corresponding scalp patterns and the averaged ERPs are shown at three degrees of regularization, i.e., $\lambda = \{0; 0.55; 0.9\}$. The scalp distributions of the source are almost identical, as well as the averaged responses, emphasizing the equality of the sources.

a state where the averaged signals prevail and the extracted component becomes "less invariant" against the single trial noise, which is reflected by the decrease of the SNR.

Summary

We have introduced a novel approach improving the decomposition of single-trial ERPs in terms of an increased SNR. To this end, we incorporated prior knowledge about the phase-locked property of the signals of interest into the source separation framework. By virtue of a linear, temporal transformation of the data we enabled the ICA method to trade off between single-trial decomposition and the separation of the averaged responses. In particular, the suggested method is incorporated into a regularization scheme, providing a parameter for controlling the degree of modification to the data prior to the application of ICA. Favorably, the proposed transformation does not depend on the specific choice of the source separation method in use and can be applied prior to any ICA algorithm. Moreover, the estimated decomposition matrix, that is determined on the basis of the transformed data is directly applicable in the decomposition of the raw single trial data.

However, the identification of different ERP sources and their allocation to a par-

ticular ERP component across different degrees of regularization remains an open issue. In the present study we solved the issue of identification by visual inspection of the spatial distribution and the averaged signal of the recovered sources.

The benefit of the proposed method was verified by an improved SNR of extracted phase-locked components from both simulated data and multichannel EEG-recordings of MNSs. Although the set-up for the simulated data is rather artificial, one could clearly observe and quantify the gain, achieved by regularizing the ICA methods, in terms of an improved SNR of the recovered ERP source. On the other hand, this example on simulated data also reveals the limitation of the method: if the underlying signal of interest is too weak compared to the noise or the number of trial is too limited, such that the ERP remains hidden even under trial averaging, then also regularization cannot help. The successful application to the multichannel EEG-recordings led to an improvement of the SNR of one extracted SEP components by a factor ranging from 1.25 to 3.7 for data from 5 different subjects.

Chapter 6

Temporal evidence accumulation

In this chapter we will present a general Bayesian classification framework that efficiently combines sequences of features across time. In particular we will introduce the algorithm by means of an example of discrimination between imaginary left and right hand movement. The effectiveness of this procedure was verified by its successful application to data from the BCI competition in 2005 [12]. Disclosure of the testing data after the competition deadline revealed this approach to be superior to all other competing algorithms. More precisely, the method that will be outlined in the following, is an upgraded version of the winning method in 2003 [11] and therefore succeeded at two different BCI competitions.

Although the feature extraction method is intended for the classification of imaginary hand movements, the general algorithm is applicable to binary decision processes on the basis of noisy multivariate feature sequences. However, to provide a decision at every time instance our method gathers information from the distinct features across time. To this end, we train sequences of weak classifiers (one per feature and time instance). Afterwards these classifiers are combined in a sophisticated probabilistic manner across time. In particular we suggest a causal weighting scheme that reflects the discriminative power of each classifier. Here the causality constraint of the decision-making process had to be satisfied as a mandatory prerequisite of a real-time BCI system.

6.1 Preliminaries

Before we introduce the algorithmic details, we will briefly review the competition objectives, the neurophysiological features and state some preliminary notes on Bayes decision theory.

6.1.1 Competition data and objectives

As previously stated our algorithm has been successfully applied in the 2003 and 2005 international data analysis competitions on BCI-tasks [12] (data set III and IIIb, respectively) for the classification of imaginary left and right hand movements. The joint objective of both competitions was to detect the respective motor intentions as early and as reliably as possible after a given cue signal. Consequently, the competing algorithms were allowed to use the provided information about the event onset. So it was not within the scope of the competition to detect the event onset itself, as it would be mandatory to operate a BCI-system in an asynchronous mode.

However, in the following we restrict ourselves to data from the competition in 2005. Here the EEG recordings from three different healthy subjects (O3, S4 and X11) were provided. Except for the first subject, each data set consists of 540 labeled (for training) and 540 unlabeled trials (for evaluation) of imaginary hand movements, with an equal number of left and right hand trials (the first data set provides just 320 trials each). Each trial has a duration of 7 s: after a 3 s preparation period a visual cue is presented for one second, indicating the ordered motor intention. This is followed by another 3 s for performing the imagination task (for details see [12]). In particular the EEG recordings from two bipolar channels (C3, C4) were provided by the Dept. of Med. Informatics, Inst. for Biomed. Eng., Univ. of Techn. Graz with bandfilter settings of 0.5 to 30 Hz and sampled at 128 Hz. The specific competition task is to provide an ongoing discrimination between left and right movements for the unlabeled single trials. More precisely, at every time instance in the interval from 3 to 7 seconds a strictly causal decision about the intended motor action must be supplied. Moreover, the magnitude of the provided feedback signal was requested to reflect the degree of confidence into the decision.

6.1.2 Neurophysiological features

The human perirolandic sensorimotor cortices show rhythmic macroscopic EEG oscillations (μ-rhythm) [32], with spectral peak energies around 10 Hz (localized predominantly over the postcentral somatosensory cortex) and 20 Hz (over the precentral motor cortex). Modulations of the μ-rhythm have been reported for different physiological manipulations, e.g., by motor activity, both actual and imagined [37, 77, 88], as well as by somatosensory stimulation [72]. Standard trial averages of μ-rhythm power show a sequence of attenuation, termed ERD [77], followed by a rebound (event-related synchronization: ERS) which often overshoots the pre-event baseline level [85].

In case of sensorimotor cortical processes accompanying finger movements Babiloni et al. [3] demonstrated that movement related potentials (MRPs) and ERD indeed show up with different spatio-temporal activation patterns across primary (sensori-)motor cortex, supplementary motor area and the posterior parietal cortex. Most importantly, the ERD response magnitude did not correlate with the amplitude of the negative MRPs slope.

Most of the pursued non-invasive BCI approaches use the accompanying EEG-rhythm perturbation in order to distinguish between single trials, e.g., of left and right hand imaginary movements [48, 70, 79, 103]. Up to now there are only a few approaches additionally using slow cortical potentials [20, 22, 61]. In the following we use both features. Thus, in order to extract the rhythmic features we map the EEG to the time-frequency domain by means of Morlet wavelets [96], whereas the slow cortical MRP are extracted by the application of a low pass filter, in form of a simple moving average.

6.1.3 Bayes decision theory

Since the modulations of the ongoing rhythmic activity and the slow cortical movement related potential are expected to be differently pronounced over time, we suggest combining these features adaptively across time on the basis of their instantaneous discriminatory power.

Bayes Formula: Suppose we know both the class conditional distributions $p(\boldsymbol{x}\,|\omega_j)$ as well as the prior distributions $p(\omega_j)$ for $j = 1, 2$. Then the joint probability to observe $\boldsymbol{x}$ along with class ω_j can be either written as $p(\boldsymbol{x}, \omega_j) = p(\boldsymbol{x}\,|\omega_j)\, p(\omega_j)$ or

$p(\boldsymbol{x}, \omega_j) = p(\omega_j \,|\boldsymbol{x})\, p(\boldsymbol{x})$. Rearranging these gives rise to

$$p\left(\omega_j \,|\boldsymbol{x},\right) = \frac{p\left(\boldsymbol{x} \,|\omega_j\right)\, p(\omega_j)}{p\left(\boldsymbol{x} \,|\omega_1\right)\, p(\omega_1) + p\left(\boldsymbol{x} \,|\omega_2\right)\, p(\omega_2)}. \tag{6.1}$$

which is known as *Bayes formula* and expresses the class posterior distribution (given an observation $\boldsymbol{x}$) in terms of the likelihoods and the priors.

Bayes error of misclassification: Let $\mathcal{R}_1 := \{\boldsymbol{x} : p\left(\omega_1 \,|\boldsymbol{x}\right) \geq p\left(\omega_2 \,|\boldsymbol{x}\right)\}$ and $\mathcal{R}_2$ denotes its complementary set, where the posterior probability of the second class exceeds those of the first, i.e., $\mathcal{R}_2 = \mathcal{R}_1^c$. Following Bayes decision rule, we decide ω_1 if $\boldsymbol{x} \in \mathcal{R}_1$ and ω_2 otherwise. This particular decision rule leads to the following definition of the *Bayes error of misclassification*:

$$\begin{aligned} P(error) &= P(\boldsymbol{x} \in \mathcal{R}_1, \omega_2) + P(\boldsymbol{x} \in \mathcal{R}_2, \omega_1) && (6.2)\\ &= \int_{\mathcal{R}_1} p\left(\boldsymbol{x} \,|\omega_2\right) p\left(\omega_2\right) \mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_2} p\left(\boldsymbol{x} \,|\omega_1\right) p\left(\omega_1\right) \mathrm{d}\boldsymbol{x}. && (6.3) \end{aligned}$$

In general, the Bayes error of misclassification cannot be calculated directly.

Remark 6.1.1. The Bayes error of misclassification is upper bounded by the minimum of the class priors $p\left(\omega_1\right)$ and $p\left(\omega_2\right)$. This can be easily seen, as

$$\begin{aligned} P(error) &= \int_{\mathcal{R}_1} p\left(\boldsymbol{x} \,|\omega_2\right) p\left(\omega_2\right) \mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_2} p\left(\boldsymbol{x} \,|\omega_1\right) p\left(\omega_1\right) \mathrm{d}\boldsymbol{x} && (6.4)\\ &\leq \int_{\mathcal{R}_1} p\left(\boldsymbol{x} \,|\omega_1\right) p\left(\omega_1\right) \mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_2} p\left(\boldsymbol{x} \,|\omega_1\right) p\left(\omega_1\right) \mathrm{d}\boldsymbol{x} && (6.5)\\ &= \int p\left(\boldsymbol{x}, \omega_1\right) \mathrm{d}\boldsymbol{x} = p\left(\omega_1\right). && (6.6) \end{aligned}$$

In a similar manner, we derive $p\left(\omega_2\right)$ as an upper bound of $P(error)$. Merging both inequations yields $P(error) \leq \min\left\{p\left(\omega_1\right), p\left(\omega_2\right)\right\}$.

Discriminative Power: In order to express the discriminative power v between the two class conditional distributions, we use the Bayes error of misclassification. In particular we propose

$$v := 1 - \frac{P\left(error\right)}{\min\left\{p\left(\omega_1\right), p\left(\omega_2\right)\right\}}. \tag{6.7}$$

According to remark 6.1.1, this definition restricts the values of the discriminative power to the interval $[0, 1]$. In this setting '0' refers to non-discriminative information, while '1' indicates perfect separability. However, as the Bayes error cannot be calculated directly

for most distributions, we approximate the Bayes error by the Chernoff bound [25], i.e.,

$$P(error) \leq \min_{0\leq\gamma\leq 1} p\left(\omega_1\right)^{\gamma} p\left(\omega_2\right)^{1-\gamma} \int p\left(\boldsymbol{x}\,|\omega_1\right)^{\gamma} p\left(\boldsymbol{x}\,|\omega_2\right)^{1-\gamma} \mathrm{d}\boldsymbol{x}. \tag{6.8}$$

Fortunately the Chernoff bound can be evaluated analytically if the class conditional distributions are Gaussian [25]. Consequently we approximate the class conditional distribution by a multivariate Gaussian. To this end, let $p\left(x\,|\omega_j\right) = \mathcal{N}\left(\mu_j, \Sigma_j\right)$, where $\mathcal{N}\left(\mu_j, \Sigma_j\right)$ denotes the PDFs of the multivariate Gaussian distributions with class mean μ_j and covariance matrix Σ_j for $j = 1, 2$. Then the integral in (6.8) can explicitly be written as (cf. [25])

$$\exp\left(-k\left(\gamma\right)\right) = \int p\left(\boldsymbol{x}\,|\omega_1\right)^{\gamma} p\left(\boldsymbol{x}\,|\omega_2\right)^{1-\gamma} \mathrm{d}\boldsymbol{x}, \tag{6.9}$$

where

$$\begin{aligned} k\left(\gamma\right) := & \frac{\gamma\left(1-\gamma\right)}{2}\ \left(\mu_1-\mu_2\right)^{\top}\left[\gamma\Sigma_1+\left(1-\gamma\right)\Sigma_2\right]^{-1}\left(\mu_1-\mu_2\right) \\ & + \frac{1}{2}\ln\frac{\det\left(\gamma\Sigma_1+\left(1-\gamma\right)\Sigma_2\right)}{\det\left(\Sigma_1\right)^{\gamma}\det\left(\Sigma_2\right)^{1-\gamma}}. \end{aligned} \tag{6.10}$$

Moreover, using an equal class prior $p(\omega_j) = \frac{1}{2}$, we finally approximate the discriminative power by

$$v \approx 1 - \min_{0\leq\gamma\leq 1} \exp\left(-k(\gamma)\right). \tag{6.11}$$

Note, that the minimum with respect to γ can be easily obtained, e.g., by application of a simple line search procedure.

6.2 The probabilistic model

Based on these preliminaries, we will now construct our probabilistic classifier. For this purpose, let $\mathbb{X} = (\boldsymbol{X}_t)_{t\in\mathcal{T}}$ denote the stochastic process representing the single trial EEG, indexed by the set $\mathcal{T}$. However, in order to guarantee strict causality, we need to restrict the feature extraction to a given observational horizon. To this end, we slightly extend our general notation. Whenever necessary, we indicate a observational horizon $s \in \mathcal{T}$ by the subscript $_{|s}$, e.g., $\boldsymbol{X}^k_{\cdot|s}$ refers to $\left(\boldsymbol{X}^k_{1\,|s}, \ldots, \boldsymbol{X}^k_{s\,|s}\right)$, where $\boldsymbol{X}^k_{t|s}$ indicates the time instance $t \leq s$ given the observational horizon s. For notational convenience we omit the index $_{|s}$ in cases when the single trial is entirely observed.

6.2.1 Feature extraction

Let $\mathcal{D} = \left\{\left(\boldsymbol{X}^k_{\cdot}, y^k\right)\right\}_{k=1}^{K}$ be the set of K labeled single trials. Where $y^k \in \{L, R\}$ indicates the label information, i.e., the intended motor action (left vs. right) of the k^{th} trial. In this context we refer to the subsets of trials labeled as *left* or *right* as $\mathcal{Y}^{(\text{L})} := \{k : y^k = L\}$ and $\mathcal{Y}^{(\text{R})}$, respectively. Furthermore, as the EEG was recorded just from two bipolar channels, namely C3 and C4, we refer to the individual components of $\boldsymbol{X}_t$ explicitly as C3_t and C4_t, respectively.

Considering ERD as a feature, we model the hand-specific time course of absolute μ-rhythm amplitudes over both sensorimotor cortices. To this end, we convolve the EEG signal with complex Morlet wavelets [96] in order to achieve a time-frequency representation of the single trials at two different frequency bands. Using the notation b_α and b_β for the wavelets centered at the individual spectral peak in the alpha (8-12 Hz) and the beta (16-24 Hz) frequency domain, the ERD feature of the k^{th} single trial, observed until time s is calculated as:

$$\text{erd}^k_{\cdot|s} = \left(\text{erd}^k_{1|s}, \ldots, \text{erd}^k_{s|s}\right),$$

where

$$\text{erd}^k_{t|s} = \begin{pmatrix} \left|(\text{C3}^k_{\cdot|s} * b_\alpha)[t]\right| \\ \left|(\text{C4}^k_{\cdot|s} * b_\alpha)[t]\right| \\ \left|(\text{C3}^k_{\cdot|s} * b_\beta)[t]\right| \\ \left|(\text{C4}^k_{\cdot|s} * b_\beta)[t]\right| \end{pmatrix}. \tag{6.12}$$

Moreover, we define the single trial feature for the MRP by convolution with a moving average filter of length $n \in \mathbb{N}$, denoted as b^n_{MA}.

$$\text{mrp}^k_{\cdot|s} = \left(\text{mrp}^k_{1|s}, \ldots, \text{mrp}^k_{s|s}\right),$$

where

$$\text{mrp}^k_{t|s} = \begin{pmatrix} \left(\text{C3}^k_{\cdot|s} * b^n_{\text{MA}}\right)[t] \\ \left(\text{C4}^k_{\cdot|s} * b^n_{\text{MA}}\right)[t] \end{pmatrix}. \tag{6.13}$$

In the final application the hyperparameters are subject to a model selection procedure, i.e., the center frequencies of the Morlet-wavelets in the alpha and beta band and the length of the moving average filter are selected by means of cross-validation.

6.2.2 Weak instantaneous classifiers

On the basis of the extracted single trial features, we model the class conditional distributions of each feature and at any time instance as a multivariate Gaussian distribution. So we have to estimate the class conditional means and covariance matrices of the features. In case of the ERD feature $\{\mathrm{erd}_t^k\}_{k=1}^K$ the moments of the multivariate Gaussian distributions at time t are estimated as:

$$\mu_t^{(y)} = \frac{1}{\left|\mathcal{Y}^{(y)}\right|} \sum_{k \in \mathcal{Y}^{(y)}} \mathrm{erd}_t^k \tag{6.14}$$

$$\Sigma_t^{(y)} = \frac{1}{\left|\mathcal{Y}^{(y)}\right|} \sum_{k \in \mathcal{Y}^{(y)}} \left(\mathrm{erd}_t^k - \mu_t^{(y)}\right)\left(\mathrm{erd}_t^k - \mu_t^{(y)}\right)^\top,\ y \in \{L, R\}. \tag{6.15}$$

For notational convenience we subsume the estimated parameters for the ERD feature at time t in $\Theta_t := \left(\mu_t{}^{(\mathrm{L})}, \Sigma_t{}^{(\mathrm{L})}, \mu_t{}^{(\mathrm{R})}, \Sigma_t{}^{(\mathrm{R})}\right)$. Accordingly $\Xi_t := \left(\eta_t{}^{(\mathrm{L})}, \Gamma_t{}^{(\mathrm{L})}\, \eta_t{}^{(\mathrm{R})}, \Gamma_t{}^{(\mathrm{R})}\right)$ denotes the estimated class conditional means and covariance matrices of the MRP features $\{\mathrm{mrp}_t^k\}_{k=1}^K$.

Given an arbitrary observation from each domain, and applying Bayes formula as introduced in (6.1), yields class posterior probability for each individual feature:

$$p\left(y \middle| \mathrm{erd}, \Theta_t\right), \quad \mathrm{erd} \in \mathbb{R}^4 \tag{6.16}$$

$$p\left(y \middle| \mathrm{mrp}, \Xi_t\right), \quad \mathrm{mrp} \in \mathbb{R}^2. \tag{6.17}$$

Additionally, on the basis of the estimated class conditional distributions and according to (6.11), we obtain estimates of the instantaneous discriminative power $(w_t)_{t \in \mathcal{T}}$ and $(v_t)_{t \in \mathcal{T}}$ of the ERD and the MRP feature, respectively.

6.2.3 Combining classifiers across time

In order to obtain a classification of the k^{th} unlabeled single trial at a specific time $s \in \mathcal{T}$, we incorporate knowledge from all preceding observations $t \leq s$, i.e., we combine the information derived from the causally extracted features: $\mathrm{erd}_{\cdot|s}^k$ and $\mathrm{mrp}_{\cdot|s}^k$. To this end, we first evaluate the class posteriors $p(y\,|\mathrm{erd}_{t|s}^k, \Theta_t)$ and $p(y\,|\mathrm{mrp}_{t|s}^k, \Xi_t)$ for all $t \leq s$. Secondly we combine the obtained class posteriors with one another and across time by taking the weighted average with respect to the discriminative power $w_{\cdot|s}$ and $v_{\cdot|s}$. In

particular, for $y \in \{\mathrm{L}, \mathrm{R}\}$ we define

$$c_s^k(y) := \frac{1}{\|w_{\cdot|s}\|_1 + \|v_{\cdot|s}\|_1} \sum_{t \le s} \begin{pmatrix} w_{t|s} \\ v_{t|s} \end{pmatrix}^\top \begin{pmatrix} p\big(y \big| \mathrm{erd}_{t|s}^k, \Theta_t\big) \\ p\big(y \big| \mathrm{mrp}_{t|s}, \Xi_t\big) \end{pmatrix} \tag{6.18}$$

$$= \sum_{t \le s} \frac{w_t \cdot p\left(y \middle| \mathrm{erd}_{t|s}^k, \Theta_t\right) + v_t \cdot p\big(y \big| \mathrm{mrp}_{t|s}, \Xi_t\big)}{\sum_{l \le s} w_l + v_l}. \tag{6.19}$$

Strictly speaking (6.19) gives the expectation value that the k^{th} single trial, observed until time s, is generated by either of the class models (L or R). This yields an evidence accumulation across time about the instantaneous decision process.

However, due to the competition requirements the final decision at time s was calculated as

$$d_s^k = 1 - 2 \cdot c_s^k(\mathrm{L}), \tag{6.20}$$

where a positive or negative sign refers to *right* or *left* movement, while the magnitude indicates the degree of confidence into the decision on a scale between 0 and 1.

6.3 Application

The competition data set for each subject consists of a training set and a testing set, each with an equal number of left and right hand trials (for further details please refer to paragraph 6.1.1).

On the basis of the labeled training data the ERD and MRP feature were extracted and the corresponding class conditional distributions along with the discriminative power were estimated. The weak classifiers (6.16) and (6.17) were then evaluated on the instantaneous (causally derived) ERD and MRP features of the unlabeled test trials. Finally, the probabilistic outputs of these classifiers were combined according to (6.19). For the k^{th} unlabeled test trial this yields a sequence of instantaneous probabilistic decisions $(d_t^k)_{t \in \mathcal{T}}$.

Remember, that the feature extraction relies on a few hyperparameters, i.e., the center frequencies of the wavelets, and the length of the MA filter. The optimal choice of these parameters was determined by a model selection procedure using a leave-one-out cross-validation scheme of the classification performance on the labeled training data. The optimal center frequencies were individually obtained at 11.5, 12.5, and 12 Hz in

the alpha band, and identically at 23 Hz in the beta band. The optimal length of the moving average filter was found at $n = 11$ samples, which corresponds to 86 ms.

After the closing deadline of the competition the true labels of the previously unlabeled trials (test set) were disclosed. Based on this label information $y^k \in \{L, R\}$ the ongoing probabilistic decisions $\{(d_t^k)_{t\in\mathcal{T}}\}_{k=1}^K$ for all test trials were evaluated in terms of the time course of the mutual information (MI)[87] that is defined as

$$\mathrm{MI}[t] = \frac{1}{2}\log_2\Big(\mathrm{SNR}[t]+1\Big) \tag{6.21}$$

$$\mathrm{SNR}[t] = \frac{\Big(\mathbb{E}[C_t\,|Y=L] - \mathbb{E}[C_t\,|Y=R]\Big)^2}{2\Big(\mathrm{Var}\,[C_t\,|Y=L] + \mathrm{Var}\,[C_t\,|Y=R]\Big)}. \tag{6.22}$$

As usual, the conditional expectations are replaced by their empirical estimates (averages over trials). Moreover, as the overall objective of the competition was to obtain a classification as fast and as accurately as possible, the maximum steepness of the MI was considered as the final evaluation criterion, i.e.,

$$\max_{t\geq 3.5\mathrm{s}} \frac{\mathrm{MI}[t]}{t-3\mathrm{s}}. \tag{6.23}$$

Note that the denominator adjusts for the 3s preparation period at the beginning of each trial, such that only the time after presenting the cue is considered.

6.3.1 Results

Disclosure of the test labels revealed our method to be superior to the competing algorithms with a MI steepness of 0.17, 0.44 and 0.35 for the three subjects O3, S4, and X11. Table 6.1 provides a comprehensive comparison of all submissions for data set IIIb of the BCI-competition in 2005. Basically this evaluation reveals that our proposed algorithm outperforms all competing approaches in terms of the achieved maximum binary classification error, the maximum MI, and the maximum steepness of MI.

The individual time courses of the MI and the steepness of the MI for the three subjects are presented in the right panel of Fig. 6.1. For all subjects the classification during the first 3.5 seconds is rather by chance. Immediately after 3.5 seconds (500 ms after the presentation of the cue signal) a steep ascent in the classification accuracy, reflected by a raising MI, can be observed for subject S4 and X11. For both subjects the maximum steepness of the MI is also obtained quite early, i.e., between $3.6-3.8$s. In

	min. error rate[%]			max. MI [bit]			max. MI/t [bit/s]		
	O3	S4	X11	O3	S4	X11	O3	S4	X11
1.	**10.69**	**11.48**	16.67	**0.603**	**0.608**	**0.486**	0.170	**0.438**	**0.349**
2.	14.47	22.96	22.22	0.447	0.232	0.307	0.163	0.417	0.172
3.	13.21	17.59	**16.48**	0.551	0.375	0.467	**0.203**	0.094	0.117
4.	23.90	24.44	24.07	0.218	0.239	0.217	0.115	0.122	0.118
5.	11.95	21.48	18.70	0.432	0.350	0.385	0.104	0.149	0.095
6.	**10.69**	13.52	25.19	0.597	0.567	0.244	0.118	0.152	0.061
7.	34.28	38.52	28.70	0.043	0.046	0.157	0.070	0.023	0.049

Table 6.1: Overall ranked performances of all competing algorithms (first row corresponds to the proposed method). For three different subjects (O3, S4 and X11) the table states different evaluation criteria, where the steepness of the MI was used as the final objective in the competition. For a description of the algorithm 2.–7. please refer to [86].

contrast to subject O3, here the maximum is achieved after 4.9 seconds, yielding a low steepness value. However, a low value is consistently achieved by all other competitors. Nevertheless, the MI constantly increases up to 0.64 bit per trial at 7 seconds. This may serve as an indicator of a delayed performance of subject O3.

Fig. 6.1-left provides the weights w_t and v_t respectively, which correspond to the instantaneous discriminative power according to (6.11). For subject S4 a switch in the regime between the ERD and the MRP feature at approximately 5 seconds is clearly observable by the crossing of the two sequences. This enables us to relate the steep increase in MI between 3 and 5 seconds mainly to the MRP feature. In contrast the continuing increase of the MI is primarily due to the ERD feature. Subject O3 provides

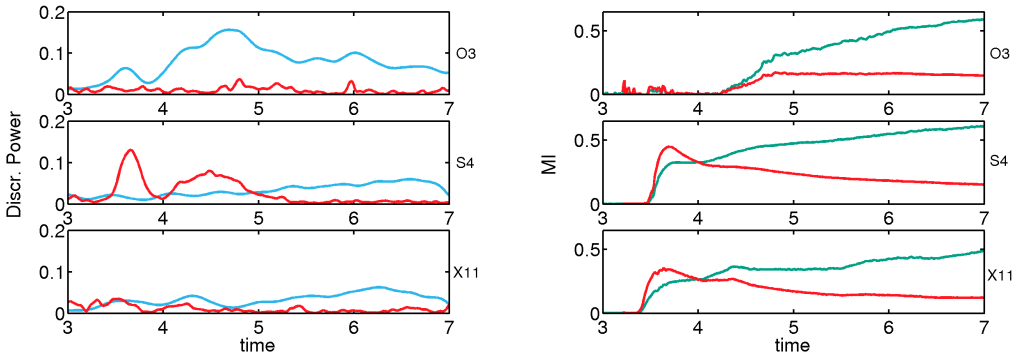


Figure 6.1: The left panel presents the estimated instantaneous discriminative power of the two different features sequences (ERD - light blue; MRP - red). The right panel depicts the time courses of the mutual information (green) and the competition criterion, i.e, the steepness of mutual information (red)

almost no discriminative MRP feature such that the classification is solely based on the ERD feature. Subject X11 exhibits a sequence of constantly low discriminative power of the class-conditional distributions. Nevertheless Fig. 6.1 depicts a MI that is continuously increasing and furthermore reports a surprisingly high steepness value of the MI. This observation clearly emphasizes the advantages of the evidence accumulation across time.

Summary

In this final chapter we proposed a general Bayesian classification framework for combining sequences of features across time. Although the method was originally intended for the real-time classification of imaginary hand movements, the general algorithm is applicable to binary decision processes on the basis of noisy multivariate feature sequences and is therefore applicable to any kind of sequential data posing a binary classification problems.

However, given a sequence of features and in order to provide an instantaneous classification at every time instance our method gathered information from the distinct features across time. To this end, we suggested training of a corresponding sequence of weak classifiers (one per feature and time instance). Afterwards these classifiers were combined in a sophisticated probabilistic manner across time. In particular we suggested a strictly causal weighting scheme that reflects the discriminative power of each feature at each time instance. Here we associated the discriminative power with the generalization error of the individual classifier. In the particular scenario of the BCI competition we used classifiers based on quadratic discriminant analysis, where the class-conditional distributions were modelled as multivariate Gaussian distributions with different covariance matrices. For this particular setting we derive the generalization error of the individual classifiers directly from the class-conditional distributions by means of the Bayes error of misclassification. However, due to the distinct covariance matrices we further approximated the Bayes error misclassification by the corresponding Chernoff bound. Integrating the information, provided by the distinct features, across time the method accumulates evidence about the binary decision.

It is worth noting that the framework does not rely on a specific choice of the classifiers and can be easily extended to arbitrary binary classifiers and arbitrary class-

conditional distributions respectively. To this end, the generalization error of the various classifiers has to be assessed implicitly, e.g., by resampling techniques such as cross-validation. Moreover, in the case of class-conditional distributions that differ from Gaussian distributions, the Bayes error of misclassification can be derived by Monte Carlo methods.

The effectiveness of this procedure was verified by its successful application to data from the BCI competitions in 2003 and 2005. In both competitions, the disclosure of the true labels of testing data after the competition deadline revealed our approach to be superior to all other competing algorithm.

Synopsis

In this thesis, we have presented novel methods for the analysis of macroscopically recorded brain signals. Here the focus was put on improved feature extraction methods, the single trial detection of mental states, and the analysis of the variability of brain responses.

After outlining the necessary mathematics in chapter 2, the third chapter introduced a novel framework of *conditional event-related (de-)synchronization* (cERD) that permits the analysis of dependencies of the ERD characteristic on external or internal explanatory variables. To this end, we generalized the conventional ERD framework with respect to the reference condition. In particular we substituted the conventionally used static reference value by a dynamic reference. Based on this generalization, we derived two novel measures for the quantification of the *averaged* and *conditional ERD*. A comparative analysis with artificially generated data of known genuine truth exposed the limitations of the conventional ERD framework and exposed the superior abilities of the novel measures with respect to the retrieval of an underlying functional relationship of the ERD characteristic on explanatory variables. Moreover, we investigated the dependencies of μ-rhythm ERD on the basis of data from one subject. Here the analysis of EEG data from a median nerve stimulation paradigm demonstrated the potential of the proposed framework. In particular we applied the novel cERD estimator in order to analyze the μ-rhythm ERD and its dependency on the magnitude of its own prestimulus activity, the magnitude of the occipital α-activity and the magnitude of the ERS in response to the preceding stimulus.

The fourth chapter presented a new feature extraction method termed *Common Spatio-Spectral Pattern* (CSSP) algorithm that extended the well known common-spatial-pattern (CSP) algorithm. In particular we transferred the CSP optimization

problem to the state space by means of the method of time delay embedding. A mathematical analysis of the obtained solution revealed that in addition to the optimization of discriminative spatial filters, CSSP optimizes simple frequency filters. By virtue of these optimal finite impulse response filters CSSP is able to adapt to the individual characteristics of the power spectrum and thus yields an improved feature extraction. The efficiency of the proposed method was demonstrated by its application in an off-line study for the classification of imaginary limb movements on a broad set of BCI experiments. Here we found the CSSP algorithm superior to the standard CSP algorithm in terms of an improved classification accuracy.

In the fifth chapter we developed a method that uses prior information about the phase-locked property of event-related potentials (ERP) in a regularization framework to bias a blind source separation algorithm towards an improved extraction of single-trial phase-locked responses in terms of an increased signal-to-noise ratio. In particular, we suggested a transformation of the data that redirects the focus of source separation methods onto the subspace of the ERP components. The practical benefit with respect to an improved separation of single trial ERP components from the ongoing background activity and extraneous noise was first illustrated on artificially generated data and finally verified in a real-world application of extracting single-trial somatosensory evoked potentials from multichannel EEG-recordings.

The sixth and last chapter introduced a Bayesian classification framework that adaptively combines sequences of features efficiently across time. For the classification task of single trials of unilateral imaginary hand movements, we particularly combined the temporally differently accentuated features of μ-rhythm ERD and movement-related potentials. The effectiveness of this approach was proven by its successful application to data from the international BCI competitions in 2003 and 2005.

Bibliography

[1] T. Allison, G. McCarthy, C. Wood, T. Darcey, D. Spencer, and P. Williamson. Human cortical potentials evoked by stimulation of the median nerve. I. cytoarchitectonic areas generating short-latency activity. *J. Neurophysiol*, 62:694–710, 1989.

[2] T. Allison, G. McCarthy, C. Wood, P. Williamson, and D. Spencer. Human cortical potentials evoked by stimulation of the median nerve. II. cytoarchitectonic areas generating long-latency activity. *J. Neurophysiol*, 62:711–722, 1989.

[3] C. Babiloni, F. Carducci, F. Cincotti, P. M. Rossini, C. Neuper, G. Pfurtscheller, and F. Babiloni. Human movement-related potentials vs desynchronization of EEG alpha rhythm: A high-resolution EEG study. *NeuroImage*, 10:658–665, 1999.

[4] S. Baker, G. Curio, and L. RN. EEG oscillations at 600 hz are macroscopic markers for cortical spike bursts. *J. Physiol.*, 550(Pt 2):529–34, Jul 2003.

[5] A. Belouchrani, K. Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second-order statistics. *IEEE Trans Signal Processing*, 45:434–444, 1997.

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

[7] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398:297–298, 1999.

[8] B. Blankertz, G. Curio, and K.-R. Müller. Classifying single trial EEG: Towards brain computer interfacing. In T. G. Diettrich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Inf. Proc. Systems (NIPS 01)*, volume 14, pages 157–164, 2002.

[9] B. Blankertz, G. Dornhege, S. Lemm, M. Krauledat, G. Curio, and K.-R. Müller. The Berlin Brain-Computer Interface: Machine learning based detection of user specific brain states. *J. Universal Computer Sci.*, 12(6):581–607, 2006.

[10] B. Blankertz, G. Dornhege, C. Schäfer, R. Krepki, J. Kohlmorgen, K.-R. Müller, V. Kunzmann, F. Losch, and G. Curio. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Trans. Neural Sys. Rehab. Eng.*, 11(2):127–131, 2003.

[11] B. Blankertz, K.-R. Müller, T. V. G. Curio, G. Schalk, J. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, and M. S. N. Birbaumer. The BCI competition 2003. *IEEE Trans. Biomed. Eng.*, 51(6):1044–51, 2004.

[12] B. Blankertz, K.-R. Müller, D. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millán, M. Schröder, and N. Birbaumer. The BCI competition III: Validating alternative approachs to actual BCI problems. *IEEE Trans. Neural Sys. Rehab. Eng.*, 14(2):153–159, 2006.

[13] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Processing*, 81(11):2353–62, 2001.

[14] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *IEE Proceedings-F*, 140(6):362–70, 1993.

[15] J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J Math Anal Appl*, 17(1):161–4, 1996.

[16] G. Chatrian, E. Lettich, and P. Nelson. Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity. *Am J EEG Technol*, 25:83–92, 1985.

[17] M. Cheng, X. Gao, S. Gao, and D. Xu. Design and implementation of a brain-computer interface with high transfer rates. *IEEE Trans. Biomed. Eng.*, 49(10):1181–1186, 2002.

[18] P. Comon. Independent component analysis, a new conept. *Signal Processing, Elsevier*, 36(3):287–314, 1994.

[19] M. M. Doppelmayr, W. Klimesch, T. Pachinger, and B. Ripper. The funcional significance of absolute power with respect to event-related desynchronization. *Brain Topogr.*, 11(2):133–140, 1998.

[20] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Combining features for BCI. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Inf. Proc. Systems (NIPS 02)*, volume 15, pages 1115–1122, 2003.

[21] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. *IEEE Trans. Biomed. Eng.*, 51(6):993–1002, June 2004.

[22] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Increase information transfer rates in BCI by CSP extension to multi-class. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 733–740. MIT Press, Cambridge, MA, 2004.

[23] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller. Optimizing spatio-temporal filters for improving brain-computer interfacing. In *Advances in Neural Inf. Proc. Systems (NIPS 05)*, volume 18, pages 315–322, Cambridge, MA, 2006. MIT Press.

[24] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors. *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, MA, 2007.

[25] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2nd edition, 2001.

[26] P. J. Durka, D. Ircha, C. Neuper, and G. Pfurtscheller. Time-frequency microstructure of event-related electro-encephalogram desynchronisation and synchronisation. *Med. Biol. Eng. Comput.*, 39:315–321, 2001.

[27] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley & Sons, Inc., 1950.

[28] J. H. Friedman. Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, 84(405):165–175, 1989.

[29] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1972.

[30] J. Gerber and H. Meinck. The effect of changing stimulus intensities on median nerve somatosensory-evoked potentials. *Electromyogr Clin Neurophysiol*, 40(8):477–82, 2000.

[31] C. Guger, W. Domej, G. Lindner, and E. G. Effects of cable car ascent to 2700 meters on mean EEG frequency and event-related desynchronization (ERD). *Wien Med Wochenschr.*, 155(7–8):143–8, 2003.

[32] R. Hari and R. Salmelin. Human cortical oscillations: a neuromagnetic view through the skull. *Trends in Neuroscience*, 20:44–9, 1997.

[33] L. Hochberg, M. Serruya, G. Friehs, J. Mukand, M. Saleh, A. Caplan, A. Branner, D. Chen, R. Penn, and J. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, July 2006.

[34] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–92, 1997.

[35] J. Intriligator and J. Polich. On the relationship between EEG and ERP variability. *Int. J. Psychophysiol.*, 20:59–74, 1995.

[36] B. H. Jansen and M. E. Brandt. The effect of the phase of prestimulus alpha activity on the averaged viusal evoked response. *Electroencephalogr. Clin. Neurophysiol.*, 80:241–50, 1991.

[37] H. Jasper and W. Penfield. Electrocorticograms in man: Effect of voluntary movement upon the electrical activity of the precentral gyrus. *Arch. Psychiatrie Zeitschrift Neurol.*, 183:163–74, 1949.

[38] T. Jung, S. Makeig, C. Humphries, T. Lee, M. McKeown, V. Iragui, and T. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37:163–78, 2000.

[39] T. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. Sejnowski. Analysis and visualization of single-trial event-related potentials. *Human Brain Mapping*, 14:166–85, 2001.

[40] J. Kalcher and G. Pfurtscheller. Discrimination between phase-locked and non-phase-locked event-related EEG activity. *Electroenceph. Clin. Neurophysiol.*, 94:381–4, 1995.

[41] M. A. Kisley and G. L. Gerstein. Trial-to-trial variability and state-dependent modulation of auditory-evoked responses in cortex. *J. Neurosci.*, 19(23):10451–60, 1999.

[42] Z. J. Koles and A. C. K. Soong. EEG source localization: implementing the spatio-temporal decomposition approach. *Electroencephalogr. Clin. Neurophysiol.*, 107:343–352, 1998.

[43] M. Krauledat, G. Dornhege, B. Blankertz, and K.-R. Müller. Robustifying EEG data analysis by removing outliers. *Chaos and Complexity Letters*, 2(3):259–274, 2007.

[44] R. Krepki. *Brain-Computer Interfaces: Design and Implementation of an Online BCI System of the Control in Gaming Applications and Virtual Limbs*. PhD thesis, Technische Universität Berlin, Fakultät IV – Elektrotechnik und Informatik, 2004.

[45] M. Laubach, J. Wessberg, and M. Nicolelis. Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. *Nature*, 405(6786):523–525, 2000.

[46] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller. Spatio-spectral filters for improving classification of single trial EEG. *IEEE Trans. Biomed. Eng.*, 52(9):1541–1548, 2005.

[47] S. Lemm, G. Curio, Y. Hlushchuk, and K.-R. Müller. Enhancing the signal to noise ratio of ICA-based extracted ERPs. *IEEE Trans. Biomed. Eng.*, 53(4):601–607, April 2006.

[48] S. Lemm, C. Schäfer, and G. Curio. Probabilistic modeling of sensorimotor μ-rhythms for classification of imaginary hand movements. *IEEE Trans. Biomed. Eng.*, 51(6):1077–1080, 2004.

[49] S. Lemm, C. Schäfer, and G. Curio. Aggregating classification accuracy across time: Application to single trial EEG. In *Advances in Neural Inf. Proc. Systems (NIPS 06)*, volume 19, pages 825–832. MIT press, 2007.

[50] E. Leuthardt, G. Schalk, J. Wolpaw, J. Ojemann, and D. Moran. A brain-computer interface using electrocorticographic signals in human. *Journal of Nerual Engineering*, 1(2):63–71, Jun 2004.

[51] S. P. Levine, J. E. Huggins, S. L. BeMent, R. K. Kushwaha, L. A. Schuh, M. M. Rohde, E. A. Passaro, D. A. Ross, K. V. Elsievich, and B. J. Smith. A direct brain interface based on event-related potentials. *IEEE Trans. Rehab. Eng.*, 8(2):180–185, 2000.

[52] S. Makeig, T.-P. Jung, D. Ghahremani, and T. Bell, A.J.and Sejnowski. Blind separation of event-related brain responses into independent components. *Proc. Natl. Acad. Sci. USA*, 94:10979–10984, 1997.

[53] S. Makeig, T.-P. Jung, D. Ghahremani, and T. Sejnowski. Independent component analysis of simulated ERP data. Technical Report INC-9606, Institute for Neural Computation, 1996.

[54] S. Makeig, M. Westerfield, T. Jung, J. Townsend, T. Sejnowski, and E. Courchesne. Functionally independent components of the late positive event-related potential during visual spatial attention. *J Neurosci.*, 19(7):2665–08, 1999.

[55] S. Makeig, M. Westerfield, T.-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295:690–694, 2002.

[56] S. Makeig, M. Westerfield, J. Townsend, T. Jung, E. Courchesne, and T. Sejnowski. Functionally independent components of early event-related potentials in a visual spatial attention task. *Philos Trans R Soc Lond B Biol Sci.*, 354, 1999.

[57] F. Mauguiere, I. Merlet, S. Vanni, V. Jousmaki, P. Adeleine, and R. Hari. Activation of a distributed somatosensory cortical network in the human brain: a dipole modelling study of magnetic fields evoked by median nerve stimulation. Part I: Location and activation timing of SEF sources. *Electroencephalogr Clin Neurophysiol*, 104(4):281–9, 1997.

[58] D. J. McFarland, W. A. Sarnacki, and J. R. Wolpaw. Brain-computer interface (BCI) operation: optimizing information transfer rates. *Biol. Psychol.*, 63:237–251, 2003.

[59] F. C. Meinecke, S. Harmeling, and K.-R. Müller. Robust ICA for super-gaussian sources. In C. G. Puntonet and A. Prieto, editors, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2004)*, 2004.

[60] F. C. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A Resampling Approach to Estimate the Stability of one- or multidimensional Independent Components. *IEEE Trans. Biomed. Eng.*, 49(12):1514–1525, 2002.

[61] B. Mensh, J. Werfer, and H. Seung. Combining gamma-band power with slow cortical potentials to improve single-trial classification of electroencephalographic signals. *IEEE Trans. Biomed. Eng.*, 51(6):1052–6, 2004.

[62] M. Middendorf, G. McMillan, G. Calhoun, and K. S. Jones. Brain-computer interface based on the steady-state visual-evoked response. *IEEE Trans. Rehab. Eng.*, 8(2):211–214, June 2000.

[63] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K.-R. Müller. Invariant feature extraction and classification in kernel spaces. In S. Solla, T. Leen, and K.-R. Müller, editors, *Proc. NIPS 12*, pages 526–532. MIT Press, 2000.

[64] J. Millán. On the need for on-line learning in brain-computer interfaces. In *Proceedings of the International Joint Conference on Neural Networks*, Budapest, Hungary, July 2004. IDIAP-RR 03-30.

[65] K.-R. Müller, C. W. Anderson, and G. E. Birch. Linear and non-linear methods for brain-computer interfaces. *IEEE Trans. Neural Sys. Rehab. Eng.*, 11(2):165–169, 2003.

[66] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Neural Networks*, 12(2):181–201, May 2001.

[67] K.-R. Müller, R. Vigario, F. C. Meinecke, and A. Ziehe. Blind source separation techniques for decomposing event-related brain signals. *International Journal of Bifurcation and Chaos*, 14(2):773–791, 2004.

[68] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clin. Neurophysiol.*, 110(5):787–798, 1999.

[69] E. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 10:186–90, 1964.

[70] C. Neuper, A. Schlögl, and G. Pfurtscheller. Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery. *Journal Clin. Neurophysiol.*, 16:373–82, 1999.

[71] E. Niedermeyer, A. Goldszmidt, and D. Ryan. "Mu rhythm status" and clinical correlates. *Clin. EEG Neurosci.*, 35(2):84–87, Apr 2004.

[72] V. Nikouline, K. Linkenkaer-Hansen, W. H., M. Kesäniemi, E. Antonova, R. Ilmoniemi, and J. Huttunen. Dynamics of mu-rhythm suppression caused by median nerve stimulation: a magnetoencephalographic study in human subjects. *Neurosci. Lett.*, 294(3):163–166, 2000.

[73] V. K. Nikouline, H. Wikström, K. Linkenkaer-Hansen, M. Kesäniemi, R. Ilmoniemi, and J. Huttunen. Somatosensory evoked magnetic fields: relation to pre-stimulus mu rhythm. *Clin. Neurophysiol.*, 111:1227–33, 2000.

[74] L. Parra, C. Alvino, A. C. Tang, B. A. Pearlmutter, N. Yeung, A. Osman, and P. Sajda. Linear spatial integration for single trial detection in encephalography. *NeuroImage*, 7(1):223–230, 2002.

[75] W. D. Penny, S. J. Roberts, E. A. Curran, and M. J. Stokes. EEG-based communication: A pattern recognition approach. *IEEE Trans. Rehab. Eng.*, 8(2):214–215, June 2000.

[76] B. O. Peters, G. Pfurtscheller, and H. Flyvbjerg. Automatic differentiation of multichannel EEG signals. *IEEE Trans. Biomed. Eng.*, 48(1):111–116, 2001.

[77] G. Pfurtscheller and A. Aranbibar. Evaluation of event-related desynchronization preceding and following voluntary self-paced movement. *Electroencephalogr. Clin. Neurophysiol.*, 46(2):138–46, 1979.

[78] G. Pfurtscheller and F. H. L. da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.*, 110(11):1842–1857, Nov 1999.

[79] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer. EEG-based discrimination between imagination of right and left hand movement. *Electroenceph. clin. Neurophysiol.*, 103:642–51, 1997.

[80] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, R. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer. Current trends in Graz brain-computer interface (BCI). *IEEE Trans. Rehab. Eng.*, 8(2):216–219, June 2000.

[81] G. Pfurtscheller, A. Stancak Jr., and C. Neuper. Event-related synchronization (ERS) in the alpha band–An electrophysiological correlate of cortical idling: a review. *Int. J. Psychophysiol.*, 1996.

[82] J. A. Pineada. The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing". *Brain Res. Brain Res. Rev.*, 50(1):57–68, Dec 2005.

[83] R. Quiroga and H. Garcia. Single-trial event-related potentials with wavelet denoising. *Clinical Neurophysiology*, 114:376–90, 2003.

[84] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehab. Eng.*, 8(4):441–446, 2000.

[85] S. Salenius, A. Schnitzler, R. Salmelin, V. Jousmäki, and R. Hari. Modulation of human cortical rolandic rhythms during natural sensorimotor tasks. *NeuroImage*, 5:221–8, 1997.

[86] A. Schlögl. `http://hci.tugraz.at/schloegl/publications/TR_BCI2005_III.pdf`.

[87] A. Schlögl, R. S. C. Keinrath, and G. Pfurtscheller. Information transfer of an EEG-based brain-computer interface. In *Proc. First Int. IEEE EMBS Conference on Neural Engineering*, pages 641–644, 2003.

[88] A. Schnitzler, S. Salenius, R. Salmelin, V. Jousmäki, and R. Hari. Involvement of primary motor cortex in motor imagery: a neuromagnetic study. *NeuroImage*, 6(3):201–208, 1997.

[89] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York, 1992.

[90] P. Shenoy, M. Krauledat, B. Blankertz, R. P. N. Rao, and K.-R. Müller. Towards adaptive classification for BCI. *J. Neural Eng.*, 3(1):R13–R23, 2006.

[91] A. Sornborger, T. Yokoo, A. Delorme, C. Sailstad, and L. Sirovich. Extraction of the average and differential dynamical response in stimulus-locked experimental data. *Journal of Neuroscience Methods*, 141(22):223–229, 2005.

[92] A. Stancák, J. Svoboda, R. Rachmanová, J. Vrána, J. Králík, and T. J. Desynchronization of cortical rhythms following cutaneous stimulation: effects of stimulus repetition and intensity, and of the size of corpus callosum. *Clin. Neurophysiol.*, 114(10):1936–47, 2003.

[93] F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L. S. Young, editors, *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, 1981.

[94] D. Taylor, S. Tillery, and A. Schwartz. Direct cortical control of 3d neuroprosthetic devices. *Science*, 5574(296):1829–32, Jun 2002.

[95] R. Tomioka, G. Dornhege, G. Nolte, B. Blankertz, K. Aihara, and K.-R. Müller. Spectrally weighted common spatial pattern algorithm for single trial EEG classification. Technical Report 40, Dept. of Mathematical Engineering, The University of Tokyo, July 2006.

[96] C. Torrence and G. Compo. A practical guide to wavelet analysis. *Bull. Am. Meterol.*, 79:61–78, 1998.

[97] J. J. Vidal. Toward direct brain-computer communication. *Annu. Rev. Biophys.*, 2:157–180, 1973.

[98] R. Vigario. Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalography and clinical Neurophysiology*, 103:395–404, 1997.

[99] R. Vigário, J.Särelä, V. Jousmäki, M. Hämäläninе, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Trans Biomed Eng*, 47(5):589–93, 2000.

[100] R. Vigário, J.Särelä, and E. Oja. Independent component analysis in wave decomposition of auditory evoked fields. In *Proc. of 8th Int. Conf. on Artificial Neural Networks*, pages 287–92. Springer, 1998.

[101] G. Watson. Smooth regression analysis. *Shankya Series A*, 26:359–72, 1964.

[102] M. Woertz, G. Pfurtscheller, and W. Klimesch. Alpha power dependent light stimulation: dynamics of event-related (de)synchronization in human electroencephalogram. *Brain Res. Cog. Brain Res.*, 20(2):256–60, 2004.

[103] J. Wolpaw and D. McFarland. Multichannel EEG-based brain-computer communication. *Electroenceph. clin. Neurophysiol.*, 90:444–9, 1994.

[104] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clin. Neurophysiol.*, 113(6):767–791, 2002.

[105] J. R. Wolpaw and D. J. McFarland. Multichannel EEG-based brain-computer communication. *Electroencephalogr. Clin. Neurophysiol.*, 90:444–449, 1994.

[106] G. Wübbeler, A. Ziehe, B.-M. Mackert, K.-R. Müller, L. Trahms, and G. Curio. Independent component analysis of non-invasively recorded cortical magnetic DC-fields in humans. *IEEE Transactions on Biomedical Engineering*, 47(5):594–599, 2000.

[107] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source. *IEEE Trans on Sig Proc*, 50, 2002.

[108] M. Zibulevsky and B. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4), 2001.

[109] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, 5:777–800, Jul 2004.

[110] A. Ziehe and K.-R. Müller. TDSEP – an efficient algorithm for blind separation using time structure. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN'98*, Perspectives in Neural Computing, pages 675 – 680, Berlin, 1998. Springer Verlag.

[111] A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Trans Biomed Eng*, 47(1):75–87, January 2000.