ON ROBUST SPATIAL FILTERING OF EEG IN NONSTATIONARY ENVIRONMENTS

WOJCIECH SAMEK

TU berlin

GRK
Sensory Computation
in Neural Systems

bccn
berlin

Bundesministerium
für Wirtschaft
und Technologie

BBCI

ON ROBUST SPATIAL FILTERING OF EEG IN NONSTATIONARY ENVIRONMENTS

vorgelegt von Dipl. Inf. Wojciech Samek geb. Wójcikiewicz

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation

Promotionsausschuss:
Vorsitzender: Prof. Dr. Benjamin Blankertz (Technische Universität Berlin)
1. Gutachter: Prof. Dr. Klaus-Robert Müller (Technische Universität Berlin)
2. Gutachter: Prof. Dr. Cuntai Guan (Institute for Infocomm Research)
3. Gutachter: Dr. Fabien Lotte (INRIA Bordeaux Sud-Ouest)

Tag der wissenschaftlichen Aussprache: 20. Juni 2014

Berlin, Deutschland 2014
D83

To my parents

ACKNOWLEDGMENTS

I am very grateful for being given the opportunity to pursue my PhD in such an inspiring environment as the Machine Learning Group of Prof. Dr. Klaus-Robert Müller. It is a great privilege to be part of a group with such a long history of cutting-edge research and to work together with so many bright minds. I would like to thank Prof. Müller for being my "doctor father" in every meaning of this term. He created an environment where I could grow scientifically, he supported me during dry periods of research and taught me to think outside the box. Above all he infected me with his enthusiasm and, with his empathy and humanity, he became a role model for being professor and leader of a research group. Thank you for all your support.

I am also deeply grateful to Dr. Motoaki Kawanabe who was my supervisor in the first two years of my PhD and now is a close collaborator. Without the support of Dr. Kawanabe this dissertation would not have been possible. He taught me the basics of machine learning, statistics and information geometry and was always available when I had questions. His dedicated help was also indispensable when writing my first papers and pursuing my first research ideas. Thank you for the generosity in sharing your knowledge with me.

I would also like to thank Dr. Carmen Vidaurre who introduced me to Brain-Computer Interfacing. Thank you very much for your patience in explaining the methodological and experimental challenges of this field more than once to me. I am aware that without your support it would be much harder for me to learn about these things. This thesis also highly benefited from scientific discussions with Duncan Blythe, Dr. Alexander Binder and Dr. Frank Meinecke. Interacting with you was very stimulating for my research and I could learn a great deal about mathematics from you. A special thanks also goes to Andrea Gerdes and Vanessa Cassagrande who supported me in all administrative matters. I am particularly indebted to the GRK "Sensory Computation in Neural Systems". Being part of this interdisciplinary graduate school was a great experience and I profited from it greatly through various soft skill courses and fruitful discussions with other PhD students. I am very thankful to Prof. Dr. Cuntai Guan and Dr. Fabien Lotte for their time and efforts in reading and evaluating this thesis and to all people who proof-read this dissertation.

I owe my deepest gratitude to my parents for their support over the years. I am particularly indebted to my father who helped me to develop first algorithmic skills by teaching me chess and the card game Bridge when I was a small child. The joy of strategic thinking and problem solving was ignited at this age but has accompanied me through all my life and influenced my decision to study Computer Science and to pursuit a PhD in Machine Learning. I am very happy about this development.

Finally, I would like to thank my wife for giving me everything that research can not provide.

ABSTRACT

Brain-Computer Interfaces (BCIs) provide a novel communication channel between a human subject and a computer application that does not rely on peripheral nerves and muscles. The core idea of a BCI is (1) to encode information by voluntarily inducing certain mental states, (2) to decode these states from recordings of brain activity and (3) to use this information for controlling a device or communicating with the environment. Reliable decoding of mental states is a very challenging task as the recorded brain signals, e.g., EEG, may be noisy, nonstationary and affected by artifacts. Furthermore these signals are usually of high dimension and for the most part contain background activity that is not related to the encoded mental state. Thus, robust extraction of informative features is essential for successful BCI application, especially when decoding mental states on a single trial basis. In motor imagery BCIs spatial filtering is a crucial step in this feature extraction process. Although many spatial filtering methods have been proposed, of which the most prominent representative is Common Spatial Patterns (CSP), these methods rarely take into account explicitly the influence of artifacts and nonstationarity. The main goal of this doctoral thesis is to refine current approaches and to develop novel methods for robust spatial filtering of EEG in nonstationary environments.

This dissertation contributes to the current state of research in several ways. The first contribution is the development of a novel regularization strategy for CSP that reduces variability of the extracted features. An algorithm, *stationary CSP*, is proposed and its regularization effect is investigated using simulations and real data sets. The second contribution is the investigation of the possibility to transfer information about changes in the data between subjects performing the same experiment. Also here a method, *stationary subspace CSP*, is developed and extensively evaluated. A key contribution of this thesis is the formulation of spatial filter computation as a divergence maximization problem. This novel view on spatial filter computation and in particular on CSP has several advantages, e.g., it easily allows one to robustify the algorithm against artifacts, to incorporate data from other subjects into the optimization process and to enforce different types of invariances on the extracted features. Conceptually, this generic formulation integrates many state-of-the-art CSP variants in a principled manner and provides an information geometric interpretation of the algorithm; this formulation also opens the door for new variants by utilizing the properties of particular divergences. The advantages and limitations of different divergence-based spatial filtering algorithms are discussed and evaluated using simulations and real EEG recordings. The fourth main contribution of this work is the derivation of a novel robust covariance estimator which takes into account trial structure and is tailored to BCI application.

All proposed algorithms are compared to several state-of-the-art approaches and the results are interpreted from a neurophysiological perspective. Future research directions are discussed at the end of this thesis.

ZUSAMMENFASSUNG

Brain-Computer Interfaces (BCIs) sind neuartige technische Systeme, die einen Kommunikationskanal zwischen Mensch und Computer zur Verfügung stellen, ohne dabei das periphere Nervensystem oder die Muskulatur einzubeziehen. Das Arbeitsprinzip eines BCIs basiert auf drei Schritten. (1) Die Versuchsperson kodiert die zu übermittelnde Anweisung als mentalen Zustand, indem sie sich z.B. das Ausführen einer speziellen Bewegung vorstellt. (2) Das BCI System dekodiert den mentalen Zustand und damit auch die zu übermittelnde Anweisung aus gemessener Hirnaktivität. (3) Das System setzt die Anweisung zur Steuerung von technischen Geräten oder zur Kommunikation mit der Umgebung um. Das zuverlässige Dekodieren der mentalen Zustände ist jedoch schwierig, da die gemessenen elektroenzephalographischen Signale sowohl verrauscht und nichtstationär sind als auch Artefakte enthalten können. Weiterhin sind diese Signale üblicherweise hochdimensional und bestehen zum größten Teil aus neuronaler Hintergrundaktivität, die keinen Bezug zu dem induzierten mentalen Zustand aufweist. Aus diesem Grund ist die Extraktion von robusten und informativen Merkmalen von zentraler Bedeutung für eine erfolgreiche Anwendung der BCI Technologie, vor allem wenn das System Einzelversuche klassifizieren können soll. Ein wesentlicher Schritt bei der Merkmalsextraktion in sogenannten motor-imagery BCI Systemen ist die räumliche Filterung. Obwohl viele Methoden zur Berechnung von räumlichen Filtern in der Literatur vorgeschlagen wurden – zu nennen ist hier vor allem der Common Spatial Patterns (CSP) Algorithmus – wurde dem Nichtstationaritäts- und dem Robustheitsproblem selten ausreichend Beachtung geschenkt. Das Ziel der vorliegenden Arbeit ist die Entwicklung von neuen Methoden für die robuste räumliche Filterung von nichtstationären EEG Signalen.

Diese Dissertation trägt in mehrerer Hinsicht zu der aktuellen Forschung im Bereich BCI bei. Der erste Beitrag dieser Arbeit ist die Entwicklung einer neuen Regularisierungsstrategie für CSP, welche die Varianz der extrahierten Merkmale reduziert. Hierfür wird ein neuer Algorithmus – stationary CSP – vorgeschlagen und die Auswirkungen der Regularisierung werden mit Hilfe von Simulationen und echten Daten erforscht. Ein weiterer Beitrag dieser Arbeit besteht in der Untersuchung von Möglichkeiten für den Austausch von Nichtstationaritätsinformationen zwischen Versuchspersonen, die am gleichen BCI-Experiment teilnehmen. In diesem Zusammenhang wird eine neue Methode – stationary subspace CSP – entwickelt und umfassend evaluiert. Der zentrale Beitrag der Dissertation ist die Formulierung der Berechnung von räumlichen Filtern als Divergenzmaximierungsproblem. Diese neue Betrachtungsweise der Filterberechnung hat verschiedene Vorteile und erlaubt es z.B. eine gegenüber Artefakten robuste Variante des CSP Algorithmus zu verfassen, Daten von zusätzlichen Versuchspersonen in den Optimierungsprozess aufzunehmen und verschiedene Arten von Stationarität zu forcieren. Die divergenzbasierte Formulierung der Filterberechnung umfasst viele CSP Varianten

und erlaubt eine informationsgeometrische Interpretation dieser Algorithmen. Auch eröffnet diese Formulierung Möglichkeiten neuartige Varianten des Algorithmus unter Ausnutzung der speziellen Eigenschaften anderer Divergenzen zu entwickeln. Die Vorteile und Grenzen der verschiedenen divergenzbasierten Methoden zur Berechnung der räumlichen Filter werden in dieser Arbeit diskutiert und mit Hilfe von Simulationen und echten EEG Aufnahmen evaluiert. Der vierte wesentliche Beitrag dieser Arbeit liegt in der Herleitung eines neuen und robusten Schätzers für Kovarianzmatrizen, welcher speziell auf die Struktur von BCI Daten abgestimmt ist.

Alle vorgeschlagenen Methoden werden mit diversen, dem Stand der Technik entsprechenden Verfahren verglichen und die Ergebnisse aus neurophysiologischer Perspektive interpretiert. Weiterführende Forschungsmöglichkeiten werden am Ende der Arbeit diskutiert.

CONTENTS

1 INTRODUCTION

Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions which we know as disease.
(Sir William Osler, 1903)

The Canadian physician Sir William Osler (1849 - 1919) revolutionized the medical world of the 19th century by advocating an approach to therapy which focuses on the needs of individual patients. He recognized the great variability among individuals with respect to their physical conditions, their mental status and their responses to drugs and concluded that focusing solely on clinical signs and symptoms does not result in the most effective therapy. The development of this novel patient-centered paradigm made him the father of personalized medicine; a branch of medicine that is also very popular today due to a better understanding of the genetic influences on a person's risk of acquiring certain diseases.

In the last two decades personalized approaches have also revolutionized the field of Brain-Computer Interfacing (BCI). A BCI system (see e.g. Wolpaw et al., 2002; Dornhege et al., 2007; Wolpaw and Wolpaw, 2012) aims to decode the intention of a user from recordings of brain activity and to use this information for controlling a computer application or a robotic device. Today's BCIs extract user-specific features from electroencelographic (EEG) recordings and adapt to the user's signal characteristics. This *subject-centered* or *machine learning* approach not only largely reduces the calibration time in comparison to classical BCI systems based on neurofeedback training (Kamiya et al., 1969) but also increases classification accuracy. Unfortunately, EEG responses to a stimulus or a task not only differ from subject to subject but also from trial to trial and from day to day. This change in the signal properties over time is termed the intrinsic *nonstationarity* of EEG (Kaplan et al., 2005) and constitutes a major challenge for data analysis and classification by violating a basic assumption of many machine learning methods, namely that data are sampled from a fixed (but unknown) distribution (Vapnik, 1998; Hastie et al., 2001). The analysis of EEG is further aggravated by the presence of artifacts in the data, e.g., eye movements, muscular activity or loose electrodes. The lack of robustness to unexpected events and the nonstationary nature of EEG are major reasons in explaining why current BCI technology is seldom used in out-of-lab scenarios and clinical practice. Since artifacts and nonstationarity can not be fully eliminated, even with the best experimental protocol, robust and invariant feature representations are required for optimal signal analysis and high classification accuracy.

1.1 STRUCTURE OF THE THESIS

This thesis is divided into three parts. The first part introduces the basic concepts of motor imagery-based Brain-Computer Interfacing and discusses the advantages and limitations of several state-of-the-art spatial filtering algorithms and other related work. The second part proposes two novel methods for the minimization of nonstationarity in the feature extraction process. The first method minimizes feature variability within an experimental session, the second utilizes data from other users to minimize the shift in feature distribution between sessions. A generic divergence-based framework for spatial filter computation and a novel robust covariance matrix estimator are introduced and evaluated in the last part of this thesis. The individual chapters are summarized as follows.

Chapter 2 introduces basic concepts of motor imagery-based Brain-Computer Interfacing and the data sets used for evaluation of the proposed methods.

Chapter 3 describes the Common Spatial Patterns algorithm, discusses the robustness and nonstationarity problem in BCI and reviews several state-of-the-art spatial filtering methods.

Chapter 4 introduces the idea of regularizing CSP towards stationarity, proposes a new algorithm, stationary CSP, and evaluates its performance.

Chapter 5 introduces the idea of transferring information about changes in the data between subjects, proposes a new algorithm, stationary subspace CSP, and evaluates its performance.

Chapter 6 derives a novel robust estimator for covariance matrices by using concepts from information geometry.

Chapter 7 proposes a divergence-based framework for spatial filter computation. This framework unifies many state-of-the-art CSP algorithms and enables us to develop novel robust and stationary CSP variants in a principled manner. Several novel CSP variants are presented and evaluated.

Chapter 8 concludes the thesis with a summary and an outlook on future work.

1.2 OWN CONTRIBUTIONS

The following contributions are made to the current state of research.

Conceptual Contributions

1. The regularization of spatial filter computation towards stationarity [4, 12, 13, 16, 18].
2. The transfer of information about changes between subjects [3, 14, 15].
3. The formulation of spatial filter computation as divergence maximization problem [1, 6, 7].

Theoretical Contributions

1. A theorem relating the CSP algorithm to symmetric Kullback-Leibler divergence maximization is developed and proved [1, 6, 7].
2. An optimization algorithm for finding a subspace with maximum sum of Kullback-Leibler and beta divergences is proposed [1, 6, 7].
3. A novel robust estimator for covariance matrices based on beta divergence minimization is derived [5].

Methodological Contributions

1. The stationary CSP algorithm is proposed [4, 12, 13, 16, 18].
2. The stationary subspace CSP algorithm is proposed [3, 14, 15].
3. The divergence-based CSP framework is developed [1, 6].
4. A robust beta divergence CSP method is proposed [7].

Contributions not included in this thesis

1. Max-Min CSP method which applies the max-min theorem to robustly compute spatial filters [2].
2. MKL + CSP method which applies Multiple Kernel Learning to optimally incorporate information from additional subjects [8].
3. GroupSSA + CSP method which projects data to a stationary subspace prior to spatial filter computation [9, 10, 19].
4. Generative model for the Stationary Subspace Analysis (SSA) algorithm [11].

1.3 LIST OF PUBLICATIONS

The following list contains all contributions made by the author to the field of robust spatial filtering in BCI. As common practice in this scientific field some of the materials presented in this thesis have been prepublished as Journal articles or presented at scientific conferences. The work presented in [1], [3], [4], [5], [6] and [7] is included in large parts into this thesis.

Journal articles

[1] Samek, W., Kawanabe, M., Müller, K.-R. *Divergence-based Framework for Common Spatial Patterns Algorithms*. **IEEE Reviews in Biomedical Engineering**, 7:50-72, 2014.

[2] Kawanabe, M., Samek, W., Müller, K.-R., Vidaurre, C. *Robust Common Spatial Filters with a Maxmin Approach*. **Neural Computation**, 26(2):1-26, 2014.

[3] Samek, W., Meinecke, F. C., Müller, K.-R. *Transferring Subspaces Between Subjects in Brain-Computer Interfacing*. **IEEE Transactions on Biomedical Engineering**, 60(8):2289–2298, 2013.

[4] Samek, W., Vidaurre, C., Müller, K.-R., Kawanabe, M. *Stationary Common Spatial Patterns for Brain-Computer Interfacing*. **Journal of Neural Engineering**, 9:026013, 2012.

Peer-reviewed contributions to conferences

[5] Samek, W., Kawanabe, M. *Robust Common Spatial Patterns by Minimum Divergence Covariance Estimator*. **Proceedings of 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2059-2062, 2014.

[6] Samek, W., Müller, K.-R. *Information Geometry meets BCI – Spatial Filtering using Divergences*. **Proceedings of 2nd IEEE International Winter Workshop on Brain-Computer Interface**, 1-4, 2014.

[7] Samek, W., Blythe, D., Müller, K.-R., Kawanabe, M. *Robust Spatial Filtering with Beta Divergence*. **Advances of Neural Information Processing 26 (NIPS)**, 1007–1015, 2013.

[8] Samek, W., Binder, A., Müller, K.-R. *Multiple Kernel Learning for Brain-Computer Interfacing*. **Proceedings of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)**, 7048–7051, 2013.

[9] Samek, W., Müller, K.-R., Kawanabe, M., Vidaurre, C. *Brain-Computer Interfacing in Discriminative and Stationary Subspaces*. **Proceedings of 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)**, 2873–2876, 2012.

[10] Samek, W., Kawanabe, M., Vidaurre, C. *Group-wise Stationary Subspace Analysis – A Novel Method for Studying Non-Stationarities*. **Proceedings of 5th International Brain-Computer Interface Conference**, 16-20, Verlag der TU Graz, 2011.

[11] Kawanabe, M., Samek, W., von Bünau, P., Meinecke, F. C. *An Information Geometrical View of Stationary Subspace Analysis*. **Artificial Neural Networks and Machine Learning – ICANN 2011**, **Lecture Notes in Computer Science**, 6792/2011:397–404, Springer-Verlag, 2011.

[12] Wojcikiewicz, W., Vidaurre, C., Kawanabe, M. *Improving Classification Performance of BCIs by using Stationary Common Spatial Patterns and Unsupervised Bias Adaptation*. **Hybrid Artificial Intelligent Systems**, **Lecture Notes in Computer

Science, 6679/2011:34–41, Springer-Verlag, 2011.

[13] Wojcikiewicz, W., Vidaurre, C., Kawanabe, M. *Stationary Common Spatial Patterns: Towards Robust Classification of Non-Stationary EEG Signals.* **Proceedings of 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 577–580, 2011.

Conference Abstracts

[14] Samek, W., Meinecke, F. C., Müller, K.-R. *Learning From Other Subjects for Robust Brain-Computer Interfacing.* **IPAM Multimodal Neuroimaging Workshop**, 2013.

[15] Samek, W., Müller, K.-R., Meinecke, F. C. *Learning Prominent Changes From Other Subjects for Robust Brain-Computer Interfacing.* **Berlin BCI Workshop 2012 - Advances in Neurotechnology**, 2012.

[16] Wojcikiewicz, W. *Single-Trial Classification of EEG data in Non-Stationary Environments.* **Workshop of the German Research Training Groups in Computer Science**, 243, GITO Verlag, 2011.

[17] Blythe, D., Samek, W., Müller, K.-R. *Stationary Linear Discriminant Analysis - Classifying Non-Stationary Features in Brain-Computer Interfacing.* **BC11 : Computational Neuroscience & Neurotechnology Bernstein Conference & Neurex Annual Meeting**, Frontiers in Neuroscience, 2011.

[18] Wojcikiewicz, W., Vidaurre, C., Kawanabe, M. *Stationary Common Spatial Patterns for Non-Stationary EEG Data.* **Bernstein Conference on Computational Neuroscience**, Frontiers in Neuroscience, 2010.

[19] Wojcikiewicz, W., von Bünau, P., Vidaurre, C. *Identifying Non-Stationarities in BCI Data.* **1st Summer School of the International Max Planck Research School on Neuroscience of Communication**, 2010.

Part I

REVISITING BCI

2 BRAIN-COMPUTER INTERFACING

On July 6, 1924, the German neurologist Hans Berger (1873 - 1941) measured the electrical activity of the human brain for the first time (Berger, 1930). With his experiments the field of human electroencephalography (EEG) was born (Collura, 1993). Soon EEG became an invaluable tool for both research and clinical diagnosis of brain diseases. Its advantages are relatively low costs, a high temporal resolution and portability due to the small size of the recording device. Today's clinical applications (see Binnie, 1995) range from coma monitoring (Fischer et al., 1999) to research on diseases such as schizophrenia (Ford, 1999), dyslexia (Baldeweg et al., 1999), ADHD (Barry et al., 2003) and dementia (Patterson et al., 1988).

However, the analysis of EEG not only attracted medical doctors but also engineers and scientists became interested, especially after the processing power of computers had significantly increased. In 1973 Jacques J. Vidal (Vidal, 1973), a professor at the University of California, Los Angeles (UCLA), had the visionary idea to use EEG

> *. . . as [a] carrier of information in man-computer communication or for the purpose of controlling such external apparatus as prosthetic devices or spaceships.*
>
> (Jacques J. Vidal, 1973)

This idea led to the development of a new research field, *Brain-Computer-Interfacing* (BCI). The first BCI systems were based on "neurofeedback" (Kamiya et al., 1969) and required weeks of training. The difficulty was that the user had to learn how to control his/her brain activity; the machine did not adapt to the user's neural signal. Today 90 years after Berger's first experiments EEG-based BCI systems have been successfully used for various applications (see Dunne et al., 2013), e.g., communication (Farwell and Donchin, 1988; Birbaumer et al., 1999; Nijboer et al., 2008; Treder and Blankertz, 2010), neurological rehabilitation (Birbaumer and Cohen, 2007; Daly and Wolpaw, 2008; Kaiser et al., 2011; Ang et al., 2011; Lim et al., 2012; Ang and Guan, 2013; Courtine et al., 2013), wheelchair control (Leeb et al., 2007; Galán et al., 2008; Rebsamen et al., 2010; del R. Millán et al., 2010; Carlson and del R. Millán, 2013), prosthesis control (Guger et al., 1999; Müller-Putz et al., 2005; Jackson et al., 2006; McFarland and Wolpaw, 2008; Hahne et al., 2012), game playing (Lalor et al., 2005; Nijholt and Tan, 2007; Tangermann et al., 2008; Nijholt et al., 2009; Lotte, 2011; Bonnet et al., 2013) and many additional non-medical applications (Müller et al., 2008; del R. Millán et al., 2009; Blankertz et al., 2010b; Haufe et al., 2011; van Erp et al., 2012; Porbadnigk et al., 2013).

A factor boosting BCI research in the recent decades was the development of novel machine learning algorithms which extract relevant, user-specific information from the recorded data. The application of these methods shifted the workload from the user to the machine. This change in paradigm from "*let the user learn*" to "*let the machine learn*" largely reduced training times and significantly increased the attraction of these systems (Blankertz et al., 2006a). Although technology has rapidly advanced during the last decade, e.g., dry electrode EEG recordings (Popescu et al., 2007), zero training systems (Krauledat et al., 2008; Fazli et al., 2009) and robust machine learning methods (Lotte and Guan, 2011; Samek et al., 2014), many challenges limiting a large scale application of BCIs in clinical practice and its usage as assistive technology for disabled people still exists (Dietrich et al., 2010; Krusienski et al., 2011; Lance et al., 2012). Several pilot studies (Ang et al., 2011; Lim et al., 2012) have demonstrated the utility of BCI for medical application, but much more research is needed in this direction.

2.1 NEUROPHYSIOLOGICAL BACKGROUND

The human brain contains around 86 billion neurons (Herculano-Houzel, 2009). These highly specialized cells form networks and are responsible for information processing. A typical neuron consists of the *soma*, *dendrites*, an *axon* and *axon terminals*. Neurons are electrically excitable and have a resting potential of approximate -70 mV across the cell membrane (Blum and Rutkove, 2007). The membrane contains voltage-gated ion channels that control the exchange of ions with the extracellular milieu. If the membrane potential increases to around -55 mV, the sodium ion channels open and sodium ions flow into the cell. This process produces a rapid rise and fall in the membrane potential, also known as *action potential*. The action potential propagates along the axon and may trigger action potentials in neighboring neurons. The flow of positively or negatively charged ions into the post- or presynaptic cell results in excitatory or inhibitory postsynaptic potentials (EPSP or IPSP) and induces an extracellular current in the opposing direction (Blum and Rutkove, 2007). This current accumulates temporally and spatially and is often strong enough to admit measurement on the scalp (Nunez, 2006). The potential differences (caused by the EPSP and IPSP currents) between two scalp locations constitute a time series which is known as the *electroencephalographic signal* or EEG signal. Vertically oriented cortical pyramidal neurons are the principal contributors to EEG.

The EEG signal often shows a characteristic $1/f$ or pink noise power spectrum with multiple peaks in specific frequency ranges. These peaks represent prominent oscillatory activity, i.e., synchrony in a large neuronal population. Traditionally, the EEG power spectrum is divided into multiple frequency bands representing specific EEG rhythms which are denoted by Greek letters. Although these rhythms have been in the focus of neuroscience research for many years, their origins and functions are often not fully understood or are subject to debate (Sterman, 1996; Schürmann and Başar, 2001; Başar et al., 2001). Table 1 gives a basic overview over the most prominent rhythms, their location on the scalp and their characteristics; note that the list is not intended to be exhaustive.

Table 1.: Overview over most prominent EEG rhythms.

Name	Band	Location	Characteristic
δ	$0.1 - 4$	frontal or posterior regions	infants, children, sleep
θ	$4 - 8$	various locations	children, sleep, drowsiness
α	$8 - 13$	posterior regions, occipital and temporal cortex	closed eyes, relaxed mental state
μ	$8 - 13$	sensorimotor cortex	attenuated by movement, motor imagery or tactile stimulation
β	$13 - 30$	frontal regions, somatosensory cortex	active concentration, attenuated by motor activity
γ	> 30	various locations	conscious attention, cognitive processes

Brain-Computer Interface (BCI) systems (see e.g. Birbaumer et al., 1999; Kübler et al., 2001; Wolpaw et al., 2002; Dornhege et al., 2007; del R. Millán et al., 2010; Lemm et al., 2011) aim to control a computer application such as a visual speller or a neuroprosthesis by decoding the intention of a subject from his/her recorded brain signals, e.g., by multi-electrode EEG. Communication becomes possible if the system is able to differentiate between at least two brain states. One popular paradigm for voluntarily inducing different brain states is motor imagery, i.e., the imagining of the execution of a movement with a particular limb such as the right or left hand or the feet. For motor imagery BCIs the μ- and β-rhythms are of outstanding importance. The μ-rhythm is typically visible over the sensorimotor cortex in absence of motor activity. Although the μ- and α-rhythms occur in the same frequency range, namely 8-13 Hz, they are thought to have different origins (Van Leeuwen et al., 1978). The key property which makes the μ-rhythm useful for BCI is its suppression in the sensorimotor cortex contralateral to the limb performing a movement (Jasper and Penfield, 1949; Pfurtscheller and Aranibar, 1979) or motor imagination (Schnitzler et al., 1997; Pineda, 2005). This phenomenon is also known as *event-related desynchronization* (ERD) (Pfurtscheller and Lopes da Silva, 1999) and can be detected as power attenuation in the mu frequency band over the cortical locations engaged in imagination, planning and execution of the movement. According to (Palva and Palva, 2007) EEG rhythms emerge as synchronous idling of a large population of neurons, thus, ERD may be interpreted as a decrease in the number of idling neurons caused by the task engagement.

Event-related desynchronization in the contralateral sensorimotor cortex has also been observed in the beta frequency range during movement execution, planning and imagination (Pfurtscheller and Neuper, 1997; Pfurtscheller and Lopes da Silva, 1999). This β-rhythm is referred to as the μ-beta or Rolandic β-rhythm; the μ-rhythm in the 8-13 Hz range is often denoted as μ-alpha or Rolandic α-rhythm. These idling oscillations over the sensorimotor cortex are often termed *sensorimotor rhythms* (SMRs). A significant increase of oscillations in the beta band can be observed after the end of a movement and after motor imagery. This *event-related synchronization* (ERS) is termed *beta rebound* (Salmelin and Hari, 1994; Pfurtscheller et al., 1996, 2005) and constitutes an additional source of information which may be used for classification purposes in Brain-Computer Interfacing. Neurophysiologically, beta rebound may be interpreted as active inhibition of motor cortical neurons after movement execution or imagination (Pfurtscheller and Neuper, 1997). Since synchronization not only manifests itself as beta rebound but is a more common phenomenon which may occur ipsilaterally during movement execution or imagination or may surround a focal desynchronization (Pfurtscheller and Lopes da Silva, 1999; Suffczynski et al., 1999), one often refers to the *ERD/ERS effect* in Brain-Computer Interfacing.

EEG rhythms other than mu and beta have also received significant attention in the motor imagery BCI literature. For instance, changes in the occipital alpha, a rhythm associated with visual processing and vigilance, have been identified as source of nonstationarity that should be actively suppressed in the feature extraction process (Blankertz et al., 2008a). Furthermore, it has been hypothesized that gamma oscillations directly affect the sensorimotor rhythm and are responsible for performance variations in BCI (Grosse-Wentrup et al., 2011); this insight may be potentially used to help BCI-illiterate subjects to learn how to control a BCI system. Other recent work has studied the impact of the delta and theta rhythms on BCI performance (Vuckovic and Sepulveda, 2008b; Ahn et al., 2013).

The somatotopic organization of the human brain is the main reason which explains why BCI systems are able to differentiate between different imagined movements. The organization manifests itself in a point-for-point correspondence between the body and the somatic sensory and motor cortex. This arrangement leads to specific spatial activation patterns when performing movements (or motor imagery) with the left and right hand; in the former case a desynchronization of the SMRs occurs in the right sensorimotor cortex in the latter case the left sensorimotor cortex is affected. In addition to the split control over the two body halves there is also a more fine-grained topological organization within the motor and somatic sensory cortices in each hemisphere. The left part of Figure 1 depicts this organization. One can see that the “control ar-

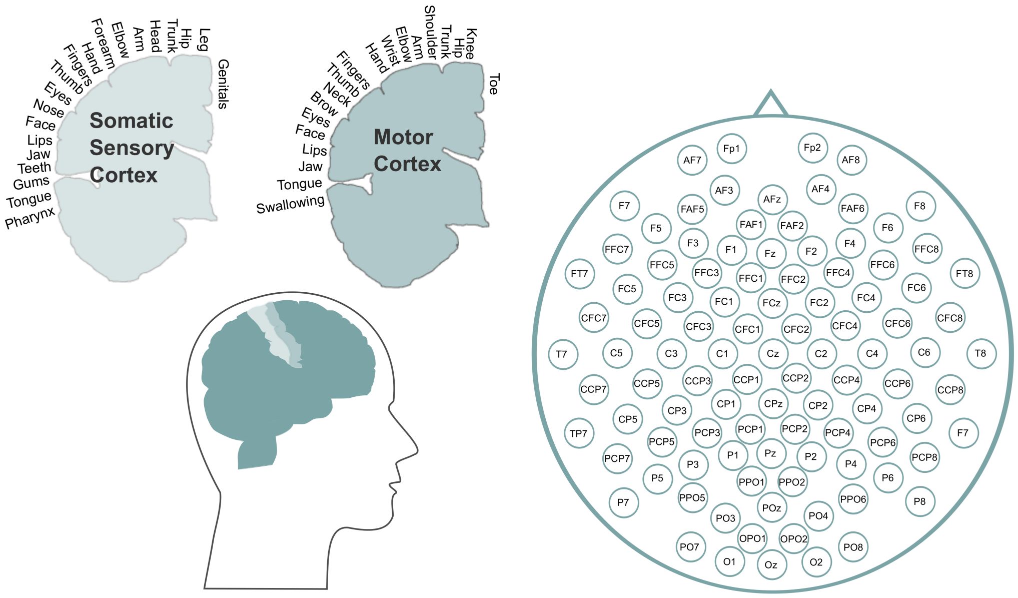

Figure 1.: *Left*: The somatic sensory and motor cortices show a somatotopic organization, i.e., neural assemblies responsible for controlling distinct body parts are spatially separated in a topology preserving manner. *Right*: The international 10-20 system specifies the electrode locations on the scalp. Electrodes on the left hemisphere have an odd number, the right hemisphere electrodes have even numbers; the numbers increase from the center to the outer electrodes. The electrode names represent the underlying cortical regions: F is frontal, P is parietal, O is occipital, T is temporal, C is central.

eas" of different limbs are spatially separated and have a meaningful topology where neighboring body parts, e.g., fingers and hand, are close together. On the other hand there is a relatively large distance between very distinct body parts such as leg, hand and tongue. This somatotopic organization enables us to build BCI systems which can discriminate between more than two motor imagery classes, (see e.g. Grosse-Wentrup and Buss, 2008; Vuckovic and Sepulveda, 2008a; Wang et al., 2012a).

The right part of Figure 1 displays the electrode montage according to the international 10-20 system (Klem et al., 1999). This system specifies the electrode locations on the scalp and assigns standardized names to them. The C3 and C4 electrodes are approximately located over the right hand and left hand area in the sensorimotor cortex whereas Cz mostly captures activity related to foot movements. This correspondence between electrodes and motor imagery classes does not always apply because of individual differences in head size, imperfect montage of the EEG cap and most importantly the effects of volume conduction blurring the underlying neural activity (Nunez, 2006). Spatial filtering reduces the effects of volume conduction and helps to extract the ERD/ERS related activity from the data.

As discussed above, the class-specific spatial signature of the ERD/ERS enables the BCI system to distinguish between different imagined movements. Figure 2 visualizes the power spectrum over the C3 and C4 electrodes when left and right hand motor imagery is performed. The two peaks in the mu and beta bands reflect "idling" oscillations of the underlying neural populations. These peaks disappear when the neurons are engaged in a motor imagery task, i.e., when imagining a movement with the contralateral hand. This power decrease or ERD is very strong for this particular subject, however, on a single-trial basis it may be much weaker and not detectable on individual channels. A multivariate signal analysis and the application of advanced machine learning methods enable us to detect these subtle changes from single trials.

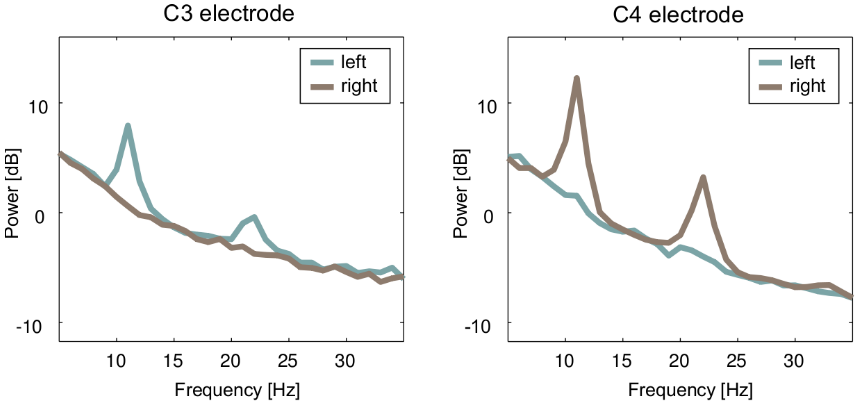

Figure 2.: *Left*: The imagination of a movement with the right hand results in a desynchronization, i.e., decrease of power in the mu and beta band, in the signal recorded by the C3 electrode. *Right*: The imagination of a movement with the left hand results in a desynchronization in the C4 channel.

2.2 MOTOR IMAGERY BCI

Before a BCI system can reliably decode and classify imagined movements it requires calibration. During this process several parameters are optimized and a classifier is learned. Despite the increasing popularity of zero-training BCIs (Krauledat et al., 2008; Fazli et al., 2009; Lotte and Guan, 2010b), subject-calibrated systems are still widely used as they are often superior in terms of classification performance. Figure 3 summarizes the calibration process of a synchronous, motor imagery BCI as typically used by the Berlin BCI (BBCI) group. In contrast to asynchronous systems the subject is only able to communicate his/her intention during a fixed time window (the *trial*) and has no control over its start and duration.

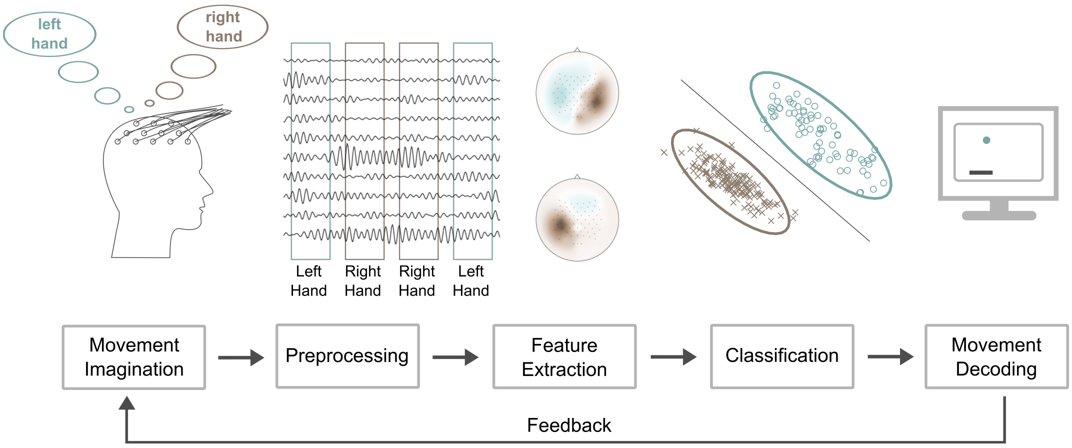

Figure 3.: BCI classification pipeline. In the first step various preprocessing algorithms are applied to the recorded EEG signal. This usually includes spectral filtering, the extraction of the time segment of interest from each trial and artifact rejection. The dimensionality of the signal is reduced in a second step by spatial filtering. Finally, log-variance features are computed from the filtered signal and a classifier is applied to decode the users' intention.

In the calibration session a cue is presented to the subject asking him/her to perform motor imagery tasks, e.g., left hand or right hand movement imagination. This procedure is repeated several times, e.g., 50 times for each motor imagery class. The recorded D-dimensional EEG signal is then cut into $i = 1 \dots n$ epochs, aligned at the start of each trial. Each epoch $\mathbf{X}_i \in \mathbb{R}^{D \times T}$ contains T sample points and at the end of the calibration session a set of labeled trials $\{\mathbf{X}_i, y_i\}_{i=1}^{n}$ is available. Note that the label $y_i = \{-1, +1\}$ indicates the motor imagery class (we assume a binary classification problem) of ith trial. With this data the training process is initiated.

In the first step of this process a time interval is determined that best captures the ERD/ERS effect. One may set this interval a priori, e.g., 500 - 2500 ms after the cue indicating the trial start, or it may be determined adaptively for each subject (Blankertz et al., 2008b). The next step in the processing pipeline consists of spectral filtering. As discussed above the SMR modulation is most prominent in the mu and beta frequency ranges. Various strategies have been proposed to optimally filter the data; among the many options are filtering in a predefined narrow or broad band, using filter banks (Ang et al., 2008), applying spatio-temporal approaches (Lemm et al., 2005; Zhang et al., 2011) and determining a subject-optimized frequency band by cross-validation (Blankertz et al., 2008b). After this step each epoch i is represented by a matrix $\tilde{\mathbf{X}}_i \in \mathbb{R}^{D \times N}$ where $N \leqslant T$ stands for the number of samples in the interval of interest.

The next step of the BCI training is the computation of spatial filters which (in the optimal case) reduce the dimensionality of the data without losing relevant information. Spatial filtering increases the signal-to-noise ratio and reduces the effects of volume conduction. One of the most popular algorithms for this task is *Common Spatial Patterns* (CSP) (Koles et al., 1990; Ramoser et al., 1998; Blankertz et al., 2008b) as it is well suited to discriminate between different mental states induced by motor imagery. A spatial filter $\mathbf{w} \in \mathbb{R}^D$ computed with CSP maximizes the difference in band power between two conditions, thus, it aims to enhance the task-related activity generating the ERD/ERS effect. The CSP spatial filters are computed by solving the generalized eigenvalue problem

$$\mathbf{\Sigma}_1 \mathbf{w}_j \quad = \quad \lambda_j \mathbf{\Sigma}_2 \mathbf{w}_j \tag{1}$$

with $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ being the $D \times D$-dimensional average covariance matrices estimated for the two motor imagery classes. The d generalized eigenvectors from both ends of the spectrum (smallest and largest λ) are often selected as spatial filters. After applying these filters $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_d] \in \mathbb{R}^{D \times d}$ to the data and computing the log-variance features, each trial i is represented by a feature vector

$$\mathbf{f}_i \quad = \quad \log\left(\operatorname{diag}\left[\mathbf{W}^\top \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^\top \mathbf{W}\right]\right) \tag{2}$$

Note that the logarithmic transformation is applied in order to make the features normally distributed (Blankertz et al., 2008b) and diag$[\cdot]$ denotes the operator extracting the diagonal of a matrix. In the final step of the BCI training a classifier $y = h(\mathbf{x})$ separating the two motor imagery classes is computed. We use *Linear Discriminant Analysis* (LDA) for this task

$$h(\mathbf{x}) \quad = \quad \operatorname{sign}(\mathbf{w}\mathbf{x} + b) \tag{3}$$

$$\text{with } \mathbf{w} \; = \; \mathbf{\Sigma}_f^{-1}(\mu_2 - \mu_1) \;\text{ and }\; b \; = \; -\frac{1}{2}\mathbf{w}^\top(\mu_1 + \mu_2)$$

where μ_1 and μ_2 denote the mean feature vectors of class 1 and 2, respectively, $\mathbf{\Sigma}_f$ stands for the covariance matrix of the features and sign$(\mathbf{x})$ is a function with +1 for $\mathbf{x} \geqslant 0$ and -1 otherwise.

In the feedback session the same preprocessing steps are applied as for the calibrating, i.e., for each trial we

(1) extract the time interval of interest

(2) apply the spectral filter

(3) spatially filter the trial

(4) compute the feature vector

(5) classify the trial

Unlike for the calibration session the result of classification may be given to the user as feedback, e.g., in the form of a bar moving on the screen. Note that a change in the feature distribution is often observed between calibration and feedback session (Shenoy et al., 2006; Arvaneh et al., 2013b). This nonstationarity may be for example due to the additional visual processing induced by the feedback or it may come from tiredness if the calibration and feedback session are performed on the same day; but also knowing the results of the classification may of course change the mental strategy of the user and consequently affect the feature distribution. Various strategies may be used in the test session to improve the classification performance of the system, among the most common are adaptation of the classifier (Li and Guan, 2006; Sugiyama et al., 2007; Vidaurre et al., 2011a) and retraining of the filters (Tomioka et al., 2006; Krauledat, 2008; Arvaneh et al., 2013b). Another successful strategy to cope with this nonstationarity is to robustify the feature extraction process (Samek et al., 2012b, 2013c; Arvaneh et al., 2013a; Kawanabe et al., 2014).

2.3 DATA SETS USED IN THIS THESIS

This section gives an overview over the data sets used for experimental evaluation in this thesis. Note that the preprocessing performed in this thesis slightly differ from some of our prepublished papers. The reason for the deviation is comparability, i.e., by considering the same setting for all experiments presented in this thesis we make the results comparable.

Vital BCI data set

This data set (Blankertz et al., 2010a) contains EEG recordings from 80 healthy subjects performing motor imagery tasks with the left and right hand or with the feet. It consists of one calibration and one feedback session, both recorded on the same day. In the calibration session visual cues (arrows pointing left, right, down) indicated which motor imagery task should be performed and three runs with 25 trials of each motor condition were recorded. Then, the best binary combination of motor imagery tasks were selected and the subjects performed feedback with three runs of 100 trials each (some users performed only one or two runs). Visual feedback, i.e., a cursor moving on the screen, was provided to the user while performing motor imagery. Note that this feedback was lacking in the calibration phase. The signals were recorded from 118 Ag/AgCl electrodes, band-pass filtered between 0.05 and 200 Hz and downsampled to 100 Hz. All subjects in this study were BCI novices.

The following preprocessing is applied in the experiments performed on this data set. We manually select 62 electrodes densely covering the motor cortex and filter the data in the frequency range 8-30 Hz with a 5th order Butterworth filter. Furthermore, we use a fixed time segment from 750 to 3500 ms after the trial start for feature extraction. Note that we do not optimize these parameters for individual subject in order to increase comparability of the results and allow between-subject information transfer.

BCI Competition III data set IVa

This data set (Dornhege et al., 2004) was used in the BCI Competition III (Blankertz et al., 2006b) and contains EEG signals from five healthy subjects performing right hand and foot motor imagery without feedback. Two types of visual cues, a letter appearing behind a fixation cross and a randomly moving object, shown for 3.5 s were used to indicate the target class. The presentation of target cues were sandwiched between periods of random length, 1.75 to 2.25 s, in which the subject could relax. The EEG signal was recorded from 118 Ag/AgCl electrodes, band-pass filtered between 0.05 and 200 Hz and downsampled to 100 Hz, so that 280 trials are available for each subject.

We manually select 68 electrodes densely covering the motor cortex and divide the data into a training and test set based on the type of cue used in the experiments. Note that this division does not coincide with the one used for the competition, however, it allows us to study between-paradigm nonstationarity as the cues in the training (letter) and test (moving object) data differ significantly. In our experiments the subjects A1 and A3 have 210 training trials (3 runs) and 70 test trials (1 run) and the remaining users have an equal number of 140 trials (2 runs) in each set. We extract a time segment located from 500 to 2500 ms after the cue instructing the subject to perform motor imagery and band-pass filtered the signal in the frequency range 8-30 Hz using a 5th order Butterworth filter.

Inhouse data set

This data set consists of two calibration (i.e. without feedback) recordings from five healthy participants. The volunteers performed motor imagery of two limbs, specifically left hand and foot. The cues indicating the stimulus were presented either visually (with an arrow appearing in the center of the screen) or as an auditory stimulus (a voice announcing the task to be performed), resulting in two data sets for each user. In our experiments (see Chapter 5) the training data set consists of the calibration recordings with visual stimuli and the test data set contains the recordings with auditory stimuli. A time segment located from 750 to 3500 ms after the cue instructing the subject to perform motor imagery is extracted from each trial and the signal is band-pass filtered in the range 8-30 Hz using a 5th order Butterworth filter. Both the training and test set contain 132 trials, equally divided between each class. We select 85 electrodes densely covering the motor cortex for the experiments presented in this thesis.

3 SPATIAL FILTERING

Electroencephalographic signals reflect not only neural voltage fluctuations underneath the recording electrodes but also capture the activity of distant current sources through volume conduction. Therefore, the signal $\mathbf{x}(t) \in \mathbb{R}^D$ recorded at the scalp is usually modeled as a (noisy) linear mixture (Blankertz et al., 2011),

$$\mathbf{x}(t) \;=\; \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t), \tag{4}$$

where $\mathbf{s}(t) \in \mathbb{R}^D$ are neural sources and $\mathbf{A} \in \mathbb{R}^{D \times D}$ is the matrix mapping the activity of each source to the electrode space. Note that many blind source separation algorithms such as ICA (for reasons of mathematical convenience) assume that the number of brain sources and electrodes coincides. Contributions not captured by $\mathbf{A}$ are modelled as normally distributed noise $\mathbf{n}(t)$. The columns of $\mathbf{A}$ are often referred to as *spatial patterns* of neural activity.

Motor imagery based BCIs aim to focus on the sources of sensorimotor rhythm modulation, i.e., on the neural populations generating the ERD/ERS effect. An estimate of these neural sources is obtained by applying *spatial filters* $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_d] \in \mathbb{R}^{D \times d}$ to the data

$$\hat{\mathbf{s}}(t) \;=\; \mathbf{W}^\top \mathbf{x}(t). \tag{5}$$

These filters project the EEG signal to a d-dimensional subspace and often assume that the estimated sources are uncorrelated, i.e., represent distinct neural populations. This zero correlation assumption can be expressed mathematically by restricting $\mathbf{W}$ to be decomposable into a whitening and an orthogonal projection part.

Note that there is an important difference in the information content of spatial filters and the corresponding spatial patterns (Blankertz et al., 2011; Biessmann et al., 2012; Haufe et al., 2014). A filter is computed by optimizing an objective, e.g., maximizing the variance ratio between classes on the recorded data (not in the source space). Thus, the noise signal $\mathbf{n}(t)$ is considered in the optimization process. For this reason high filter values must not be interpreted as indicator for relevant scalp locations, on the contrary, high values may be solely due to cancelling of noise correlations. A spatial pattern on the other hand allows for a physiologically meaningful interpretation as its coefficients directly reflect the correlation between the channel and the source. In other words when displayed in a scalp plot a spatial pattern enables us to infer the 2D locations of active neural sources. Figure 4 visualizes the difference between filters and patterns; in contrast to the filter which mainly attenuates the noise in the data, a pattern depicts the location of the underlying neuronal activity. Note that spatial patterns can be computed from the filters under the assumption of uncorrelated sources (Biessmann et al., 2012) using the following formula

$$\mathbf{A} \;=\; \mathbf{X}\mathbf{X}^\top \mathbf{W}, \tag{6}$$

where $\mathbf{X} = [\mathbf{x}(1)\ \mathbf{x}(2)\ \ldots]$ represents the data matrix.

3.1 COMMON SPATIAL PATTERNS ALGORITHM

Common Spatial Patterns (CSP) have been widely used in BCI systems (Koles et al., 1990; Ramoser et al., 1998; Blankertz et al., 2008b) as they are well suited to discrim-

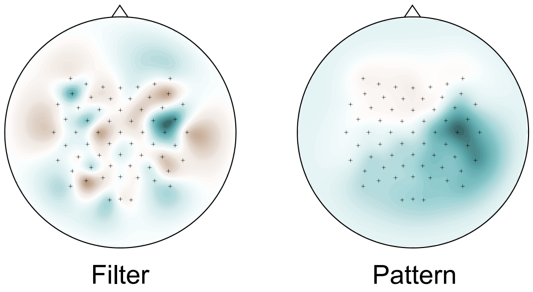

Figure 4.: Visualization of a spatial filter and the corresponding spatial pattern. Only the spatial pattern enables us to infer the 2D locations of active neural sources.

inate between two motor imagery classes. A CSP spatial filter $\mathbf{w} \in \mathbb{R}^D$ maximizes the variance of band-pass filtered EEG signals in one condition while minimizing this quantity in the other condition (or equivalently minimizing the common variance). By definition the variance of a band-pass filtered signal is equal to band power, thus CSP enhances the differences in band power between two conditions (preferably in the frequency band in which ERD/ERS occurs). The CSP spatial filters can be computed by maximizing the Rayleigh quotient

$$R(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w}}, \tag{7}$$

where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are the average covariance matrices from class 1 and 2, respectively. The maximization of this quotient can be formulated as a constrained optimization problem

$$\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w} \tag{8}$$
$$\text{subject to } \mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w} - C = 0$$

where C is an arbitrary constant; the constant is arbitrary because the norm of $\mathbf{w}$ is not fixed. When solving this problem using Lagrange multipliers we arrive at the solution $\mathbf{w}_1^*$ satisfying

$$\mathbf{\Sigma}_1 \mathbf{w}_1^* = \lambda \mathbf{\Sigma}_2 \mathbf{w}_1^* \tag{9}$$

This equation has the form of a generalized eigenvalue problem where the generalized eigenvector with largest eigenvalue λ corresponds to the spatial filter $\mathbf{w}_1^*$ maximizing the Rayleigh quotient. One can show that the CSP filter $\mathbf{w}_2^*$ minimizing Eq. (7) is simply the generalized eigenvector with smallest eigenvalue (Ramoser et al., 1998). The discriminative information is best preserved when selecting d spatial filters from both ends of the spectrum, i.e., $\mathbf{w}$ with largest and smallest eigenvalues λ in Eq. (9). Note that various selection schemes are used in practice:

(1) Selecting the same number of filters from both ends of the spectrum.

(2) Sorting the filters according to their discriminativity value $\alpha_i = \max\left\{\lambda_i, \frac{1}{\lambda_i}\right\}$ and selecting the top d filters.

(3) Selecting the spatial filters adaptively (d not fixed) by applying a heuristic, e.g., median variance criterion (Blankertz et al., 2008b).

Note that we use the second selection scheme (α-sorting) with $d = 6$ throughout this thesis.

3.2 WHY IS CSP NOT ROBUST ?

The CSP method computes spatial filters in a naive data-driven manner. This makes the algorithm vulnerable whenever artifacts are present in the data and may produce suboptimal results due to overfitting.

One major source of error results from the difficulty in proper estimating the class covariance matrices. Since poorly estimated covariance matrices do not well represent the underlying neural processes this directly affects the spatial filter computation. The increasing number of electrodes used in BCI experiments further aggravates the estimation problem. Thus, if data is scarce it becomes very challenging to reliably estimate the high-dimensional covariance matrices without prior information or regularization. Furthermore, the covariance matrix estimation may be negatively affected by EEG artifacts such as eye blinks or loose electrodes. These artifacts often have substantially more signal power than the BCI related activity, thus, if not properly removed they may dominate the covariance matrix estimation and lead to overfitted CSP solutions. It is well know that the standard covariance estimator (mean of signal is zero)

$$C = \frac{1}{N-1} \sum_{t=1}^{N} x(t)x(t)^{\top} \tag{10}$$

is not robust (Huber, 1981) in the sense that artifacts with extreme values of $x(t)$ dominate the averaging. If the artifacts are not similarly distributed between both motor imagery classes they will largely influence the CSP solution. Problems may also occur in small sample settings where the largest eigenvalues of the covariance matrix are overestimated whereas the smallest ones are underestimated (Bartz et al., 2013); this has a direct influence on the variance ratio estimation and therefore may negatively affect CSP. Various techniques, e.g., normalization, shrinkage, robust estimators or incorporating data from additional subjects, have been developed to improve the estimation of covariance matrices in BCI.

Another source of problems is the nonstationarity of the EEG (Kaplan et al., 2005). A time series $\{X_t\}$ is called *strictly stationary* (Priestley, 1981) if for all $(t_1, \dots, t_n) \in \mathbb{Z}^n$, the joint distributions of

$$X_{t_1}, \dots, X_{t_n} \text{ and those of } X_{t_1+h}, \dots, X_{t_n+h}$$

coincide for all time shifts h. Intuitively, stationarity means that the data of two different epochs comes from the same distribution. Note that this property hardly holds for brain signals such as EEG because different processes are active in the brain at different times and the sensory input also changes constantly. However, one may assume that the underlying neural process responsible for movement imagination produces stationary activity or at least is less nonstationary than remaining processes, e.g., responsible for visual information processing. The nonstationary nature of EEG affects all stages of the BCI processing pipeline and leads to changing feature distributions which may negatively affect performance as standard machine learning methods such as Linear Discriminant Analysis assume that data are sampled from a fix (but unknown) distribution (Hastie et al., 2001). Two strategies exist to cope with this problem, namely adaptation to the changes and the usage of robust representations that are invariant to the changes. Note that we use the term *nonstationarity* in a very general sense to indicate all types of changes in feature distribution, i.e., we do not restrict the definition to a time series measured in a particular experimental session from a single subject. We also use the term in the context of data recorded in different sessions or subjects. For instance, we say that there exist a nonstationarity (although the term heterogeneity would fit better) between subjects if their data distributions do not match.

Changes may be measured on various time scales and between different data sources, e.g., epochs, sessions or subjects. They may be class-specific or class-independent. They may occur frequently or very seldom. Finally, they may adversely affect classification performance, do not have an effect or be discriminative (class-related nonstationarity). In the following we discuss three types of nonstationarity which is often present in BCI data.

Within-session changes in the signal distribution are very common and occur on different time scales (Reuderink, 2011; Samek et al., 2012b; Arvaneh et al., 2013a). For instance, artifacts such as loose electrodes, muscle movements, blinking, swallowing, jaw clenching or sudden shifts of attention are often present in EEG recordings and typically affect the signal of one or few trials. Tiredness, changes in electrode impedance or learning effects (Vidaurre et al., 2011c) are on the other hand only visible on larger time scales. Both types of changes may not only corrupt the covariance matrix estimation but also lead to overfitted CSP solutions and increase the variability of the extracted features. Therefore, the application of robust algorithm is crucial for successful BCI operation under nonstationarity.

Between-session nonstationarity are often observed in BCI experiments (Shenoy et al., 2006; Samek et al., 2013c). There are several reasons why data recorded in one session[1] are substantially different from data recorded in other sessions (which may be on a different day), e.g., the calibration of the system may be different, the subjects' state of mind may differ and the position of the electrodes may not match exactly. A particular type of between-session nonstationarity are changes related to the transition from calibration to application phase of a BCI experiment. These changes may be due to addition processing induced by the visual or auditory feedback which is often lacking when calibrating the system. Users may also change the strategy to control the BCI when knowing the result of the classification. Another scenario where CSP may produce results that do not generalize well is that in which the method focuses on discriminative but not motor imagery related activity. For instance assume we use a visual cue (arrow pointing to the left or right) in the training phase to indicate the motor imagery class. A subject may involuntarily perform tiny eye movements when observing the cue, i.e., move the eyes to the direction of the arrow. These ocular movements may induce task-related EEG activity which will be captured by the CSP spatial filters. However, this activity is not related to motor imagery, thus, it becomes meaningless and may deteriorate classification performance if the cue is lacking in subsequent sessions.

Finally, nonstationarity may also be defined in terms of differences between subjects. This type of nonstationarity is not relevant when training a single-subject BCI system, but certainly becomes important when aiming for user-independent BCIs or shorter calibration times (Krauledat et al., 2008; Fazli et al., 2009). A typical scenario where this nonstationarity matters is when utilizing data from additional subjects in order to improve spatial filter computation (Kang et al., 2009; Lotte and Guan, 2010b; Samek et al., 2014). Differences in the signal distribution of different subjects may have many reasons. They may be due to differences in the electrode positions or the user's state-of-mind, but also anatomical differences, e.g., size of the head, may play a significant role. It is advisable to weight the contributions from other subjects according to their relevances (Samek et al., 2013a).

3.3 REGULARIZATION FRAMEWORK

Regularization is a popular strategy to robustify machine learning algorithms. The Common Spatial Patterns method can be regularized by adding a penalty term $P(\mathbf{w})$

[1] Note that we use the term *sessions* for recordings conducted on different days or experiments performed on the same day with a difference in paradigm, cue or feedback.

to the denominator of the Rayleigh quotient (Blankertz et al., 2008a; Lotte and Guan, 2011). This leads to an objective function which maximizes the variance ratio between classes and at the same time aims to minimize the penalty term. It was shown (Mika et al., 2000) that this form of regularization can be used to compute invariant features because spatial filters $\mathbf{w}$ with large $P(\mathbf{w})$ values will be effectively discarded. One can enforce various types of invariances on the features by designing specific penalty terms. The regularized CSP method maximizes the following objectives

$$\mathbf{w}_1^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{w} + \lambda P(\mathbf{w})}, \tag{11}$$

$$\mathbf{w}_2^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\top \boldsymbol{\Sigma}_2 \mathbf{w}}{\mathbf{w}^\top (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{w} + \lambda P(\mathbf{w})}, \tag{12}$$

where $\lambda \geqslant 0$ is a parameter trading-off the maximization of the CSP variance ratio and the minimization of the penalty term. If $\lambda = 0$, then the regularized CSP method coincides with standard CSP.

Note three differences between this algorithm and the CSP objective in Eq. (7). First, the spatial filters computed by the regularized CSP method minimize the overall variance (i.e. sum of class covariance matrices) whereas the original CSP formulation only considers the variance of the other class in the denominator of the Rayleigh quotient. This is not a real difference as it can be shown (Ramoser et al., 1998) that both formulations are equivalent in the sense that the same spatial filters are computed (at least for $\lambda = 0$); the only difference is the scaling, in the former case the objective value is restricted to lie between 0 and 1 whereas it is not upper bounded in the latter formulation. Another seeming discrepancy is that the regularized CSP method maximizes two objectives, one for each class, whereas CSP computes spatial filters by solving only one generalized eigenvalue problem and selecting the eigenvectors from both ends of the eigenvalue spectrum. Optimizing two objectives is necessary (for $\lambda > 0$) because the regularized CSP filter $\mathbf{w}_2^*$ (same holds for $\mathbf{w}_1^*$) maximizing the variance ratio for the second class does not coincide with the filter which minimizes the quotient in Eq. (11), i.e., the filter which minimizes the variance ratio for the first class. The equivalence between eigenvectors minimizing and maximizing the Rayleigh quotients for the two classes only holds for CSP. The third difference between both methods concerns the zero correlation assumption. In the case of CSP the spatial filters $\mathbf{w}_i$ and $\mathbf{w}_j$ are orthogonal (with respect to $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$) and the estimated sources $\hat{s}_i(t)$ and $\hat{s}_j(t)$ are uncorrelated, i.e.,

$$\mathbf{w}_i(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{w}_j = \mathbf{w}_i(\mathbf{A}\mathbf{s}(t)(\mathbf{s}(t))^\top \mathbf{A}^\top)\mathbf{w}_j = \hat{s}_i(t)\hat{s}_j(t) = 0 \quad \text{for} \quad i \neq j. \tag{13}$$

The penalty term in the regularized CSP framework disrupts the orthogonality, consequently, the correlation of the extracted sources may increase or decrease slightly and is not zero anymore. By using the divergence-based framework introduced in Chapter 7 we can regularize the spatial filters without sacrificing the orthogonality property.

In order to compute the regularized CSP filters by solving a generalized eigenvalue problem the penalty term is required to be a quadratic form $P(\mathbf{w}) = \mathbf{w}^\top \mathbf{K}\mathbf{w}$. Only for such penalty terms are we able to write the objective functions in Eq. (11) and Eq. (12) as a Rayleigh quotient $\frac{\mathbf{w}^\top \mathbf{A}\mathbf{w}}{\mathbf{w}^\top \mathbf{B}\mathbf{w}}$ which is maximized by solving a generalized eigenvalue problem. Alternative penalty terms using, e.g., Kullback-Leibler divergences or Euclidean distances, do not allow us to represent the objective in this form. Mathematically, the penalty term translates into an additional quadratic constraint in the Lagrange formulation of the objective (see Eq. (8)). A geometrical interpretation of the optimization is given in Figure 5. For a constant C_1 we can formulate CSP as maximization of $\mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}$ subjected to $\mathbf{w}^\top (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{w} = C_1$. The spatial filter maximizing

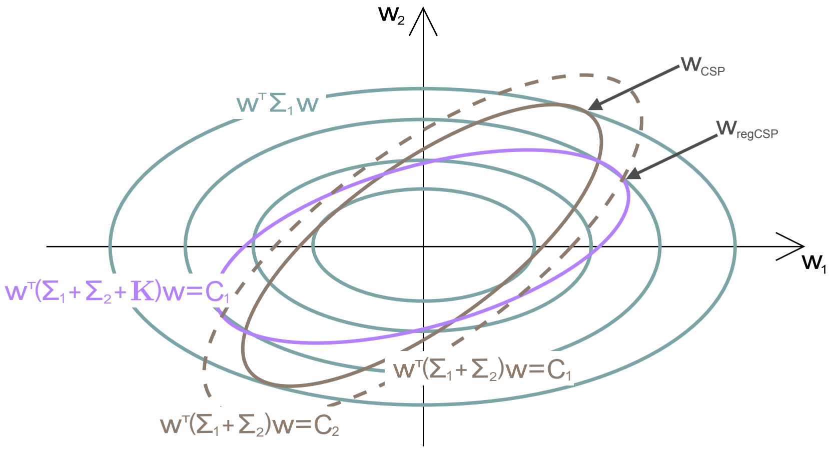
w_2
w_{CSP}
$w^T\Sigma_1 w$
w_{regCSP}
w_1
$w^T(\Sigma_1+\Sigma_2+K)w=C_1$
$w^T(\Sigma_1+\Sigma_2)w=C_1$
$w^T(\Sigma_1+\Sigma_2)w=C_2$

Figure 5.: The maximization of the Rayleigh quotient in Eq. (7) can be interpreted as finding the largest cyan ellipse while satisfying the constraints represented by the brown ellipse. By adding a regularization term $\mathbf{K}$ to the CSP denominator (purple curve) we change the solution from $\mathbf{w}_{CSP}$ to $\mathbf{w}_{regCSP}$. The dashed curve represents the CSP denominator and contains the point $\mathbf{w}_{regCSP}$. One clearly sees that this point does not maximize the Rayleigh quotient. In other words the introduction of a penalty term decreases the CSP objective function but potentially regularizes the optimization to a "better" (e.g. more stable, sparse, stationary) solution.

this optimization problem can be interpreted as point of intersection $\mathbf{w}_{CSP}$ between the brown ellipse representing the constraint and one of the cyan ellipses representing the objective value (outer ellipses stand for high objective values). The inclusion of the penalty matrix $\mathbf{K}$ in the denominator of the Rayleigh quotient can be interpreted as rotation of the brown ellipse towards the purple ellipse. This rotation changes the solution, i.e., we arrive at a new point of intersection $\mathbf{w}_{regCSP}$. Note that the (average) discriminativity decreases by adding the constraint in the sense that

$$\frac{\mathbf{w}_{CSP}^\top \boldsymbol{\Sigma}_1 \mathbf{w}_{CSP}}{\mathbf{w}_{CSP}^\top (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \mathbf{w}_{CSP}} \geqslant \frac{\mathbf{w}_{regCSP}^\top \boldsymbol{\Sigma}_1 \mathbf{w}_{regCSP}}{\mathbf{w}_{regCSP}^\top (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \mathbf{w}_{regCSP}}$$

From a geometrical perspective, the spatial filter $\mathbf{w}_{regCSP}$ lies on the dashed constraint ellipse (which has a different constant C_2), but the point $\mathbf{w}_{regCSP}$ does not maximize the objective value $\mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}$ as intersections with "higher" cyan curves exist. However, by sacrificing a little (average) discriminativity regularization aims to find filters which are in some other sense "better" (e.g. more stable, sparse, stationary) than the solution with maximum variance ratio.

3.4 STATE-OF-THE-ART CSP VARIANTS

In this section we review some of the recently proposed algorithms for spatial filter computation. Note that we focus on CSP-like methods and do not discuss spatio-spectral algorithms such as (Lemm et al., 2005; Zhang et al., 2011) or adaptation strategies such as (Shenoy et al., 2006; Vidaurre et al., 2011b). Furthermore, we do not review the work on artifact identification and removal such as (Jung et al., 2000; Winkler et al., 2011). Figure 6 gives an overview over the presented CSP variants.

Robust Estimation

Several strategies have been proposed to improve the estimation of the covariance matrices used in CSP. For instance, the authors of (Yong et al., 2008; Kawanabe and Vidaurre, 2009) robustly estimate the covariance matrices by using M-estimators. A novel robust estimator for covariance matrices is presented in this thesis and has been prepublished in (Samek and Kawanabe, 2014). Regularization of the covariance matrix (Lu et al., 2009; Kang et al., 2009; Lu et al., 2010; Lotte and Guan, 2011) is also a common approach to increase robustness, especially in small-sample settings. Other authors (Lal et al., 2004; Arvaneh et al., 2011; Goksu et al., 2011) propose to improve the CSP solution by performing channel selection or enforcing sparsity on the spatial filters. The idea of computing CSP in a region of interest was used in (Grosse-Wentrup et al., 2007, 2009; Sannelli et al., 2011). The authors of (Kawanabe et al., 2009, 2014) propose a maxmin approach to robustify the CSP algorithm. A generative CSP model using the robust Student-t distribution was proposed in (Wu et al., 2009). Other methods robustify the variance estimation in CSP by applying L_p-norms (Wang et al., 2012b; Park and Chung, 2013). The authors of (Sannelli et al., 2009) apply trial pruning in order to separate signal from noise and (Parra et al., 2005) discuss several methods for minimum noise estimation. A beta divergence method for robust spatial filtering is proposed in this thesis and has been prepublished in (Samek et al., 2013b).

Another popular way for robustifying the spatial filter computation is the regularization of the CSP algorithm. Many CSP variants use regularization in order to incorporate a priori information (Lotte and Guan, 2010a), avoid overfitting (Lotte and Guan, 2011) or reduce ocular artifacts (Blankertz et al., 2008a). In this thesis we propose a generic divergence-based framework for spatial filter regularization; the framework has been prepublished in (Samek et al., 2014; Samek and Müller, 2014).

Stationary Features

Recently, the development of methods compensating for nonstationarities (Quionero-Candela et al., 2009; Sugiyama and Kawanabe, 2011) has gained increased attention in many application fields of machine learning including Brain-Computer Interfacing. The stationary CSP approach presented in this thesis (Samek et al., 2012b) regularizes the CSP solution towards stationarity in a data-driven manner. The authors of (Arvaneh et al., 2013a) use the same ansatz but apply Kullback-Leibler divergence to measure the changes in the data. Two-step approaches (von Bünau et al., 2010; Samek et al., 2011, 2012a) have also been suggested for computing stationary features. They first estimate and remove the nonstationary contributions and apply CSP to the remaining part of the data in a second step. Furthermore, a second-order baseline was used (Reuderink, 2011) to robustify the algorithm against time and subject related variations. Robust feature extraction methods have also been proposed for reducing between-session nonstationarities (Bamdadian et al., 2012; Arvaneh et al., 2013b). The method presented in this thesis has been prepublished in (Samek et al., 2013c). Some alternative approaches for tackling the between-session nonstationarity problem (Krauledat et al., 2008; Krauledat, 2008; Fazli et al., 2009) utilize data collected in previous sessions, others (Sugiyama et al., 2007; Vidaurre et al., 2011a) update the trained model using adaptation.

Multi-Subject Methods

Many recent algorithms improve the CSP solution by incorporating data from additional subjects. Such approaches are especially important when the objective is to train a subject-independent BCI system or reduce calibration time. The authors of (Devlaminck et al., 2011) jointly train the spatial filters of several subjects by applying a multi-task learning algorithm. A Bayesian method for subject-to-subject information transfer has been proposed in (Kang and Choi, 2011). Data from other users have also been used as regularization target by (Kang et al., 2009; Lotte and Guan, 2010b). A recently proposed method (Samek et al., 2013a) incorporates information from additional subjects by applying Multiple Kernel Learning. An unsupervised BCI based on inter-subject information has been proposed in (Lu et al., 2008).

Other approaches

Some CSP variants improve the quality of the solution by explicitly considering the temporally local structure of observed samples (Wang and Zheng, 2008; Wang and Xu, 2012; Wang, 2013). Other algorithms were specifically designed for multi-class problems and optimize the solution by using information theory (Grosse-Wentrup and Buss, 2008), joint approximate diagonalization (Gouy-Pailler et al., 2010; Nguyen et al., 2012) or Kullback-Leibler Divergence (Wang, 2012). Recently, a spatial filtering method directly linking to Bayes classification error was proposed in (Zhang et al., 2013). Wavelet CSP methods have been considered in (Mousavi et al., 2011; Robinson et al., 2013) whereas the authors of (Falzon et al., 2012) improve the discriminative capability of CSP by taking into account both the amplitude and phase components of the EEG signal. A CSP variant directly optimizing the discriminativity of the features was proposed in (Thomas et al., 2009; Fattahi et al., 2013). A recently published approach (Li et al., 2013) learns spatial filters by considering signal propagation and volume conduction effects. Note that many existing methods try to improve the classification step rather than the CSP computation. For instance, the method presented in (Alamgir et al., 2010) incorporates information from additional subjects by applying multi-task learning whereas the authors of (Sugiyama et al., 2007; Vidaurre et al., 2011a) propose alternative adaptation strategies to cope with nonstationarity. Some authors omit the CSP computation step and suggest to jointly perform feature extraction and classification (e.g. Li and Guan, 2006; Tomioka and Müller, 2010). Other approaches (Barachant et al., 2010, 2012) omit spatial filtering and directly perform classification on the manifold of covariance matrices. Ensemble classification methods have also been used for nonstationary EEG processing (Liyanage et al., 2013).

3.5 BASELINE METHODS USED IN THIS THESIS

In the following we introduce several state-of-the-art spatial filtering methods which are used as baseline in this thesis. These methods tackle different problems which we abbreviate as: robustness to artifacts (A), within- and between-session stationarity (WS and BS) and integration of multi-subject data (MS). Table 2 summarizes the baseline approaches and describes the main property of each method. Note that the methods proposed in this thesis as well as the baseline methods (except shrinkCSP) have one or more free parameters. We determine these parameters (if not stated otherwise) by applying five-fold cross-validation to the training data with minimum error rate as selection criterion.

Tikhonov Regularized CSP (TRCSP): This method (Lotte and Guan, 2011) belongs to the class of regularized CSP approaches introduced in Section 3.3 and uses a regularization function $P(\mathbf{w}) = \mathbf{w}^\top \mathbf{K}\mathbf{w} = ||\mathbf{w}||^2$ which penalizes the norm of spatial filters, i.e., $\mathbf{K}$ is the identity matrix. This penalty term mitigates the influence of artifacts and reduces the tendency to overfitting. As all regularized CSP algorithms TRCSP has a regularization parameter λ trading-off the maximization of the CSP objective and the minimization of the penalty term. We select the regularization

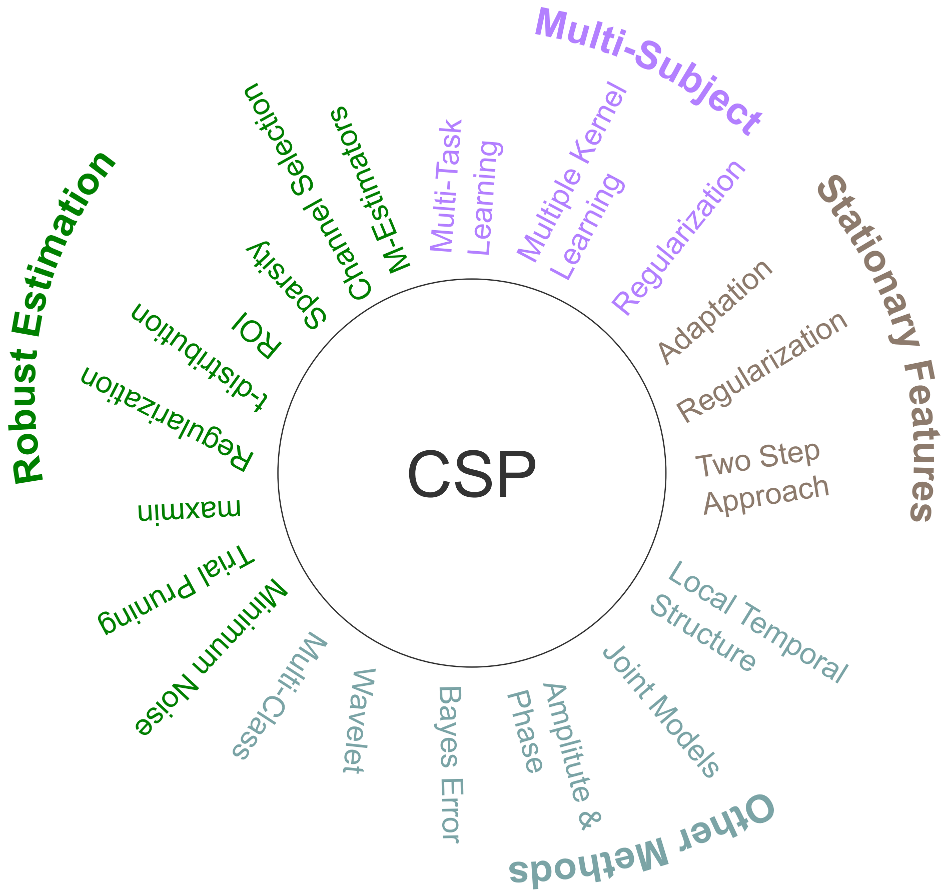

Figure 6.: Overview over several state-of-the-art CSP variants.

parameter from $\{0, 2^{-10}, \dots, 2^{-1}, 2^{0}\}$.

Stationary Subspace Analysis CSP (SSA+CSP): This two-step method projects the data to a stationary subspace prior to CSP computation. The underlying assumption is that the observed signal $\mathbf{x}(t)$ is a linear superposition of stationary $\mathbf{s}^{\mathfrak{s}}(t)$ and nonstationary $\mathbf{s}^{\mathfrak{n}}(t)$ sources

$$\mathbf{x}(t) \;=\; \mathbf{A}\,\mathbf{s}(t) \;=\; \begin{bmatrix} \mathbf{A}^{\mathfrak{s}} & \mathbf{A}^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} \mathbf{s}^{\mathfrak{s}}(t) \\ \mathbf{s}^{\mathfrak{n}}(t) \end{bmatrix}, \qquad (14)$$

and that the BCI-related information is contained in the stationary subspace. The *Stationary Subspace Analysis* (SSA) (von Bünau et al., 2009; Samek et al., 2011) method is applied to separate the $\mathfrak{s}$-sources from the $\mathfrak{n}$-sources and the data is projected to the stationary subspace before applying CSP. The free parameter is the number of directions removed and is selected from {0, 1,. . . , 22}.

Covariance-based CSP (covCSP): This method (Lotte and Guan, 2010b) regularizes the estimated covariance matrix towards the average covariance matrix of the remaining subjects. The CSP filters are computed using these regularized covariance matrices. The covariance matrix of subject i^* is estimated as

$$\tilde{\mathbf{\Sigma}}_{i^*,c} \;=\; (1-\lambda)\mathbf{\Sigma}_{i^*,c} + \lambda \sum_{i \neq i^*} \omega_{i,c} \mathbf{\Sigma}_{i,c}, \qquad (15)$$

Table 2.: State-of-the-art methods used as baseline in this thesis. We apply these methods to several scenarios abbreviated as A = artifacts, WS = within-session stationarity, BS = between-session stationarity, MS = multi-subject.

Name	Scenario	Description
TRCSP (Lotte and Guan, 2011)	WS	Adds identity matrix to denominator of Rayleigh quotient in order to penalize spatial filter norm.
covCSP (Lotte and Guan, 2010b)	BS, MS	Regularizes the class covariance matrices towards the covariance matrices of other subjects.
klcovCSP (Kang et al., 2009)	BS, MS	As covCSP but weights the contributions of other subjects by inverse KL divergence.
SSA+CSP (Samek et al., 2011)	WS	Projects the data to a stationary subspace prior to CSP computation.
KLCSP (Arvaneh et al., 2013a)	WS	Adds KL divergence regularization term to CSP in order to extract stationary features.
shrinkCSP (Lotte and Guan, 2011)	A, WS	Improves estimation of covariance matrix by regularization towards identity and applies CSP.
MCDE+CSP (Yong et al., 2008)	A	Robustly estimates covariance matrix in the presence of outliers and applies CSP.

where $\mathbf{\Sigma}_{i^*,c}$ is the sample covariance matrix of class c for the subject of interest, $\mathbf{\Sigma}_{i,c}$ are the covariance matrices of the other $i = 1 \ldots K$, $i \neq i^*$ subjects and $\lambda \in [0\ 1]$ is a regularization parameter controlling the amount of information incorporated from additional users. The contributions of additional subjects are weighted equally with $\omega_{i,c} = \frac{1}{K-1}$. Note that in this thesis we use all other available subjects, i.e., we do not apply the subject selection proposed in (Lotte and Guan, 2010b). We use the regularization parameters $\{0, 10^{-5}, \ldots, 10^{-1}, 0.2, \ldots, 0.9, 1\}$.

Kullback-Leibler divergence covariance-based CSP (klcovCSP): This method (Kang et al., 2009) applies the same ansatz as covCSP but weights the contributions of additional subjects by the inverse Kullback-Leibler (KL) divergence[2] between the data distribution of subject i and i*. The weights are computed as

$$\omega_{i,c} \quad = \quad \frac{1}{\sum_{i \neq i^*} \frac{1}{D_{kl}(\mathbf{\Sigma}_{i,c} \,\|\, \mathbf{\Sigma}_{i^*,c})}} \frac{1}{D_{kl}(\mathbf{\Sigma}_{i,c} \,\|\, \mathbf{\Sigma}_{i^*,c})}, \qquad (16)$$

where $D_{kl}(\mathbf{\Sigma}_{i,c} \,\|\, \mathbf{\Sigma}_{i^*,c})$ denotes the KL divergence between zero-mean Gaussian distributions with covariances $\mathbf{\Sigma}_{i,c}$ and $\mathbf{\Sigma}_{i^*,c}$. We use the same regularization parameters as for covCSP.

Kullback-Leibler CSP (KLCSP): This method (Arvaneh et al., 2013a) optimizes the CSP objective and simultaneously minimizes within-class variability of the extracted features. Variability is measured as average Kullback-Leibler divergence between the data distribution of a trial and the average distribution of the corresponding class. The following objective function is optimized

$$\min_{\mathbf{w}} \left((1-\lambda)\mathbf{w}^\top \mathbf{\Sigma}_c \mathbf{w} \; + \; \lambda \frac{1}{2n} \sum_{c=1}^{2} \sum_{i=i}^{n} D_{kl}(\mathbf{w}^\top \mathbf{\Sigma}_c^i \mathbf{w} \;\|\; \mathbf{w}^\top \mathbf{\Sigma}_c \mathbf{w}) \right), \qquad (17)$$

$$\text{subject to} \;\; \mathbf{w}^\top (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\mathbf{w} \; = \; 1$$

[2] See definition in Eq. (32) in Chapter 6.

where $\mathbf{\Sigma}_c^i$ is the covariance matrix estimated from ith trial of class c and $\mathbf{\Sigma}_c$ is the average class covariance matrix. The parameter λ trades-off discriminativity and stationarity. When extracting multiple spatial filters we need to include orthogonality constraints. This method is a special case of the divergence-based algorithm proposed in this thesis, namely deflation divCSP-WS (see Chapter 7). We will extensively evaluate the divergence methods in Chapter 7. We select λ from the set $\{0, 0.1, 0.2, \dots, 1\}$.

Shrinkage covariance CSP (shrinkCSP): This method computes spatial filters by applying CSP with regularized covariance matrices (cf. shrinkage LDA (Blankertz et al., 2011)). More precisely, shrinkCSP (Lotte and Guan, 2011) regularizes the estimated covariance matrix towards the identity matrix before computing the spatial filters, i.e.,

$$\tilde{\mathbf{\Sigma}}_c = \mathbf{\Sigma}_c + \lambda \mathbf{I}. \tag{18}$$

The parameter λ is determined analytically by minimizing the bias-variance trade-off (Ledoit and Wolf, 2004), thus, this method has no free parameters.

Minimum Covariance Determinant Estimator CSP (MCDE+CSP): This method (Yong et al., 2008) uses robustly estimated covariance matrices and CSP for spatial filter computation. The minimum covariance determinant estimator (MCDE) (Rousseeuw and Driessen, 1999) finds $h \leqslant N$ samples which have a covariance matrix with the lowest possible determinant, i.e.,

$$\tilde{\mathbf{\Sigma}}_c^h = \underset{\mathbf{\Sigma}_c^h}{\operatorname{argmin}} \, |\mathbf{\Sigma}_c^h| \tag{19}$$

$$\text{with} \quad \mathbf{\Sigma}_c^h = \frac{1}{h-1} \sum_{i=1}^{h} \left(\mathbf{x}_{\sigma(i)} - \frac{1}{h} \sum_{i=1}^{h} \mathbf{x}_{\sigma(i)} \right) \left(\mathbf{x}_{\sigma(i)} - \frac{1}{h} \sum_{i=1}^{h} \mathbf{x}_{\sigma(i)} \right)^{\top},$$

where $\sigma(i)$ is the ith element of the permutation on the samples, thus MCDE resists $(N - h)$ outliers. Based on this estimate, the algorithm assigns weights to the observations such that outliers get zero weight (see Rousseeuw and Driessen, 1999). The final covariance matrix is then estimated from samples with non-zero weight. The parameter h trades-off accuracy of estimation and robustness. We select the free parameter $\lambda = \frac{h}{N}$ from $\{1, 0.95, 0.9, \dots, 0.5\}$.

Part II

NOVEL SPATIAL FILTERING METHODS

4 STATIONARY COMMON SPATIAL PATTERNS

Classification under nonstationarity has recently gained increased attention in the field of machine learning (Quionero-Candela et al., 2009; Sugiyama and Kawanabe, 2011). It is a challenging task because classifiers learned at time t may not perform well at time t + Δt due to changes in the feature distribution. Figure 7 visualizes the problem; each circle and cross represents a feature vector and the color encodes class membership. The three plots depict feature distributions at different times. One can see clearly that there is a change in distribution across time which negatively affects classifiability, i.e., the hyperplane (black line) separating the classes at the beginning gives poor results at later times.

There are (at least) two ways to tackle this nonstationarity problem, namely adaptation of the classifier and extraction of features which are stationary. In this chapter we introduce a novel CSP variant termed *stationary Common Spatial Patterns* (sCSP) which regularizes the CSP solution towards stationary subspaces, i.e., extracts features which are (more) invariant to variations of the signal properties.

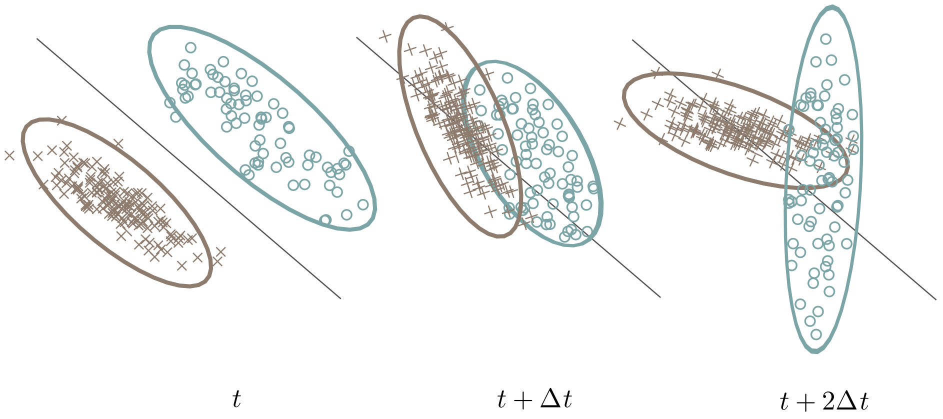

Figure 7.: The feature distributions estimated at different times vary significantly. This nonstationarity negatively affects classifiability of the data. Two ways to tackle this problem are adaptation of the classifier and extraction of invariant features.

4.1 MEASURING NONSTATIONARITY

The sCSP algorithm proposed in this chapter belongs to the class of regularized CSP methods presented in Section 3.3. This novel CSP variant includes a term $P(\mathbf{w})$ in the denominator of the Rayleigh quotient that assigns large penalties to filters $\mathbf{w}$ which lead to a nonstationary feature distribution. Thus, it regularizes the feature extraction step towards stationarity. From a conceptual point of view sCSP applies a similar idea as Linear Discriminant Analysis (LDA) (Hastie et al., 2001), namely it trades-off between-class distance (discriminativity) and within-class variance (nonstationarity). However, in contrast to LDA it does not maximize the difference in class means and minimize the variance of the extracted features, but it rather maximizes the difference in class variances and minimizes the variance of variances. In the following we discuss how to define the penalty term $P(\mathbf{w})$.

One measure commonly used for capturing the variability of features is the average squared difference between the projected average variance and the projected variance of ith trial

$$P(\mathbf{w}) \quad = \quad \frac{1}{2n} \sum_{c=1}^{2} \sum_{i=1}^{n} \left(\mathbf{w}^\top \mathbf{\Sigma}_c^i \mathbf{w} \; - \; \mathbf{w}^\top \mathbf{\Sigma}_c \mathbf{w} \right)^2, \tag{20}$$

where $\mathbf{\Sigma}_c^i$ is the covariance matrix of ith trial of class c, $\mathbf{\Sigma}_c$ is the average covariance matrix of class c and n is the number of trials per class (we assume that it is the same for both classes). Ideally, we would like to maximize the average band power ratio between classes (as done by CSP) and simultaneously minimize $P(\mathbf{w})$ to keep the variance estimation along the projected direction as stable as possible across trials (i.e. minimize the variance of variances). However, as discussed in Section 3.3 this penalty term can not be used in a regularized CSP framework which computes spatial filters by solving a generalized eigenvalue problem because it is not a quadratic form. The same is true for other nonlinear measures of variation, e.g., Kullback-Leibler divergence (Samek et al., 2011; Arvaneh et al., 2013a; Samek et al., 2014).

Another simple measure of nonstationarity is the average absolute difference

$$P(\mathbf{w}) \quad = \quad \frac{1}{2n} \sum_{c=1}^{2} \sum_{i=1}^{n} \left| \mathbf{w}^\top \mathbf{\Sigma}_c^i \mathbf{w} \; - \; \mathbf{w}^\top \mathbf{\Sigma}_c \mathbf{w} \right|. \tag{21}$$

This measure is also not a quadratic form, thus, we can not use it if we want to keep the Rayleigh quotient formulation of the algorithm. Note that the main advantages of formulating the algorithm as generalized eigenvalue problem are the very fast computation and the uniqueness of the solution (global optimum). Thus, in order to keep the Rayleigh quotient formulation we propose to use a related quantity as penalty term. By taking the vector $\mathbf{w}$ out of the absolute value function and ensuring that the difference matrix is positive definite we approximate each term in the absolute difference penalty as

$$\left| \mathbf{w}^\top \mathbf{\Sigma}_c^i \mathbf{w} \; - \; \mathbf{w}^\top \mathbf{\Sigma}_c \mathbf{w} \right| \quad \approx \quad \mathbf{w}^\top \mathcal{F} \left(\mathbf{\Sigma}_c^i \; - \; \mathbf{\Sigma}_c \right) \mathbf{w}, \tag{22}$$

where $\mathcal{F}$ is an operator to make symmetric matrices be positive definite. If a symmetric matrix $\mathbf{M}$ has eigendecomposition $\mathbf{M} = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^\top$, the operator returns $\mathcal{F}(\mathbf{M}) = \mathbf{V}\,|\mathbf{D}|\,\mathbf{V}^\top$, i.e., the signs of all the negative eigenvalues are flipped. The intuition behind this operation is to ensure that the penalty term is always positive, even in the case that the variance in the ith trial is smaller than the global average. Thus, one may think of this operator as an absolute value operator for matrices. The penalty term used in the sCSP method is

$$P(\mathbf{w}) \quad = \quad \mathbf{w}^\top \underbrace{\left(\frac{1}{2n} \sum_{c=1}^{2} \sum_{i=1}^{n} \mathcal{F}(\mathbf{\Sigma}_c^i \; - \; \mathbf{\Sigma}_c) \right)}_{\mathbf{K}} \mathbf{w}. \tag{23}$$

Since $\mathbf{K}$ is a positive definite matrix, our stationary CSP algorithm can be computed within the regularized CSP framework. Although the quantities in Eq. (21) and Eq. (23) are not equivalent, they both measure absolute deviations, in the latter case before and in Eq. (21) after projecting the data. In fact, it can be shown that our new measure is an upper bound for the quantity in Eq. (21).

The following theorems assume that the trial covariance matrices $\boldsymbol{\Sigma}_c^i$ are *jointly diagonalizable*. Mathematically, this means that the eigendecomposition of $\boldsymbol{\Sigma}_c^i$ is

$$\boldsymbol{\Sigma}_c^i = \mathbf{V}\mathbf{D}_c^i\mathbf{V}^\top, \tag{24}$$

with $\mathbf{D}_c^i$ being a diagonal matrix and $\mathbf{V}$ being eigenvectors which are same for all $\boldsymbol{\Sigma}_c^i$. This assumption holds if the signal in different trials is generated by the same mixture model with uncorrelated sources.

Theorem 1. *The average absolute difference penalty defined in Eq. (21) is upper bounded by the sCSP penalty defined in Eq. (23), i.e., the following inequality holds*

$$\frac{1}{2n}\sum_{c=1}^{2}\sum_{i=1}^{n}\left|\mathbf{w}^\top\boldsymbol{\Sigma}_c^i\mathbf{w} - \mathbf{w}^\top\boldsymbol{\Sigma}_c\mathbf{w}\right| \leqslant \mathbf{w}^\top\left(\frac{1}{2n}\sum_{c=1}^{2}\sum_{i=1}^{n}\mathcal{F}(\boldsymbol{\Sigma}_c^i - \boldsymbol{\Sigma}_c)\right)\mathbf{w}$$

under the assumption that all $\boldsymbol{\Sigma}_c^i$ are jointly diagonalizable.

Proof. We show that the inequality holds for each single term

$$\left|\mathbf{w}^\top\boldsymbol{\Sigma}_c^i\mathbf{w} - \mathbf{w}^\top\boldsymbol{\Sigma}_c\mathbf{w}\right| \leqslant \mathbf{w}^\top\mathcal{F}(\boldsymbol{\Sigma}_c^i - \boldsymbol{\Sigma}_c)\mathbf{w}.$$

Let $\mathbf{V}\mathbf{D}\mathbf{V}^\top$ be the eigendecomposition of the difference matrix $\boldsymbol{\Sigma}_c^i - \boldsymbol{\Sigma}_c$ and let $\mathbf{u} = \mathbf{V}^\top\mathbf{w}$. We rewrite the inequality as $|\mathbf{u}^\top\mathbf{D}\mathbf{u}| \leqslant \mathbf{u}^\top|\mathbf{D}|\mathbf{u}$ and $|u_1^2d_1 + u_2^2d_2 + \ldots + u_D^2d_D| \leqslant u_1^2|d_1| + u_2^2|d_2| + \ldots + u_D^2|d_D|$ with u_j being the jth element of vector $\mathbf{u}$ and d_j being the jth diagonal element of matrix $\mathbf{D}$. This is a Jensen's inequality because the absolute value function is convex. Since this inequality holds for each single term, it also holds when summing over the trials and classes. □

The following theorem shows that the spatial filter $\mathbf{w}$ maximizing Eq. (23) also maximizes Eq. (21). This relation ensures that sCSP assigns high penalties to directions with large absolute differences, thus, it allows sCSP to effectively regularize the solution towards stationarity.

Theorem 2. *The sCSP penalty in Eq. (23) and the average absolute difference penalty in Eq. (21) are maximized by the same spatial filter $\mathbf{w}^*$, i.e.,*

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\sum_{c=1}^{2}\sum_{i=1}^{n}\left|\mathbf{w}^\top\boldsymbol{\Sigma}_c^i\mathbf{w} - \mathbf{w}^\top\boldsymbol{\Sigma}_c\mathbf{w}\right| = \underset{\mathbf{w}}{\operatorname{argmax}}\left(\mathbf{w}^\top\left(\sum_{c=1}^{2}\sum_{i=1}^{n}\mathcal{F}(\boldsymbol{\Sigma}_c^i - \boldsymbol{\Sigma}_c)\right)\mathbf{w}\right)$$

under the assumption that all $\boldsymbol{\Sigma}_c^i$ are jointly diagonalizable.

Proof. Let $\mathbf{V}\mathbf{D}_c^i\mathbf{V}^\top$ be the eigendecomposition of the difference matrix $\boldsymbol{\Sigma}_c^i - \boldsymbol{\Sigma}_c$ and let $\mathbf{u} = \mathbf{V}^\top\mathbf{w}$. From Theorem 1 we know that

$$\sum_{c=1}^{2}\sum_{i=1}^{n}\left|\mathbf{u}^\top\mathbf{D}_c^i\mathbf{u}\right| \leqslant \mathbf{u}^\top\underbrace{\left(\sum_{c=1}^{2}\sum_{i=1}^{n}\left|\mathbf{D}_c^i\right|\right)}_{\mathbf{K}}\mathbf{u}.$$

Both terms are equal iff $\mathbf{u}$ has exactly one non-zero entry u_j in the jth row, i.e.,

$$\sum_{c=1}^{2}\sum_{i=1}^{n}\left|u_j^2d_{c,j}^i\right| = u_j^2\left(\sum_{c=1}^{2}\sum_{i=1}^{n}\left|d_{c,j}^i\right|\right),$$

where $d_{c,j}^i$ is the jth diagonal element of $\mathbf{D}_c^i$. The vector $\mathbf{u}$ has exactly one non-zero entry iff $\mathbf{w}$ is orthogonal to all columns of $\mathbf{V}$ except to one, i.e., $\mathbf{w}$ is itself an eigenvector of $\mathbf{V}\mathbf{K}\mathbf{V}^\top$. The eigenvector $\mathbf{w}^*$ of $\mathbf{V}\mathbf{K}\mathbf{V}^\top$ with largest eigenvalue maximizes the quadratic norm $\mathbf{w}^\top \mathbf{V}\mathbf{K}\mathbf{V}^\top \mathbf{w}$, i.e., the right side of the inequality. The corresponding $\mathbf{u}^* = \mathbf{V}\mathbf{w}^*$ has exactly one non-zero entry (i.e., the inequality becomes an equality), thus, $\mathbf{w}^*$ also maximizes the left side of the inequality. □

Since the assumption that all trial covariance matrices are jointly diagonalizable is often not fulfilled in practice, the sCSP penalty will not give the same solutions as the average absolute deviation measure. One can even construct examples (see Section 4.3) in which the sCSP penalty fails to capture the true nonstationarity. When applying sCSP to real data, however, it gives reasonable performance improvements (see Section 4.4). This indicates that Theorem 2 also holds when the covariance matrices are *approximately* jointly diagonalizable. Note also that in contrast to, e.g., the divergence-based nonstationarity measure which is presented later in this thesis, the sCSP penalty allows one to compute the spatial filters very efficiently by solving a generalized eigenvalue problem.

4.2 STATIONARY CSP ALGORITHM

The stationary CSP (sCSP) method is summarized in Algorithm 1. Its input parameters are a set of covariance matrices estimated from the trials $\{\mathbf{\Sigma}_c^i\}$, the number of spatial filters to return d, a regularization parameter $\lambda \geqslant 0$ trading-off stationarity and discriminativity and a parameter ν controlling the time scale of the nonstationarities captured by the penalty matrix. Note that when $\lambda = 0$, then sCSP reduces to CSP.

In the first step of the algorithm average covariance matrices $\mathbf{\Sigma}_c$ are computed for each class. Then, the signal properties of chunks of data are captured by estimating covariance matrices $\tilde{\mathbf{\Sigma}}_c^i$ in epochs of size ν. An epoch of size ν is a set of ν consecutive trials from the same class. For simplicity we assume that ν is a divisor of the number of trials of each class n; if this is not the case, then the residual trials are assigned to the last epoch. The computation of epochs with different ν parameters allows one to capture nonstationarities on various time scales (Wojcikiewicz et al., 2011; Samek et al., 2012b). Note that we use the same method as presented in last section to compute the penalty matrix, however, $\tilde{\mathbf{\Sigma}}_c^i$ does no longer denote the covariance matrix in ith trial, but the covariance matrix estimated from ith epoch. By taking into account nonstationarities on various time scales we are able to cope with different types of changes, e.g., estimating the covariance matrix from individual trials allows one to capture changes such as muscular artifacts which occur on a trial-by-trial basis whereas if the chunk size increases the focus shifts towards slower changes such as variations of task involvement or electrode impedance. We evaluate the influence of the time scale on the penalty matrix later in this chapter.

After computing the penalty terms for both classes we normalize the average covariance matrices $\mathbf{\Sigma}_c$ and the penalty matrices $\mathbf{\Delta}_c$. Note that we compute the sCSP penalty matrix $\mathbf{\Delta}_c$ for each class separately, i.e.,

$$\mathbf{\Delta}_c = \frac{1}{n}\sum_{i=1}^{n} \mathcal{F}(\mathbf{\Sigma}_c^i - \mathbf{\Sigma}_c). \qquad (25)$$

The normalization step is optional but setting the regularization parameter λ is easier if all matrices have approximately the same scale. One possible normalization strategy is to divide the matrices by their traces. If the regularization parameter λ is not specified, then it can be determined by cross-validation. In this thesis we determine the λ parameter from $\{0, 2^{-10}, \ldots, 2^0]$ and the chunk size parameter ν from $\{1, 5, 10\}$ by applying five-fold cross-validation with minimum error as selection criterion.

In the next step of the algorithm the two Rayleigh quotients in Eq. (11) and Eq. (12) are maximized by solving two generalized eigenvalue problems. Finally, we select d

eigenvectors from both eigendecompositions by using one of the selection schemes described in Section 3.1. In this thesis we use scheme (2) for all experiments, e.g., we pool the solutions of the two generalized eigenvalue problems, sort the eigenvectors by decreasing eigenvalue and select the top d eigenvectors as spatial filters. Note that we do not combine sCSP with additional regularization targets, e.g., with Tikhonov Regularization as done in (Samek et al., 2012b), because we aim to separately study the effects of regularization towards stationarity.

Algorithm 1 Stationary Common Spatial Patterns

```
function sCSP($\{\boldsymbol{\Sigma}_c^i\}$, $d$, $\lambda$, $\nu$)
    Compute average covariance matrices $\boldsymbol{\Sigma}_c = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Sigma}_c^i$.
    Compute covariance matrices $\tilde{\boldsymbol{\Sigma}}_c^i = \frac{1}{\nu}\sum_{j=(i-1)\cdot\nu+1}^{i\cdot\nu} \boldsymbol{\Sigma}_c^j$ of epochs.
    Compute penalty matrices $\boldsymbol{\Delta}_c = \frac{\nu}{n}\sum_{i=1}^{\frac{n}{\nu}} \mathcal{F}(\tilde{\boldsymbol{\Sigma}}_c^i - \boldsymbol{\Sigma}_c)$.
    Normalize average covariance matrices and penalty matrices.
    Compute eigenvectors $\mathbf{V}_c = \mathrm{eig}(\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 + \lambda(\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2))$
    Select $d$ columns $\mathbf{W} \in \mathbb{R}^{D\times d}$ from $\mathbf{V}_1$ and $\mathbf{V}_2$.
    return W
end function
```

4.3 FAILURE OF APPROXIMATION

The following example[1] demonstrates that the heuristic ($\mathcal{F}$ operator) used by sCSP to construct the penalty matrix may fail, i.e., sCSP does not penalize the true nonstationarities in the data. Assume we have the following matrices

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.9 & 0.15 \\ 0.15 & 0.1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.9 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_1^1 = \begin{bmatrix} 0.9 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1^2 = \begin{bmatrix} 0.9 & 0.25 \\ 0.25 & 0.1 \end{bmatrix}$$

where $\boldsymbol{\Sigma}_c$ denotes the average covariance matrix of class c and $\boldsymbol{\Sigma}_1^1$ and $\boldsymbol{\Sigma}_1^2$ stand for the covariance matrices estimated from trial 1 and 2 of class 1, respectively. Note that we only assume class 1 to be nonstationary, i.e., the trial covariance matrices of class 2 coincide with $\boldsymbol{\Sigma}_2$. If we aim to maximize the ratio between the variance of class 1 and 2 and simultaneously want to minimize nonstationarity, then the optimal spatial filter is $\mathbf{w}^\top = [1 \;\; 0]^\top$. Considering the class differences in the off-diagonal elements of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ leads to a higher Rayleigh quotient (therefore it is preferred by CSP), but introduces variability to the extracted features. The penalty term of sCSP is computed as

$$\boldsymbol{\Delta} = 0.5 \cdot \mathcal{F}\left(\boldsymbol{\Sigma}_1^1 - \boldsymbol{\Sigma}_1\right) + 0.5 \cdot \mathcal{F}\left(\boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_1\right) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},$$

where $\mathcal{F}$ is the operator that flips the negative eigenvalues of a matrix. Since adding this matrix to the denominator of the Rayleigh quotient will not penalize the off-diagonal elements, sCSP will not extract the filter $\mathbf{w}^\top = [1 \;\; 0]^\top$. In other words the

[1] We thank the authors of (Arvaneh et al., 2013a) for mentioning the deficits of sSCP to us.

flipping sign heuristic fails in this example because the assumption which says that the covariance matrices are jointly diagonalizable is violated. Stationary CSP variants based on Kullback-Leibler divergence (Samek et al., 2011; Arvaneh et al., 2013a; Samek et al., 2014) will penalize the off-diagonal terms because they do not rely on heuristics but rather evaluate nonstationarity in a principled manner (see Section 7.5.2).

4.4 EXPERIMENTAL EVALUATION

This section empirically investigates the effects of regularizing the CSP solution towards stationarity. We apply sCSP to two data sets and compare the results with several state-of-the-art methods. We demonstrate that our novel regularization strategy is in many situations superior to standard techniques and interpret the reasons for the improvement in classification accuracy from a neurophysiological perspective. Before we present results on real data, we investigate the impact of different parameters on the penalty matrix computation in a simulation study; in particular we analyse the influence of chunk size.

4.4.1 *Simulations*

When introducing the sCSP regularization term (see Eq. (23)) we mentioned that sCSP only approximately minimizes the quantity we would like to minimize, namely the variance of the variances (see Eq. (20)). In the following controlled simulation experiment we investigate the quality of this approximation.

Consider the observed signal $\mathbf{x}(t) \in \mathbb{R}^D$ generated as noisy mixture of D sources $\mathbf{s}(t) = [s_1(t) \dots s_D(t)]^\top$ with a random orthogonal mixing matrix $\mathbf{A} \in \mathbb{R}^{D \times D}$

$$\mathbf{x}(t) \quad = \quad \mathbf{A}\mathbf{s}(t) \; + \; \xi(t).$$

We generate n trials with N samples per trial and each source signal $s_2(t) \dots s_D(t)$ is sampled from a zero mean normal distribution with standard deviation 1 and the noise $\xi(t)$ is sample from an isotropic Gaussian with standard deviation 2. The first source $s_1(t)$ is nonstationary, i.e., for each trial i it is sampled from a Gaussian with a different variance parameter $\sigma_i^2 = 1 + \epsilon$ and $\epsilon \sim \mathcal{N}\left(0, \omega^2\right)$. Note that we ensure that $\sigma_i^2 > 0$. In summary, we generate noisy data with one nonstationary source.

The sCSP regularization term $P(\mathbf{w})$ aims to penalize a projection $\mathbf{w}$ to the nonstationary source s_1, i.e., the sCSP penalty matrix $\boldsymbol{\Delta}$ should be such that $P(\mathbf{w}) = \mathbf{w}^\top \boldsymbol{\Delta} \mathbf{w}$ is much larger for $\mathbf{w}$ being the projection to s_1 than for $\mathbf{w}$ being the projection to the other sources. Figure 8 depicts the median (100 repetitions) ratio of the penalty value when projecting to source s_1 and the maximum penalty value when projecting to the other sources for different n, D, N and ω. More precisely, if $\mathbf{B} = \mathbf{A}^{-1}$ and $\mathbf{b}_i$ is the ith row of $\mathbf{B}$, we plot the variance ratio

$$r = \frac{\mathbf{b}_1 \boldsymbol{\Delta} \mathbf{b}_1^\top}{\max_{i=2 \dots D} \left\{ \mathbf{b}_i \boldsymbol{\Delta} \mathbf{b}_i^\top \right\}}$$

The black dashed line represents value 1, i.e., the case when the penalty of the projection to source 1 equals the penalty of the projection to another source. The sCSP penalty matrix should assign higher penalty values to the projection to source 1 because this source is nonstationary, thus, values above the dashed line represent well estimated penalty matrices. One clearly sees that the quality of the sCSP penalty matrix varies largely with the number of trials n, the number of samples per trial N, the dimensionality of the data D and the strength of the nonstationarity ω. This is not very surprising because the estimation of the covariance matrices used for the penalty matrix computation also largely depends on these parameters. On the other hand even when the sCSP regularization term does not assign the largest penalty value to the true nonstationary projection, sCSP may still perform well as long as $P(\mathbf{w})$ assigns a large enough penalty to this projection direction and does not penalize the BCI-relevant directions in the data. In summary, this example illustrates that sCSP does what it is supposed to do as long as the covariance matrices are reliably estimated.

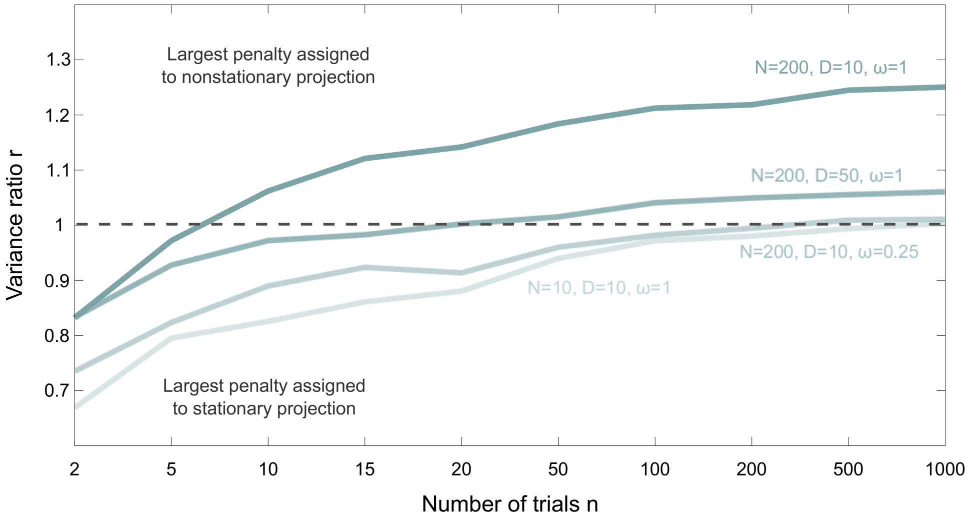

Figure 8.: Ratio of the penalty value when projecting to the true nonstationary source and the maximum penalty value when projection to another source for different parameters such as number of trials n, number of samples per trial N, dimensionality of the data D and strength of the nonstationarity ω. The dashed line represents the equality of both penalty terms (ratio of 1); if the ratio is smaller than 1, then sCSP finds a projection direction which is supposed to be more nonstationary than the projection to the true nonstationary source.

An important parameter of the sCSP algorithm is the chunk size. We mentioned before that only a small chunk size can capture single-trial nonstationarities such as electrode artifacts whereas larger chunk sizes focus on larger time scales and may oversee single-trial events. We want to investigate this difference from a theoretical perspective here. Assume that all trial covariance matrices $\mathbf{\Sigma}^i$ share the same eigenspace, i.e., are jointly diagonalizable. Then, for chunk size 1 the contribution of the first k trials to the sCSP penalty matrix (assume number of trials $n > k$) is

$$\mathbf{\Delta}_{\nu=1} = \frac{1}{k}\sum_{i=1}^{k} \mathcal{F}(\mathbf{\Sigma}^i - \mathbf{\Sigma}) = \mathbf{V}\left(\frac{1}{k}\sum_{i=1}^{k} |\mathbf{D}^i - \mathbf{D}|\right)\mathbf{V}^\top,$$

where $\mathbf{\Sigma}$ is the average covariance matrix, $\mathbf{V}$ are the common eigenvectors and $\mathbf{D}^i$ and $\mathbf{D}$ are diagonal matrices containing the eigenvalues of $\mathbf{\Sigma}^i$ and $\mathbf{\Sigma}$, respectively. When using chunk size k we obtain

$$\mathbf{\Delta}_{\nu=k} = \mathcal{F}\left(\frac{1}{k}\sum_{i=1}^{k} \mathbf{\Sigma}^i - \mathbf{\Sigma}\right) = \mathbf{V}\left|\frac{1}{k}\sum_{i=1}^{k} \mathbf{D}^i - \mathbf{D}\right|\mathbf{V}^\top.$$

According to Jensen's inequality the following relation holds $\mathbf{\Delta}_{\nu=k} \leqslant \mathbf{\Delta}_{\nu=1}$. In order to demonstrate the difference between small and large chunk sizes we construct an example where $\mathbf{w}^\top\mathbf{\Delta}_{\nu=1}\mathbf{w}$ is large but $\mathbf{w}^\top\mathbf{\Delta}_{\nu=k}\mathbf{w}$ is small, i.e., sCSP with chunk size

1 penalizes a spatial filter that is not penalized by sCSP with chunk size k. Assume we have covariance matrices of four consecutive trials

$$\boldsymbol{\Sigma}^1 = \begin{bmatrix} 0.9 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{\Sigma}^2 = \begin{bmatrix} 1.1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^3 = \begin{bmatrix} 0.9 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{\Sigma}^4 = \begin{bmatrix} 1.1 & 0 \\ 0 & 1 \end{bmatrix}$$

The average covariance matrix is then

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and the penalty matrix for chunk size 1 can be computed as

$$\boldsymbol{\Delta}_{\nu=1} = \frac{1}{4}\sum_{i=1}^{4} \mathcal{F}(\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0 \end{bmatrix}$$

Thus, the projection $\mathbf{w}^\top = [1 \;\; 0]$ that leads to nonstationary features has a non-zero penalty. In other words the trial-wise variations are penalized by sCSP with chunk size 1. On the other hand when computing the penalty matrix for chunk size 2 we obtain

$$\boldsymbol{\Delta}_{\nu=2} = \frac{1}{2}\sum_{j=1}^{2} \mathcal{F}\left(\frac{1}{2}\sum_{i=2\cdot(j-1)+1}^{2\cdot j} \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}\right) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

The projection $\mathbf{w}^\top = [1 \;\; 0]$ is not penalized when applying sCSP with chunk size 2. In this sense there is a qualitative difference between both penalty matrices, small chunk sizes capture trial-wise variations which may be overlooked when using larger chunk sizes. Note that in some situations it may be advantageous to ignore single-trial variations and focus on slower changes; in other data sets the opposite may be true. In our experiments we select the optimal chunk size from $\{1, 5, 10\}$ by cross-validation in order to adapt to the nonstationarities present in the data of each subject.

4.4.2 *Performance Results*

In the following we evaluate the sCSP method on two data sets described in Section 2.3, namely the Vital BCI data set and the BCI Competition III data set IVa. We compare the performance results[2] to CSP and four baseline methods which also use regularization in the spatial filter computation process. Figure 9 displays the results for the first data set. The error rate of sCSP is shown on the y-axis whereas the error rates of the baseline methods are depicted on the x-axes. Each subject is represented by a circle and when the circle is below the solid line our method outperforms the baseline approach for this particular subject. We display the p-values of the one-sided Wilcoxon sign rank test in the bottom right corner. Note that the null hypothesis of the test is that the median of the error rate differences (our method - baseline method) is greater or equal to zero. For $p < 0.05$ we reject this null hypothesis, thus, we say that our method significantly outperforms the baseline. The free parameters of all methods are selected by five-fold cross-validation on the training data.

[2] These results differ from the results in (Samek et al., 2012b) because we use a different spatial filter selection scheme and different preprocessing here in order to be consistent with the results in later chapters.

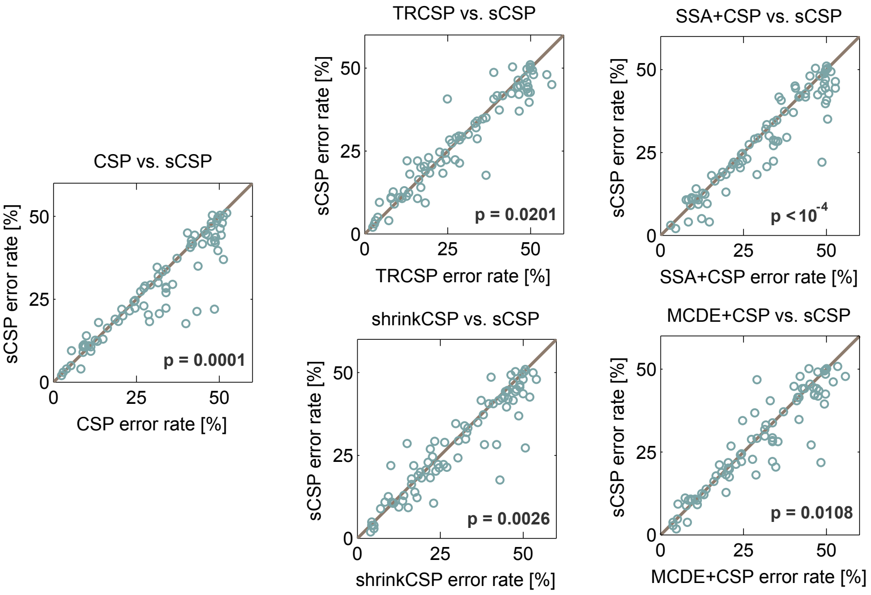

Figure 9.: Scatter plots showing error rates of sCSP and five baseline methods. Each circle represents one subject and if the circle is below the solid line, then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner.

One clearly sees that the regularization performed by sCSP significantly improves the classification accuracy for all baselines. This means that the sCSP regularization (for this data set) is more efficient than the Tikhonov regularization (TRCSP) presented in (Lotte and Guan, 2011), the two-step method (SSA+CSP) introduced in (Samek et al., 2011) and CSP with improved estimates of the class covariance matrix (shrinkCSP (Lotte and Guan, 2011) and MCDE+CSP (Yong et al., 2008)). Thus, it seems that nonstationarity is a major problem in this data set. We conjecture that one reason may be that all subjects in the Vital BCI data set are BCI novices, thus they probably have not developed a stable strategy for performing motor imagery yet. Note that every baseline method has its own area of application. For instance, TRCSP penalizes the norm of the spatial filters, i.e., avoids overfitting, the shrinkage estimator provides optimal (in terms of bias-variance trade-off) covariance matrix estimates when data is scarce and MCDE+CSP is designed to perform well when the signal is affected by artifacts. The sCSP algorithm[3] is the first CSP variant to regularize the solution towards stationarity; SSA+CSP is a two step method and has its own deficits (discussed in Section 7.5.1). For the Vital BCI data nonstationarity is a large problem, therefore sCSP performs very well. For other data sets different challenges such as lack of data or large artifacts may be more relevant and the other approaches may outperform sCSP. In practice it may be advantageous to combine several regularization strategies (c.f. Samek et al., 2012b).

Figure 10 visualizes the sCSP parameters selected by cross-validation on the Vital BCI data set. It is interesting to note that chunk size 1 is preferred over chunk size 5 and 10. Thus, it seems that changes are mostly present on a trial-by-trial basis in the data. This preference also occurs when selecting the parameters a posteriori, i.e., by minimizing test error. The regularization parameter $\lambda = 2^{-3}$ is the most popular choice. When applying sCSP with the most popular parameters $\lambda = 2^{-3}$ and $\nu = 1$ to all subjects we obtain a performance increase over CSP which is significant with

[3] Another CSP variant which regularizes the solution towards stationarity, KLCSP (Arvaneh et al., 2013a), was proposed after sCSP and will be discussed in the context of our divergence CSP framework in Chapter 7

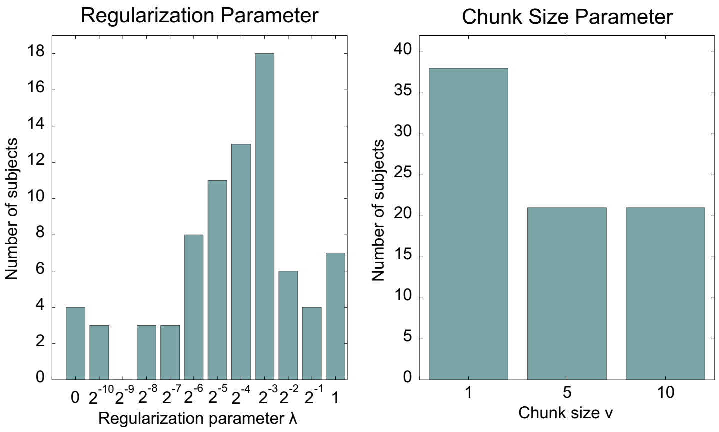

Figure 10.: *Left*: Selected regularization parameter. *Right*: Selected chunk size.

Table 3.: Comparison of classification accuracies for sCSP and different baselines on the BCI Competition III data set IVa.

	BCI Competition III					Overall		
Subject	A1	A2	A3	A4	A5	Mean	Median	Std
CSP	66.1	94.6	58.2	87.9	90.5	79.5	87.9	16.3
TRCSP	66.1	94.6	61.7	87.5	**91.3**	80.2	87.5	15.2
SSA+CSP	66.1	94.6	58.7	85.3	83.3	77.6	83.3	14.8
shrinkCSP	**74.1**	94.6	62.8	82.6	90.5	80.9	82.6	12.8
MCDE+CSP	68.8	94.6	59.2	**89.7**	88.5	80.2	88.5	15.3
sCSP	66.1	**96.4**	**65.8**	**89.7**	90.9	**81.8**	**89.7**	14.7

$p = 0.0482$. For the other popular values of λ, namely $2^{-6}, 2^{-5}$ and 2^{-4}, the p-values are also highly significant ($p = 0.0027$, $p = 0.0473$ and $p = 0.0070$). This result suggests that sCSP can be applied with fixed parameters in practice.

Finally, we compare the six spatial filter computation methods on the second data set. Table 3 displays the results. Also here sCSP outperforms the baselines; it gives best performance (bold font indicates highest classification accuracy) for three out of five subjects. We do not provide p-values because they can not be reliably estimated for this limited number of subjects.

4.4.3 *Further Analysis*

In the following we analyse and interpret the sCSP results in more detail. First we demonstrate that sCSP reduces the nonstationarity of the extracted features, i.e., the flipping sign heuristic works in practice. Note that the original goal of sCSP was to reduce the variance of the features, however, we could not directly use this penalty in the Rayleigh quotient formulation of the algorithm. Therefore we minimize a different quantity which was shown to be related (see Theorem 2) to average absolute deviation. Figure 11 visualizes how well the original objective, i.e., the minimization of the feature variance, is achieved on the Vital BCI data set. The figure displays the relative change in the variance when applying sCSP with various regularization parameters (we fix the chunk size to 1). The left boxplot depicts the changes in mean variance (over all six spatial filters), i.e.,

$$\kappa_\lambda = \frac{\sum_{j=1}^{6}\sum_{c=1}^{2}\sum_{i=1}^{n}\left(\mathbf{w}_\lambda^j \boldsymbol{\Sigma}_c^i \mathbf{w}_\lambda^j - \mathbf{w}_\lambda^j \boldsymbol{\Sigma}_c \mathbf{w}_\lambda^j\right)^2}{\sum_{j=1}^{6}\sum_{c=1}^{2}\sum_{i=1}^{n}\left(\mathbf{w}_0^j \boldsymbol{\Sigma}_c^i \mathbf{w}_0^j - \mathbf{w}_0^j \boldsymbol{\Sigma}_c \mathbf{w}_0^j\right)^2},$$

where $\mathbf{w}_\lambda^j$ denotes the jth spatial filter computed by sCSP with parameter λ and $\boldsymbol{\Sigma}_c^i$ and $\boldsymbol{\Sigma}_c$ stand for the trial and average covariance matrices of class c, respectively. The right boxplot displays the changes in maximum variance

$$\kappa_\lambda = \frac{\max_{j=1\ldots6}\left\{\sum_{c=1}^{2}\sum_{i=1}^{n}\left(\mathbf{w}_\lambda^j \boldsymbol{\Sigma}_c^i \mathbf{w}_\lambda^j - \mathbf{w}_\lambda^j \boldsymbol{\Sigma}_c \mathbf{w}_\lambda^j\right)^2\right\}}{\max_{j=1\ldots6}\left\{\sum_{c=1}^{2}\sum_{i=1}^{n}\left(\mathbf{w}_0^j \boldsymbol{\Sigma}_c^i \mathbf{w}_0^j - \mathbf{w}_0^j \boldsymbol{\Sigma}_c \mathbf{w}_0^j\right)^2\right\}}.$$

In both cases one sees a steady decrease of these quantities, i.e., the variance of the extracted features is reduced in comparison to the CSP solution ($\lambda = 0$). In this sense sCSP does exactly what it is supposed to do, namely reducing the nonstationarity of the features. For some subjects, especially in the left panel, the sCSP heuristic fails i.e., the variance of the features increases. One potential explanation for this is that the trial covariance matrices are not jointly diagonalizable, thus, the sCSP approximation is suboptimal (Theorem 2 does not hold); another reason may be that this effect is due to averaging over all spatial filters, i.e., the decrease in variance for the first spatial filters goes along with a larger increase for the other filters. This may be the case as the variance increase is less present in the right boxplot which displays the change for one spatial filter. The outlier points marked by the circles represent subjects who have a large increase in variance of the features. These subjects not only shows a significant increase in nonstationarity but also a performance gain, namely in the one case from almost chance level to error rates of 33% and 22% and in the the other case from an error rates of 21% to error rates of 13% and 16%. As suggested these subjects show a reduction in variance for some of its spatial filters and an increase for others. Note also that the initial nonstationarity level (for $\lambda = 0$) is very low for the filters of one of the subjects, thus, the sCSP penalty raises the variance to "normal level". Therefore, for these two subjects an average increase in nonstationarity goes along with an performance increase; for most other users the opposite is true.

In the next analysis we investigate the impact of chunk size on performance. For that we compare the results of sCSP when selecting the chunk size by cross-validation to the results of sCSP when using a fixed chunk size of 1, 5 and 10. Table 4 gives an overview over the average classification accuracies and provides p-values of the one-sided Wilcoxon signrank test. A p-value smaller than 0.05 means that sCSP with chunk size selection (sCSP with CV) significantly outperforms sCSP with fixed chunk size parameter. From the results we see that selecting a subject-specific chunk size, i.e., measuring nonstationarities on a subject-optimized time scale, is very important as all p-values are below the significance level of 0.05.

As discussed earlier selecting an appropriate chunk size allows us to focus on the type of changes present in the data, e.g., trial-wise nonstationarities or changes on a larger time scale. Figure 12 displays the differences of the sCSP penalty matrices when computed with chunk size 1 and 5. Here we visualize the top 10 eigenvectors of subject's 21 penalty matrix. Note that the top eigenvectors of the penalty matrix are the directions which are mostly penalized. One clearly sees that very similar directions are penalized when using chunk size 1 and chunk size 5. Despite this general similarity there are differences in the ordering of the eigenvectors (i.e. strength of penalty) as well as differences in the eigenvectors themselves. It is hard to say why sCSP with

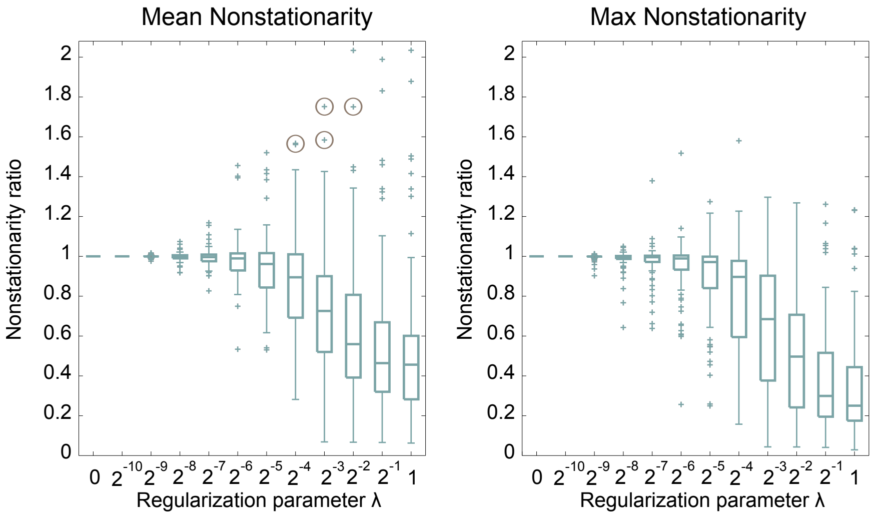

Figure 11.: *Left*: Change in mean variance of the extracted features over all six spatial filters. *Right*: Change in maximum variance of the extracted features.

Table 4.: Comparison of classification accuracies of sCSP with different chunk sizes. The last column displays the p-values of the one-sided Wilcoxon sign-rank test when comparing sCSP with CV to the other methods.

Method	Mean	Median	p-value
CSP	68.7	66.5	0.0001
sCSP with $\nu = 1$	70.7	71.2	0.0270
sCSP with $\nu = 5$	70.1	70.4	0.0098
sCSP with $\nu = 10$	70.4	70.3	0.0331
sCSP with CV	71.2	71.5	–

chunk size 5 gives better results than sCSP with chunk size 1 for this subject. We conjecture that sCSP with chunk size 5 penalizes activity patterns which are very similar to the artifactual CSP activity patterns displayed in Figure 13, especially the third and seventh eigenvector of the penalty matrix are similar to CSP patterns 1 and 6 in Figure 13. There is also a similarity for the chunk size 1 case (3th and 9th eigenvector), however, the similarity is lower than when using a larger chunk size. It seems that the chunk size 1 penalty matrix is simply more noisy (single-trial effects) for this particular subject.

Subject 21 performs left hand vs. right hand motor imagery and has the following error rates: 40% (CSP), 36.7% (TRCSP), 18% (SSA+CSP), 43% (shrinkCSP), 21% (MCDE+CSP), 17.7% (sCSP). It is interesting to mention that neither TRCSP nor shrinkCSP significantly improves classification performance for this subject, even when selecting the regularization parameter a posteriori by minimizing test error. On the other hand SSA+CSP, a method that does not perform very well on average because it is a two-step method (see discussion in Section 7.5.1), improves classification accuracy. Thus, it seems that this subject has a large nonstationarity problem that can neither be tackled by penalizing the norm of the spatial filters nor by using improved covariance matrix estimation; only regularization towards stationarity helps for this subject. Figure 13 visualizes the activation patterns of CSP and sCSP. We display the area under the ROC curve (AUC) measure (Bradley, 1997) below each spatial pattern;

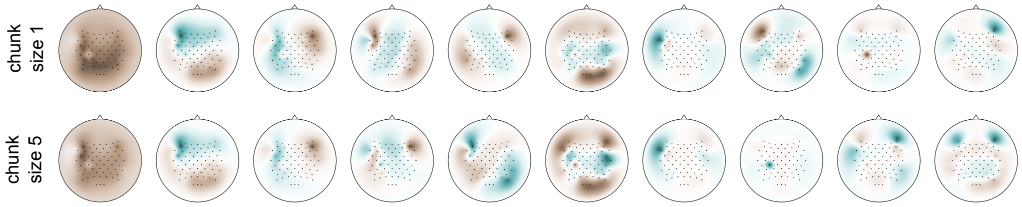

Figure 12.: *Top row*: Top eigenvectors of the penalty matrix for chunk size 1. *Bottom row*: Top eigenvectors of the penalty matrix for chunk size 5.

higher values represent better performance. The AUC measure better summarizes the quality of the spatial filter than error rate because it is robust to bias shifts. Note that both CSP and sCSP compute a neurophysiologically meaningful spatial pattern for the right hand motor imagery class (second pattern of CSP, first pattern of sCSP). These patterns also have relatively large AUC values, i.e., are discriminative. However, for the left hand condition the CSP algorithm fails to compute meaningful patterns, i.e., no pattern captures activity over the right sensorimotor cortex. The fourth sCSP pattern captures this activity and has a high AUC value of 0.8. Therefore sCSP provides much better performance results than CSP; it captures activity related to both motor imagery classes. Figure 14 visualizes (parts of) the filtered EEG signal in channel FFC6. The artifacts in this electrode are the reason which explains why CSP fails to capture activity related to left hand motor imagery. In other words the main difference between the CSP and sCSP patterns in Figure 13 is that CSP contains an artifact pattern (6th pattern) instead of capturing neurophysiologically meaningful activity (4th pattern of sCSP). The sCSP method is able to penalize spatial filters which focus on the artifactual activity. Since the left hand motor imagery related activity is substantially more stationary than the signal at electrode FFC6, it is preferred by sCSP. Note that sCSP was able to penalize the artifactual activity in an unsupervised (data-driven) manner by minimizing nonstationarity. It is interesting to note that electrodes with high channel-wise variance (right panel in Figure 14) are preferably penalized by sCSP (see Figure 12).

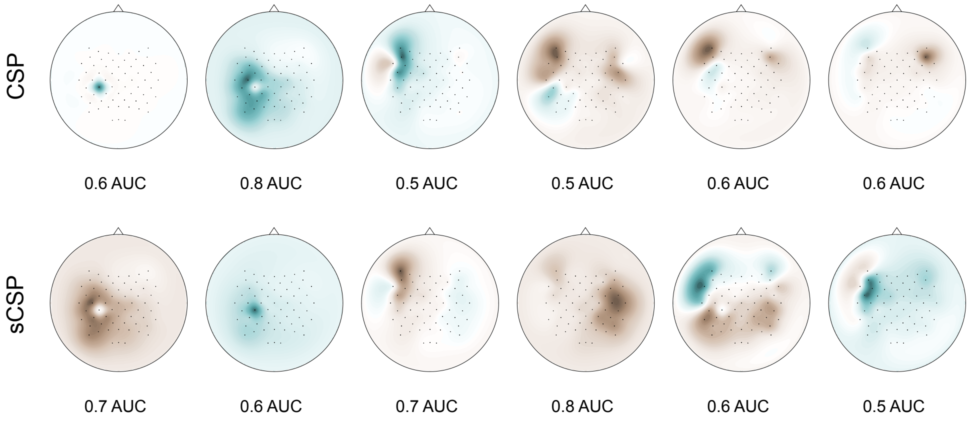

Figure 13.: *Top row*: Activation patterns of CSP. *Bottom row*: Activation patterns of sCSP.

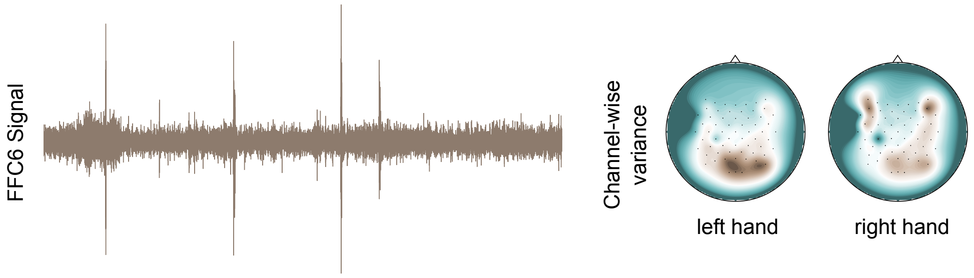

Figure 14.: *Left*: Signal at electrode FFC6 containing artifacts. *Right*: Scalp maps visualizing channel-wise variance for both motor imagery classes.

Lessons learned in this chapter

- Stationarity is an effective regularization target with intriguing properties.
- sCSP outperforms all baseline methods in terms of error rate and is superior to CSP even when using fixed parameters.
- Simultaneous optimization of two objectives is superior to the application of two-step approaches such as SSA+CSP.
- Selecting an appropriate chunk size, i.e., capturing nonstationarities on the right time scale, is important.
- The approximation applied by sCSP, namely flipping the sign of negative eigenvalues, works well in practice.

Future work

- Combination of sCSP with other regularization targets and robust estimators; stationarity on multiple time scales.
- Criterion for parameter selection; Bayesian approach.
- Evaluation of different measures of nonstationarity.
- Theoretical analysis of penalty matrix properties for approximately joint diagonalizable covariance matrices.
- Analysis of neural sources responsible for nonstationarity.

5 TRANSFERRING INFORMATION BETWEEN SUBJECTS

Incorporating data from additional subjects (or sessions) into the training process allows one to reduce calibration times and to construct subject-independent BCI systems (Krauledat, 2008; Fazli et al., 2009, 2011). Recently, several methods have been proposed to transfer information between users, e.g., by regularization of covariance matrices (Kang et al., 2009; Lotte and Guan, 2010b) or by constructing a common feature space (Devlaminck et al., 2011). The recordings acquired from additional sessions or subjects are helpful for improving the estimation of relevant parameters (especially if training data is scarce) and serve as source of prior information, e.g., for guiding the CSP algorithm towards regions of interest. However, large variations between subjects may prevent the BCI system from learning a common representation or classification model. Regions of interest may vary substantially from subject to subject and relying on information acquired from a user with very different signal characteristics may cause a deterioration in performance. Therefore many transfer learning approaches (Kang et al., 2009; Devlaminck et al., 2011; Samek et al., 2013a) weight the contribution of each additional subject or perform subject selection.

In this chapter we propose an alternative approach, namely rather than transferring information relating to regions of interest we transfer information relating to prominent nonstationarities in the data. We develop a method, *stationary subspace Common Spatial Patterns* (ssCSP), which does not regularize the CSP solution towards a region of interest but rather regularizes the spatial filters away from sources responsible for the changes in the signal. Conceptually, this novel approach is similar to sCSP, but the penalty matrix is estimated from recordings of additional subjects and has low rank.

5.1 WHAT INFORMATION CAN BE TRANSFERRED ?

Data from subjects performing the same experiment prove very helpful in estimating important parameters of the BCI pipeline. For instance, additional recordings may help in computing the spectral filter, may improve the estimation of the high-dimensional covariance matrices and help in extracting superior spatial filters or they may robustify the classifier training by providing additional examples of a particular motor imagery class. Incorporating brain recordings from additional subjects is only reasonable if these recordings have similar properties, abstractly speaking are sampled from the same distribution as the data at hand. Note that this assumption is very restrictive and will hardly be satisfied in practice. Besides differences in the electrode positions or head geometries, there often exist discrepancy in the users' mental states, attention levels or strategies to perform the motor imagery tasks. Several authors (Kang et al., 2009; Lotte and Guan, 2010b; Devlaminck et al., 2011; Samek et al., 2013a) recognized the subject heterogeneity problem and proposed to weight the contributions of individual subjects or to only include data from users that are "similar enough" to the subject of interest.

An alternative approach to reduce the risk of training the model on data which do not fit the subject of interest is to transfer information about what sources not to include instead of what regions to rely on. The underlying idea is that although users may have very different motor imagery patterns, some of their non-task related activity may still be similar and may be used for regularizing the CSP solution. In the case of ssCSP we propose to regularize the spatial filters away from a common nonstationary subspace rather than regularizing them towards a region of interest. Of course regularization towards and away from a subspace are two sides of the same coin because when regularizing away from a subspace we by definition regularize towards its orthogonal complement. However, when regularizing towards a small subspace, e.g., in the case of CSP we are typically interested in 6 out of say 60 dimensions, we discard a large amount ($\geqslant 90\%$) of information. Thus, if the regularization target is not the right one, we very likely lose relevant information. On the other hand when regularizing away from a small subspace we keep most of the information in the data and remove only a small fraction. Thus, if the regularization target is suboptimal (e.g. because subjects are very different) we lose little discriminative information on average.

There are several types of nonstationary information which may be transferred between subjects. For instance, a change in the experimental paradigm between sessions may be induced by additional feedback, differences in the cue indicating the stimulus (visual vs. auditory) or by changes in the task (movement execution vs. movement imagination vs. robot assisted movement). This between-session nonstationarity often leads to a shift in the feature distribution and may negatively affect classification accuracy (Shenoy et al., 2006; Krauledat, 2008; Samek et al., 2013c; Arvaneh et al., 2013b). This type of experimentally induced nonstationarity (if similar among subjects) may be transferred between users and may act as regularization target in the ssCSP algorithm. Potentially transferable is also the activity responsible for the day-to-day variability of EEG recordings. Patterns associated with most of this variability may be identified on data from additional subjects and the corresponding spatial filters may be penalized when performing experiments with the subject of interest. If the patterns are similar among subjects, then this step will increase the stationarity of the signal. Finally, one may also transfer information about common artifacts between users. Approaches which extract the most prominent artifact patterns in a data-driven manner on recordings of additional subjects and incorporate these pieces of information into ssCSP, may potentially regularize the spatial filter computation towards robust solutions and improve classification performance. In this thesis we only consider experimentally induced between-session changes as this type of nonstationarity is most stable between subjects (see Samek et al., 2013c).

5.2 COMMON NONSTATIONARY SUBSPACE

The ssCSP algorithm uses the regularized CSP framework for spatial filter computation. The penalty term of the proposed method is a quadratic form $P(\mathbf{w}) = \mathbf{w}^\top \mathbf{K}\mathbf{w}$ with a low rank matrix $\mathbf{K}$ specifying the nonstationary subspace. The low-rank assumption ensures that we only remove a small part of the information, optimally the most prominent common changes, from the data. In the following we describe how to compute the common nonstationary subspace from recordings of additional subjects.

In the first step we extract the dominant directions of change from every additional subject $k = 1 \ldots K$ by computing the eigendecomposition of the difference of the training and test covariance matrix $\mathbf{\Sigma}_{tr}^k - \mathbf{\Sigma}_{te}^k$. Note that the ℓ eigenvectors $\mathbf{v}_1^k, \mathbf{v}_2^k \ldots \mathbf{v}_\ell^k$ with largest absolute eigenvalues $|d_1^k|, |d_2^k| \ldots |d_\ell^k|$ capture most of the changes occurring between training and test phase. In other words a spatial filter $\mathbf{w}$ that is orthogonal to the subspace spanned by these eigenvectors will extract features with a relatively small between-session shift. Parameter ℓ may be a fixed value or may be individually determined for each subject, e.g., by setting a threshold on the eigenvalue spectrum.

In a second step we aggregate the eigenvectors into a matrix $\mathbf{V} = \left[\mathbf{v}_1^1 \ldots \mathbf{v}_\ell^K\right]$; the columns of this matrix span the subspace of common nonstationarities $\mathcal{S}_\mathbf{V} = \text{span}(\mathbf{V})$. The dimensionality of the subspace $\mathcal{S}_\mathbf{V}$ can be reduced by applying *Principal Component*

Analysis (PCA) to matrix $\mathbf{V}$. This step is important as the dimensionality of $\mathcal{S}_{\mathbf{V}}$ grows linearly (with factor ℓ) with the number of additional subjects. By applying PCA we extract a subspace of dimensionality $\delta \leqslant \mathrm{dim}(\mathcal{S}_{\mathbf{V}})$ that contains common nonstationary directions. We denote the basis of this low-dimensional subspace as $\mathbf{V}_\delta$. The penalty term of ssCSP is then

$$P(\mathbf{w}) \quad = \quad \mathbf{w}^\top \mathbf{V}_\delta \mathbf{V}_\delta^\top \mathbf{w}. \tag{26}$$

Spatial filters $\mathbf{w}$ that are orthogonal to the subspace $\mathcal{S}_{\mathbf{V}_\delta}$ are not penalized by the ssCSP algorithm. Note that PCA must be applied without mean subtraction as the column vectors of $\mathbf{V}$ are directional vectors without a common zero point. Figure 15 visualizes the problem of mean subtraction.

The goal of the PCA step is to approximate the information contained in $\mathbf{V}$ by a lower rank matrix $\mathbf{V}_\delta$. Each column vector of $\mathbf{V}$ is represented by a dark point in Figure 15. The dimensionality of the space is equal to the dimensionality of the data. Note that in this example the common direction of change (in signal space) is the direction represented by the center of the point cloud (brown cross). If we apply standard PCA to $\mathbf{V}$, we first center the data, i.e., shift the point cloud to the origin (light points), and then determine the direction of largest variance (left panel). Note that the direction of largest variance does not coincide with the direction of the common nonstationarity in signal space, thus, PCA with mean subtraction does not give the desired solution. The right panel in Figure 15 visualizes the application of PCA to the data and its mirrored copy $\tilde{\mathbf{V}} = [\mathbf{V}\ -\mathbf{V}]$, i.e., we mirror the dark point cloud and apply PCA to the pooled data. Note that the pooled data have mean zero, i.e., no centering is required. It is easy to show that the vector maximizing the variance of the pooled data, i.e., the eigenvector with largest eigenvalue of $[\mathbf{V}\ -\mathbf{V}][\mathbf{V}\ -\mathbf{V}]^\top$, coincides with the PCA solution when no mean is subtracted, i.e., the eigenvector with largest eigenvalue of $\mathbf{V}\mathbf{V}^\top$. This eigenvector gives us the direction of common nonstationarity.

In order to find the subspace of common nonstationarity we select the top δ eigenvectors of $\mathbf{V}\mathbf{V}^\top$. The matrix containing these vectors as columns is denoted as $\mathbf{V}_\delta$ and approximates $\mathbf{V}$ in the sense that

$$\mathbf{w}^\top \mathbf{V}_\delta \mathbf{V}_\delta^\top \mathbf{w} \text{ is large} \quad \Leftrightarrow \quad \mathbf{w}^\top \mathbf{V}\mathbf{V}^\top \mathbf{w} \text{ is large}$$

5.3 STATIONARY SUBSPACE CSP ALGORITHM

The goal of the stationary subspace CSP method is to remove the subspace that contains the principal nonstationary directions common to most subjects prior to CSP computation. The method is summarized in Algorithm 2.

The input parameters of the algorithm are the class covariance matrices (only training data) of the subject of interest $\{\boldsymbol{\Sigma}_c\}$, training and test covariance matrices $\{\boldsymbol{\Sigma}_{tr}^k\}$ and $\{\boldsymbol{\Sigma}_{te}^k\}$ of the additional subjects $k = 1 \dots K$, parameter d determining the number of spatial filters to return, parameters ℓ and δ specifying the common nonstationary subspace and the regularization parameter λ. In the first step of the algorithm all covariance matrices are normalized to have approximately the same scale; a commonly used approach is to divide the matrices by their traces. We select the parameters ℓ and δ from $\{0, 1, \dots, 10\}$. The first parameter controls the number of nonstationary directions extracted per subject. This parameter may have the same value for all subjects or be subject-specific, e.g., by defining a threshold on the amount of changes one wants to capture. The second parameter determines the rank of the ssCSP penalty matrix, i.e., the dimensionality of the nonstationary subspace $\mathcal{S}_{\mathbf{V}_\delta}$. Note that the parameters can not be determined by cross-validation on the subject of interest as the goal of our method is to reduce the shift between training and test data and this does not

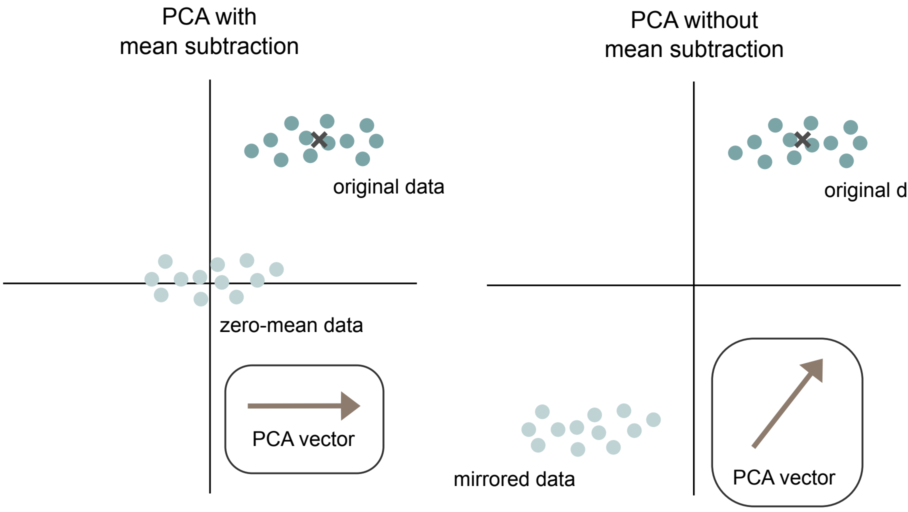

Figure 15.: *Left*: The application of PCA to the dark points provides the direction of largest variance within the point cloud (horizontal direction). This PCA direction does not coincide with the common nonstationary direction (dark cross). *Right*: When applying PCA to the data and its mirrored copy, then the direction of largest variance (PCA vector) captures the common change. Note that rather than using the mirrored copy we may also apply PCA to the original data without subtracting the mean.

necessarily correlate with a performance increase on the training data. Therefore, we determine the parameters by minimizing test error of the additional subjects.

In line 5 of Algorithm 2 we extract the ℓ eigenvectors with largest absolute eigenvalues of the difference matrix $\mathbf{\Sigma}_{tr}^{k} - \mathbf{\Sigma}_{te}^{k}$ for each additional subject k. In the subsequent steps the dimensionality of the common nonstationary subspace $\mathcal{S}_{\mathbf{V}} = \text{span}(\mathbf{V})$ is reduced to δ by applying PCA without mean subtraction, i.e., via SVD-based rank reduction of $\mathbf{V}$. This step ensures that the spatial filters $\mathbf{w}$ which lead to a large shift between training and test features are penalized. Note that we measure the shift in an unsupervised manner, i.e., we assume that it is not class-specific. One can apply the same algorithm to compute class-specific nonstationary subspaces. In the final steps we add the penalty matrix $\mathbf{\Delta} = \mathbf{V}_{\delta}\mathbf{V}_{\delta}^{\top}$ to the Rayleigh quotient. Note that $\mathbf{\Delta}$ has a rank $\delta << D$, i.e., spatial filters which are orthogonal to the common nonstationary subspace $\mathcal{S}_{\mathbf{V}_{\delta}}$ are not penalized. In contrast to sCSP we completely remove the nonstationary directions from the data by setting $\lambda = 10^5$. We use the same spatial filter selection scheme as for sCSP.

5.4 EXPERIMENTAL EVALUATION

In this section we evaluate the ssCSP method using simulations and real EEG recordings. We compare the performance to covCSP and klcovCSP, two methods which utilize information from additional subjects by regularization of the subject-specific covariance matrices. The experimental evaluation is performed on the three data sets described in Section 2.3.

5.4.1 *Simulations*

A common assumption of multi-subject methods used in BCI is that the EEG recordings of different users have the same underlying data generating process. For instance,

Algorithm 2 Stationary Subspace Common Spatial Patterns

```
function ssCSP($\{\boldsymbol{\Sigma}_c\}$, $\{\boldsymbol{\Sigma}^k_{tr}\}$, $\{\boldsymbol{\Sigma}^k_{te}\}$, $d$, $\ell$, $\delta$, $\lambda$)
    Normalize all covariance matrices.
    for all subjects $k = 1 \dots K$ do
        Compute $\ell$ top eigenvectors $\mathbf{v}^k_1 \dots \mathbf{v}^k_\ell$ of $\boldsymbol{\Sigma}^k_{tr} - \boldsymbol{\Sigma}^k_{te}$.
    end for
    Aggregate nonstationary directions of all subjects $\mathbf{V} = [\mathbf{v}^1_1 \dots \mathbf{v}^K_\ell]$.
    Compute a matrix $\mathbf{V}_\delta$ consisting of $\delta$ top eigenvectors of $\mathbf{V}\mathbf{V}^\top$.
    Compute penalty term $\boldsymbol{\Delta} = \mathbf{V}_\delta \mathbf{V}_\delta^\top$.
    Compute eigenvectors $\mathbf{V}_c = \mathrm{eig}(\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 + \lambda\boldsymbol{\Delta})$.
    Select d columns $\mathbf{W} \in \mathbb{R}^{D \times d}$ from $\mathbf{V}_1$ and $\mathbf{V}_2$.
    return W
end function
```

regularization of covariance matrices as done by covCSP and klcovCSP is only reasonable if there is a structural similarity between the covariance matrices of different users. This assumption is very restrictive and only holds approximately in practice. The simulation study described in this section analyses the stability of ssCSP under increasing dissimilarity between subjects and compares it to the stability of covCSP. In other words we evaluate the impact on classification performance when moving from transferring relevant information (subjects are similar) to transferring meaningless information (subjects are not similar).

The data set used for the evaluation consists of artificially generated training and test recordings of five subjects. We use a mixture model with (non)discriminative and (non)stationary sources. In order to separately study the effect of dissimilarity of the discriminative subspace and the nonstationary subspace, we generate the data as sum of two independent mixtures. More precisely, data $\mathbf{x}(t)$ is generated as a sum of a stationary noise-signal term and a nonstationary noise term

$$\mathbf{x}(t) \quad = \quad \underbrace{\mathbf{A}\begin{bmatrix}\mathbf{s}^{dis}(t)\\ \mathbf{s}^{ndis}(t)\end{bmatrix}}_{noise-signal\ term} \quad + \quad \underbrace{\mathbf{B}\begin{bmatrix}\mathbf{s}^{stat}(t)\\ \mathbf{s}^{nstat}(t)\end{bmatrix}}_{noise\ term}. \tag{27}$$

Note that we call the first mixture the "noise-signal term" as it contains contributions from sources which contain information about the simulated BCI task (signal) as well as contributions from nonrelevant sources (noise). The second mixture is termed "noise term" as its sources are not important for classification. Note that this model can also be written as mixture model in Eq. (4) with nonstationary noise. The matrices $\mathbf{A}$ and $\mathbf{B}$ are random rotation matrices projecting the source activity to channel space and the sources generate i.i.d. normally distributed (with zero mean) samples. In order to approximate the properties of real data we restrict the discriminative and nonstationary subspaces to be low-dimensional.

The following parameters are used for the experiments. The 6 discriminative sources $\mathbf{s}^{dis}$ are sampled from a zero mean Gaussian with standard deviation 0.8 in one condition and 0.1 in the other condition, whereas the 74 nondiscriminative sources $\mathbf{s}^{ndis}$ have standard deviation 0.1 irrespectively of class label. The 75 stationary sources $\mathbf{s}^{stat}$ have standard deviation 1 in both the training and test data set; the standard deviation of the 5 nonstationary sources $\mathbf{s}^{nstat}$ is 1 in the training data set and 3 in the test data set. For each artificial subject we generate 100 trials per condition, each consisting of 100 data points, for both the training and the test set. We extract three CSP filters per class, use log-variance features and a LDA classifier. The parameters of the multi-subject methods are determined by cross-validating classification performance in a leave-one-subject-out manner on the other users. The following experiments were performed on this toy data set using 50 repetitions.

In the first experiment we fix matrix $\mathbf{B}$ for all subjects, but increase the distance between the mixing matrix $\mathbf{A} = e^{\mathbf{M}}$ of subject 1 and the mixing matrices of the other subjects by adding an increasing amount of randomness[1]. By adding a random matrix Ξ to $\mathbf{M}$ we obtain $\mathbf{M}_2 = \mathbf{M} + \eta\Xi$. The new rotation matrix $\mathbf{A}_2$ is computed as $\mathbf{A}_2 = e^{\frac{1}{2}(\mathbf{M}_2 - \mathbf{M}_2')}$. The weight η controls the distance between $\mathbf{A}$ and $\mathbf{A}_2$. In other words we simulate the case of increasing dissimilarity between discriminative subspaces of subject 1 and the other artificial users. The results of the two multi-subject methods are summarized in the first column of Figure 16. Each boxplot visualizes the distribution of classification error rates of subject 1 for increasing dissimilarity values η. Furthermore, the median CSP error rate is plotted as brown curve. We see from the figure that the application of covCSP significantly decreases error rates when the dissimilarity between the mixing matrices $\mathbf{A}$ of subject 1 and the others is low; this performance gain is due to the transfer of discriminative information between subjects. However, if the information that is transferred becomes more and more random, then the performance deteriorates dramatically. The stationary subspace CSP method is not affected by increased dissimilarity of the mixing matrices $\mathbf{A}$ as it does not transfer discriminative information. It is able to improve classification performance as the nonstationary subspace remains the same for all subjects (i.e. matrix $\mathbf{B}$ is constant).

In the second experiment we simulate the opposite scenario, namely we fix $\mathbf{A}$ and increase the dissimilarity of $\mathbf{B}$ between subject 1 and subjects 2-5. The second column of Figure 16 depicts the results for this case. We observe a stable improvement of covCSP because the discriminative subspaces are the same for all subjects irrespectively of $\mathbf{B}$. The figure shows an improved performance (decrease in error rates) for the ssCSP method when the dissimilarity between the nonstationary subspaces is low and a performance drop when the dissimilarity is high. However, the important point here is that in contrast to the transfer of discriminative information in the last experiment the performance loss is minimal, actually the performance goes back to CSP level. This increased robustness can be explained with a lower risk of losing important information when regularizing the solution away from a small subspace. Although the transferred nonstationary information becomes more and more meaningless when distance between the mixing matrices $\mathbf{B}$ increases, classification accuracy does not decrease on average since only few directions are removed from data. The different behaviour of covCSP and ssCSP highly depends on the size of the discriminative and nonstationary subspaces, the selection of regularization parameters and of course if subject (pre)selection is used or not. Thus, on some data sets and in some settings, e.g., if training data is scarce, covCSP may perform very well whereas in other settings, e.g., in the presence of large nonstationarity between sessions, ssCSP may be the method of choice.

In the final experiment we generate both matrices $\mathbf{A}$ and $\mathbf{B}$ such that they are either different or the same for subject 1 and the other users (third column of Figure 16). In the first case multi-subject methods have no advantage over CSP as there is no meaningful information to be transferred. On the contrary, the method transferring discriminative information may even lose performance as the solution is regularized towards a noninformative subspace. In the other case when both subspaces stay constant over subjects we observe a significant performance gain of all multi-subject meth-

[1] Matrix $\mathbf{A}$ is constructed as a matrix exponent of a random antisymmetric matrix $\mathbf{M}$, i.e., $\mathbf{A} = e^{\mathbf{M}}$. This ensures that $\mathbf{A}$ is a rotation matrix, i.e., $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$ as $\mathbf{A}^\top = (e^{\mathbf{M}})^\top = e^{-\mathbf{M}} = \mathbf{A}^{-1}$.

ods. Since the nonstationarity problem is more severe than the estimation problem (in small sample settings this may be vice versa), we obtain best results for ssCSP.

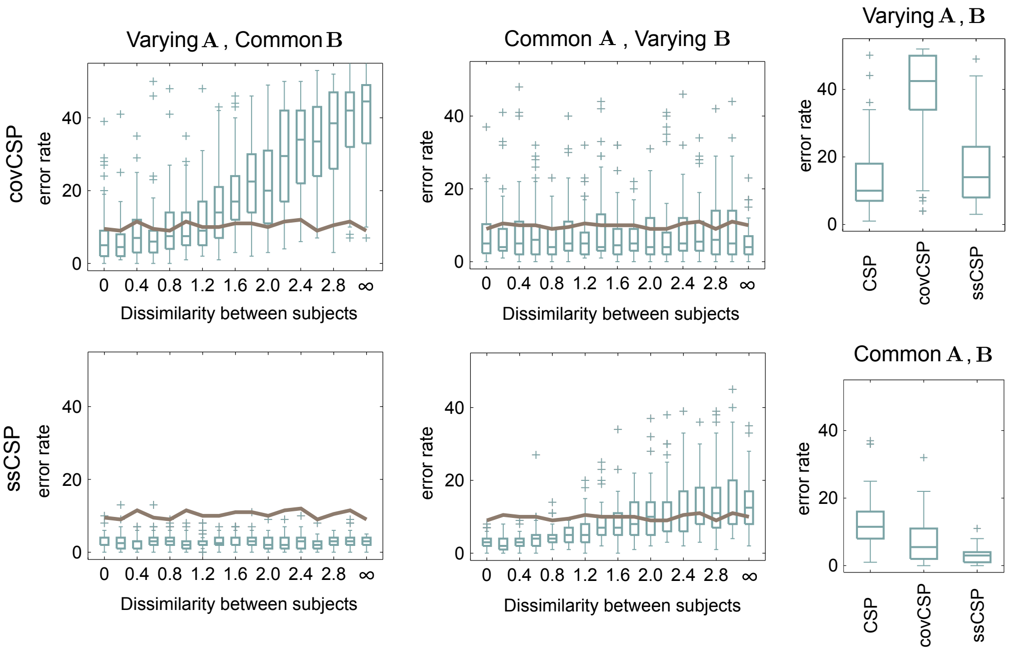

Figure 16.: *First column*: Results when discriminative subspaces become more and more dissimilar but the nonstationarities stay the same for all subjects. One can see that covCSP improves classification performance when subjects are similar, but when the difference between them becomes larger, then the information transferred becomes more and more meaningless and error rates increase almost to chance level. The ssCSP method improves classification accuracy as it penalizes nonstationary directions in the spatial filter computation and is not affected by differences in the discriminative subspaces. *Second column*: Results for the opposite case, namely constant discriminative subspaces but different nonstationary directions. The ssCSP method improves classification accuracy when the transferred information are meaningful, but does not lead to a significant increase in error rates when this is not the case. *Third column*: Performance of all methods for the case when both subspaces are either different or same for all subjects.

5.4.2 *Subspace Similarity*

In an initial analysis we study the similarity between signals recorded in different experimental runs and with different users. We quantify the similarity between two sets of recordings (indexed with i and j) by using symmetric Kullback-Leibler (KL) divergence. More precisely, we model the signals as zero-mean D-dimensional Gaussian distributions $\mathcal{N}(0, \boldsymbol{\Sigma}_i)$ and $\mathcal{N}(0, \boldsymbol{\Sigma}_j)$ and compute

$$\tilde{D}_{kl}\left(\mathcal{N}(0,\boldsymbol{\Sigma}_i) \parallel \mathcal{N}(0,\boldsymbol{\Sigma}_j)\right) = \frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\right)\right) + D. \quad (28)$$

The symmetric Kullback-Leibler divergence is a popular measure of discrepancy between probability distributions, the divergence is always positive (or zero) and low values indicate high similarity between the distributions.

Table 5.: This table displays the average symmetric Kullback-Leibler divergence between the recordings of different subjects (first and second row) and between the training and test data of the same subject.

	Inhouse data set					BCI Comp. data set					Vital data set
Divergence	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	Median
Subject-Subject (train)	490	799	650	853	657	205	344	373	254	498	1072.5
Subject-Subject (test)	995	1803	1799	1947	1377	286	483	468	291	600	925.5
Train-Test	62	27	57	110	15	6	7	7	6	14	11

Table 5 displays the similarity values for the three data sets used for experimental evaluation in this chapter. The first row displays the average symmetric KL divergence between the distributions estimated on training data of different subjects. For instance, the value in the first column is the average divergence between the training data distribution of subject A1 and the training data distributions of subjects A2-A5. The second row contains the same quantities for the test data sets whereas the last row displays the discrepancy between the training and test recordings of the same user. Note that the values can not be directly compared across the data sets because the dimensionality of the signal varies from data set to data set. However, a tendency can be observed in all three data sets, namely that the variations between subjects are up to two orders larger than the differences between training and test sessions. Although this large discrepancy demonstrates that transferring information between subjects is a very challenging task, one should keep in mind that the similarity values in Table 5 are computed in the full signal space, thus may in large part reflect differences in the subject-specific noise (task-unrelated activity) and be affected by the high dimensionality of the signal. In the "signal space", i.e., when only considering subsignals which are relevant for the BCI task, the between-subject dissimilarity may be much lower. In other words discrepancies in the full data space may not correlate with differences in the feature distribution after spatial filtering. It is interesting to note that the users with largest train-test discrepancy, subject A4 in the Inhouse data set and subject B5 in the BCI Competition data set, are users for which ssCSP provides a higher classification accuracy than the baseline methods (see Table 6). Furthermore, ssCSP reduces the train-test shift in the feature distributions by 8% for A4 and by 12% for B5. Unfortunately, we could not find a correlation between the reduction of train-test divergence and performance gain in the Vital BCI data set. In the following we move from an analysis in the full signal space to an analysis where we quantify the similarity between relevant subspaces of different subjects.

Figure 17 depicts the average similarity values between different subspaces for the Inhouse (left panel), BCI Competition (middle panel) and Vital BCI (right panel) data set. We measure similarity as the mean of squared cosines of the principal angles θ_k between the subspaces[2]. This quantity corresponds to the amount of energy preserved when projecting data from one subspace to the other, thus, higher values indicate closer subspaces. Considering all principal angles gives a clearer picture of the relation between two subspaces than when restricting the analysis to the largest principal angle as the latter one tends to become 90° very fast. In Figure 17 four curves are depicted in each panel. The cyan curve with lowest color intensity represents the

[2] Principal angles are defined recursively as $\cos(\theta_k) = \max_{\mathbf{u} \in \mathcal{F}} \max_{\mathbf{v} \in \mathcal{G}} \mathbf{u}^\top \mathbf{v} = \mathbf{u}_k^\top \mathbf{v}_k$ subject to $\|\mathbf{u}\| = \|\mathbf{v}\| = 1, \quad \mathbf{u}^\top \mathbf{u}_i = 0, \quad \mathbf{v}^\top \mathbf{v}_i = 0, \quad i = 1, \ldots, k-1$. Note that there exist an equality between the canonical correlation and the cosine of principal angles.

similarity between subspaces spanned by CSP filters, the curve with medium color intensity stands for similarity values between the corresponding CSP patterns, the curve with highest color intensity depicts the similarity between subspaces constructed from the prominent nonstationary directions between training and test data and the black dashed line represents similarity between random subspaces. Note that higher values indicate closer subspaces. From the plot we see that the subspaces spanned by the CSP filters of different users have an average similarity value which is close to the similarity between random subspaces. We conjecture that the reason for this low similarity is that CSP filters capture subject-specific noise (see discussion in Section 3.1). The similarity is much higher for the CSP patterns as they represent the underlying neural activity and are less affected by the noise. Interestingly, the nonstationary subspaces show the highest similarity values for all three data sets, i.e., the directions of largest change between training and test data are very similar between different subjects. This important result is the main motivation of our ssCSP method.

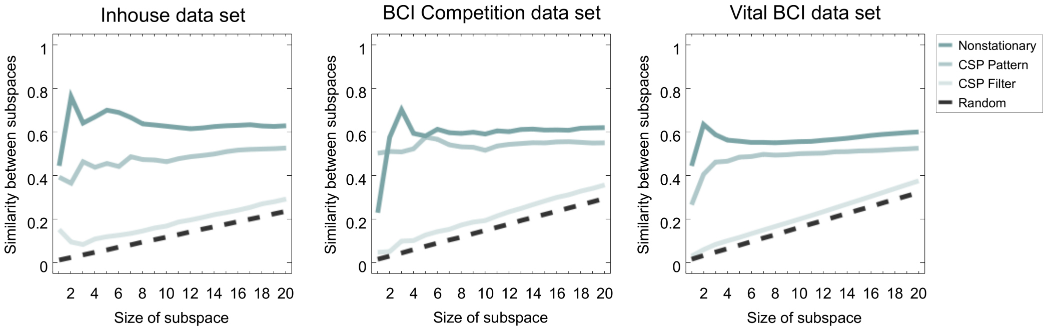

Figure 17.: Similarity between subspaces measured as mean of squared cosines of the principal angles. For all three data sets the similarity between the nonstationary subspaces is significantly higher than the similarity between subspaces spanned by CSP filters or CSP patterns.

5.4.3 *Performance Results*

Although we have demonstrated that common nonstationary directions are present in the data sets, we have not yet shown that removing these directions improves classification accuracy. In fact, the impact of nonstationarity on classification performance is not straight forward and largely depends on whether nonstationary directions are discriminative or do not contain class-related information, whether they are parallel or orthogonal to the separating hyperplane and whether they are considered by CSP or do not affect the feature distribution. In last chapter we have seen that regularizing CSP towards (within-session) stationarity significantly improves classification accuracy. In the following we evaluate the ssCSP method which regularizes CSP towards stationary subspaces estimated on additional users. Note that the results presented in this chapter differ from the results in (Samek et al., 2013c) because we use a different spatial filter and parameter selection scheme in order to make all results presented in this thesis comparable to each other.

Table 6 summarizes the performance results for the Inhouse and the BCI Competition data set. We see that almost all subjects benefit from incorporating data of additional users. Our novel ssCSP algorithm significantly improves classification accuracy over CSP according to the Wilcoxon signed-rank test, however, it is not able to outperform covCSP and klcovCSP. As mentioned before ssCSP has a different focus than covCSP and klcovCSP, namely it tackles the nonstationarity problem and not the

Table 6.: Comparison of classification accuracies for different multi-subject CSP variants. The last column displays the p-values of the one-sided Wilcoxon signed-rank test when comparing ssCSP with the method in the row.

	Audio-Visual Data Set					BCI Competition III					Overall		
Methods	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	Mean	Median	p-value
CSP	79.5	80.0	**69.2**	79.2	**94.2**	66.1	**94.6**	58.2	87.9	90.5	79.9	79.8	**0.0234**
covCSP	**89.4**	**84.2**	**69.2**	81.7	**94.2**	66.1	**94.6**	**73.0**	87.1	92.1	**83.1**	**85.6**	0.7109
klcovCSP	88.6	72.5	68.3	79.2	**94.2**	66.1	**94.6**	68.9	**90.2**	89.7	81.2	83.9	0.4727
ssCSP	78.8	81.7	**69.2**	**82.5**	**94.2**	**69.6**	**94.6**	58.7	88.8	**92.9**	81.1	82.1	–

estimation problem. Note that we have included covCSP and klcovCSP as baseline methods in order to see which problem is more relevant in practice; our primarily goal is not to outperform these approaches with ssCSP. Thus, it is not surprising that the classification accuracy of some users such as A1, A2 and B3 significantly improves when covCSP is applied whereas other subjects such as A4, B1 and B5 benefit from the application of ssCSP. From Table 6 one also sees that in contrast to klcovCSP (subject A2) there is no significant decrease in performance when applying the ssCSP method. This observation is in line with the results from the toy experiment, however, this property of course largely depends on the size of the common nonstationary subspace penalized by ssCSP.

Figure 18 depicts the results when applying the three multi-subject methods to the Vital BCI data set. Each circle represents the error rate of a subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner. On this data set our ssCSP algorithm improves the classification accuracy for few subjects, but the improvement is not significant over all 80 users. Thus, although the nonstationary subspaces are similar between users (see Figure 17) we can not obtain a significant performance increase over the CSP baseline. However, when restricting the analysis to subjects who have an error rate of 30% or larger, then our method significantly outperforms CSP with $p = 0.0232$. There are several potential explanations for the relatively small performance improvement of ssCSP. First, the nonstationarities present in the Vital BCI data differ from the train-test changes in the other two data sets. In the case of Vital BCI there is additional feedback present in the test session but the cue presenting the stimulus does not change between the sessions. For the other two data sets a different cue is used in training and test phase but no feedback is provided to the user. Potentially, the latter type of nonstationarity may be better transferable between users or may have more impact on performance. Second, the Vital BCI data have a higher between-subject variability (see Table 5) and are potentially more noisy as all users are BCI novices. Finally, the number of additional subjects used for the computation of the common nonstationary subspace is significantly higher in the Vital BCI data set. Since we do not differentiate between real nonstationarity and artifact-related changes, a larger number of users with lots of artifacts may largely affect the computation of the common nonstationary subspace. We believe that advanced clustering methods may perform better than our PCA without mean subtraction algorithm in such a scenario. On the other hand the covCSP and klcovCSP methods may benefit from incorporating data from many additional user because this averages out individual differences. We investigate this point in the subsequent subsection. In fact, we see that covCSP and klcovCSP largely improve classification accuracy, klcovCSP is even significantly better than ssCSP according to the one-sided Wilcoxon signed-rank test. However, note that covCSP and klcovCSP not only solve a different problem and include different data than ssCSP but the parameters are also selected differently. In the case of covCSP and klcovCSP we select the parameters by minimizing the cross-validation error on the subject of interest whereas in the case of ssCSP we select the parameters by minimizing average test error of the additional subjects (as we assume that test data of the subject of interest is not available). We believe that more advanced parameter selection schemes (e.g., based on the eigenvalue spectrum), subject selection and a more elaborate approach to the computation of the common nonstationary subspace (e.g., clustering method) may improve the performance of ssCSP on the Vital BCI data set.

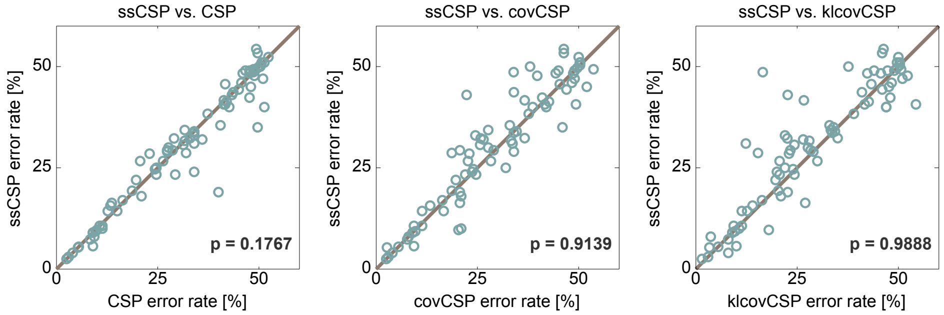

Figure 18.: Scatter plots showing error rates of ssCSP and three baseline methods. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner.

5.4.4 *Further Analysis*

In the following we analyse and interpret the ssCSP results in more detail. First, we evaluate the ability of ssCSP to reduce the shift in feature distribution by transferring information from additional users. Figure 19 displays the relative change in nonstationarity for all five users of the Inhouse data set and all regularization parameters of ssCSP. More precisely, we compute the symmetric KL divergence $\tilde{D}$ between the training and test feature distribution for CSP and for ssCSP with parameters $\varphi = \{\ell, \delta\}$ and plot the following relative change for all subjects and parameters

$$\kappa = \frac{\tilde{D}^{\varphi}_{sscsp} - \tilde{D}_{csp}}{\tilde{D}_{csp}}$$

In Figure 19 we see that the relative change in nonstationarity is negative for all subjects except A3, i.e., the feature distribution becomes more stationary when applying ssCSP. Since the algorithm penalizes nonstationary directions without actually "seeing" the test data of the subject of interest, we may conclude information about nonstationarity is really transferred between subjects. Figure 19 also shows that a reduction of nonstationarity does not necessarily correlate with an increase in classification accuracy. For instance, the reduction of nonstationarity is largest for subject A1, however, this user does not improve classification accuracy in Table 6. The CSP error rate of this subject is 20.5%. When applying ssCSP and using the parameters selected by minimizing test error of other users (result in Table 6) we obtain an error rate of 21.2% and $\kappa = -46\%$. When using the parameters which minimize nonstationarity in Figure 19 we obtain an error rate of 24.2% and $\kappa = -88\%$. The minimal ssCSP test error for this subjects is 8.3% and $\kappa = -68\%$. Thus, reduction of the between-session shift does not necessarily correspond to a classification improvement.

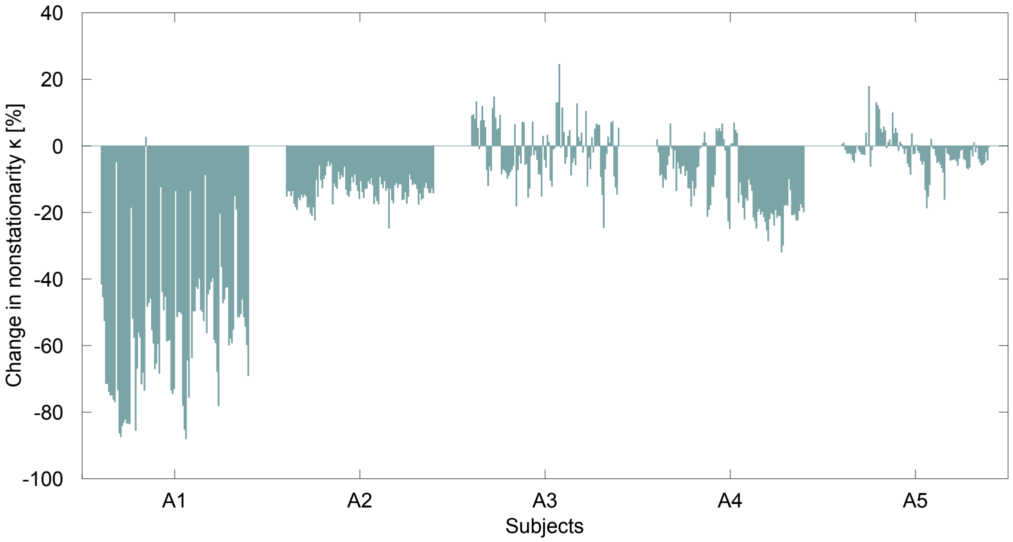

Figure 19.: Relative change in the train-test nonstationarity for subjects A1-A5 and all ssCSP parameters. A negative value corresponds to a relative decrease in nonstationarity compared to CSP.

Figure 20 visualizes the change between the training and test features of subject A1 for CSP and ssCSP (with parameters minimizing the shift). We plot the two feature dimensions that best visualize the shift between training (cyan circles) and test (brown crosses) data. In the case of CSP the feature distribution obtained from training data contains several outliers and consequently largely differs from the distribution estimated on test data. On the other hand when applying ssCSP there is only little difference between both distributions. Note that this increase in stationarity is obtained by regularizing the spatial filter computation away from the common nonstationary subspace, i.e., without using test data of the subject of interest.

In Figure 21 we visualize the five most nonstationary directions of subjects A1-A5. One can see that the nonstationarity patterns of different users are highly similar. This similarity is also reflected in the results in Figure 17. The nonstationarity patterns show mainly activity in the occipital and temporal regions. We believe that the activity in the occipital region reflects the change in the visual processing between training and test session, i.e. the transition from a visual mode of stimulus presentation to an auditory one. Since the visual cortex is highly involved in visual processing we expect to see changes related to visual processing in the occipital region. The activity in the temporal regions may be related to auditory processing or be due to muscle activity.

In the following we analyse the results of the Vital BCI data set in more detail. Figure 22 depicts the relative changes in nonstationarity (κ values) for all subjects and ssCSP parameters. Note that we plot the same quantity as in Figure 19 but as histogram instead of plotting it for each subject separately. Although the histogram covers more area on the negative side, i.e., most κ values are negative, it is much more balanced than the results for the Inhouse data set. The fact that regularization towards the common nonstationary subspace often increases the shift between training and test session rather than reducing it is certainly one explanation for the limited performance improvement of ssCSP.

In the last part of this section we analyse the results of subject 13 from the Vital BCI data set. This subject has the second largest improvement in classification accuracy over CSP in Figure 18. We do not consider the subject with the largest improvement

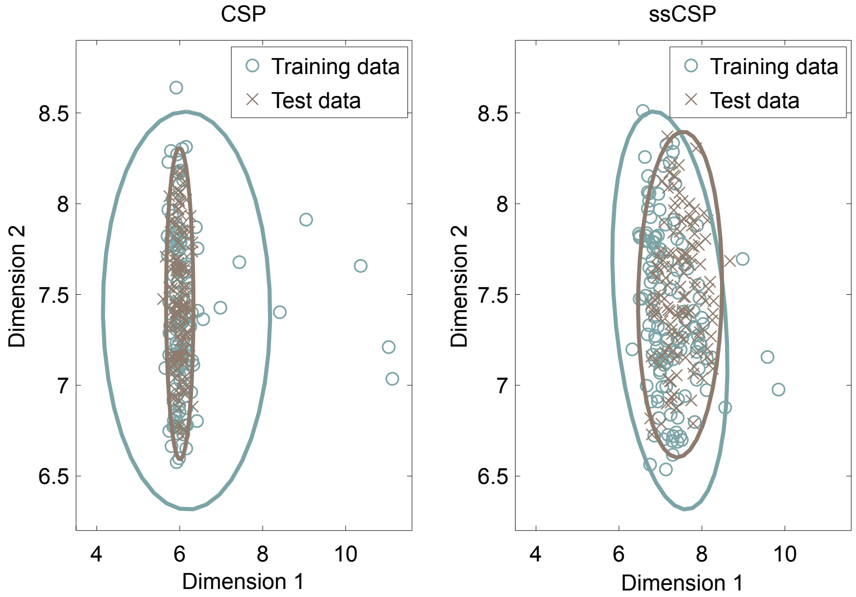

Figure 20.: *Left*: Visualization of the two most discriminative dimensions for subject A1. A significant change in the feature distribution between training (cyan circles) and test (brown crosses) can be observed for CSP. *Right*: When applying ssCSP the change in feature distribution becomes almost negligible.

(user 21) as we have already performed an analysis for this subject in last chapter. The error rates of subject 13 are 50% when applying CSP and 35% for the ssCSP method. Figure 23 visualizes the activation patterns of CSP and ssCSP with the corresponding area under the ROC curve (AUC) value. The third pattern is the most interesting one as the AUC value is larger for ssCSP than for CSP. Qualitatively, there is no large difference between the ssCSP solution and the pattern computed by CSP except that the ssCSP pattern has less activation in the frontal and occipital region than the CSP pattern. Thus, the features computed by CSP contain more occipital activity than the ssCSP features. Since occipital activity is often related to visual processing and the occipital alpha rhythm, these features are more nonstationary.

In Figure 24 we plot the feature values corresponding to the third patterns, i.e., we visualize the third dimension of the feature vectors after applying CSP and ssCSP, respectively. The cyan dots represent training trials whereas the brown dots stand for test trails. In order to increase the comparability of the results we normalize the feature distribution such that the training points have mean zero and variance one. From the figure one sees that the CSP features exhibit a significant shift between training and test stage, i.e., the distribution of the cyan and brown points changes drastically. Since the CSP filter captures activity from the occipital region, the additional visual processing that is induced by the feedback provided in the test session is directly affecting the feature distribution. On the other hand when applying ssCSP we obtain a substantially more stationary feature distribution with no significant changes between training and test session. Thus, a small difference in the patterns in Figure 23, i.e., less weight on occipital and frontal area, largely influences the stationarity of the feature.

Finally, we come back to the issue raised in the previous subsection, namely that covCSP largely benefits from the incorporation of data of a large number of additional subjects. The left panel in Figure 25 compares the classification performance of covCSP for subject 37 when utilizing data from one additional subject (thin curves) and incorporating data from all 29 additional subjects performing the same motor imagery task (cyan curve). Note that the thick brown line is the median of the thin curves. One can

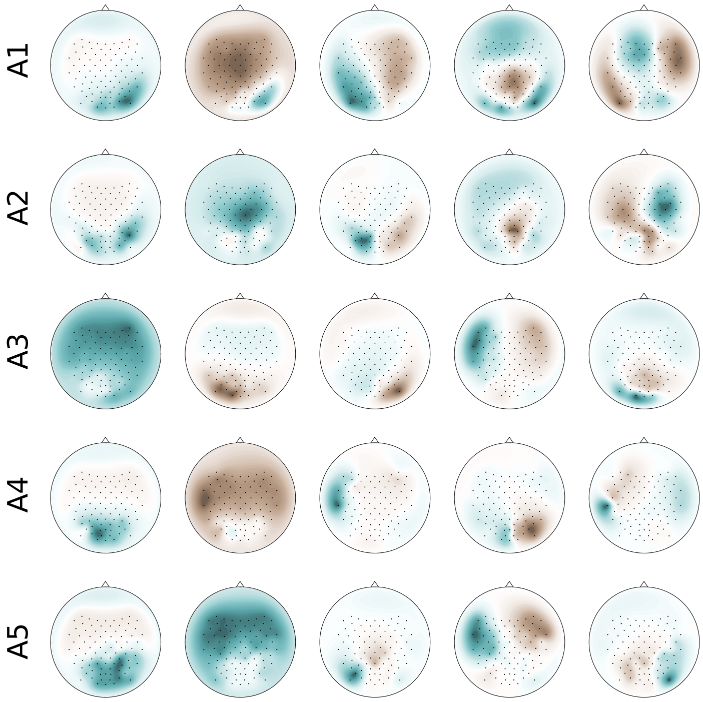

Figure 21.: Visualization of five most nonstationary directions for each subject.

see from the figure that using the average of all the 29 covariance matrices from the additional subjects as regularization target in covCSP is better than when using the data of each individual subject separately (cyan line below thin lines). The right panel of Figure 25 displays the error rates when incorporating data from $k = 1 \dots 29$ additional subjects. Note that we depict the median error rate over 100 randomly selected subsets of k subjects. Also here we see a clear tendency towards large regularization parameters and a large number of subjects. Thus, covCSP benefits from using many additional subjects. We could not find a similar averaging effect for the ssCSP method.

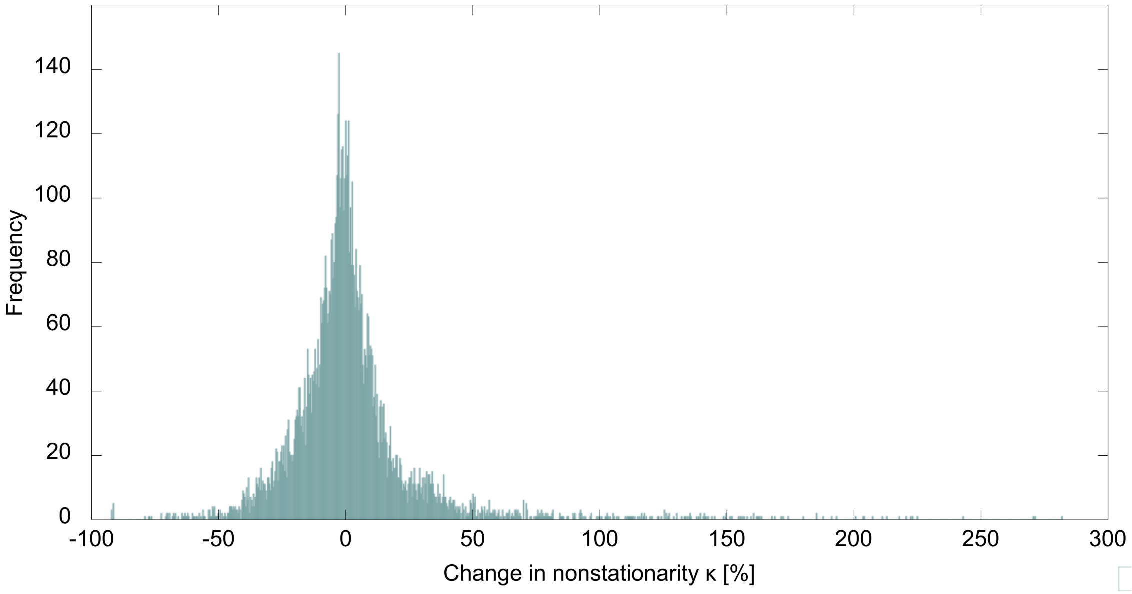

Figure 22.: Histogram of relative changes in nonstationarity over all subjects and ssCSP parameters for the Vital BCI data set.

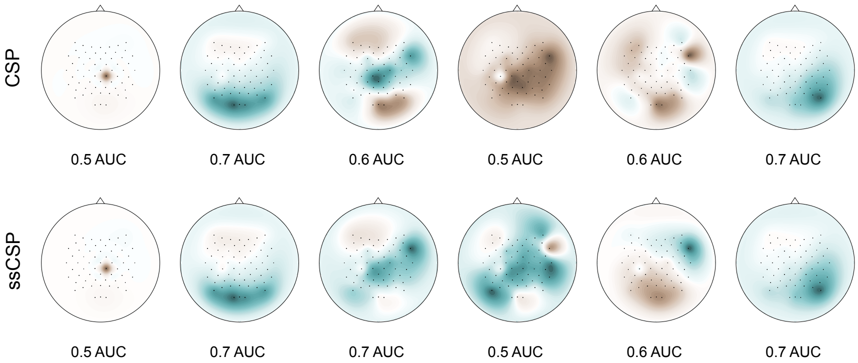

Figure 23.: *Top row*: Activation patterns computed by CSP. *Bottom row*: Activation patterns computed by ssCSP.

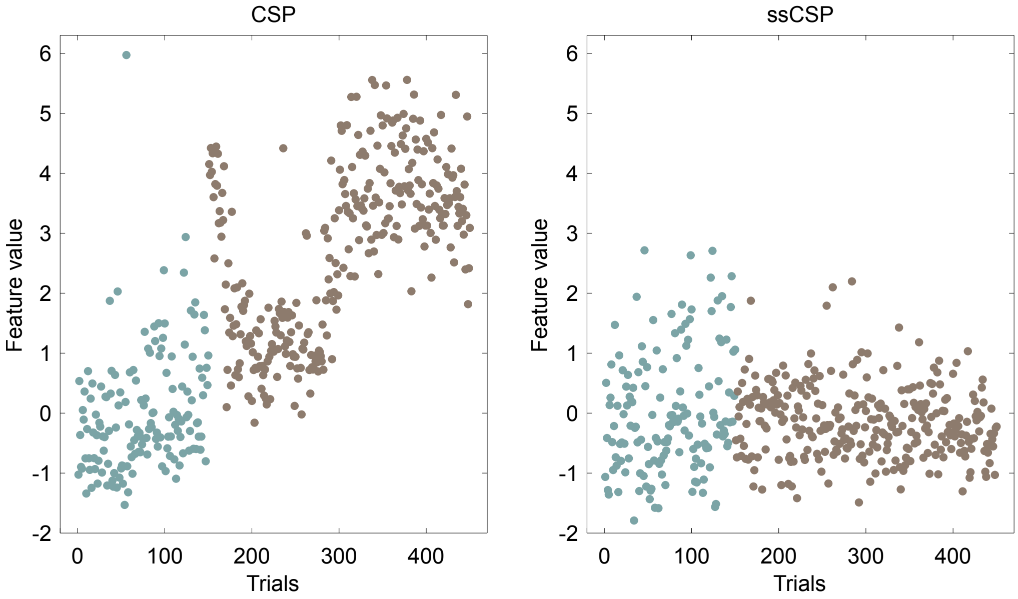

Figure 24.: *Left*: Normalized feature values of training (cyan) and test (brown) trials for CSP. *Right*: Normalized feature values for ssCSP.

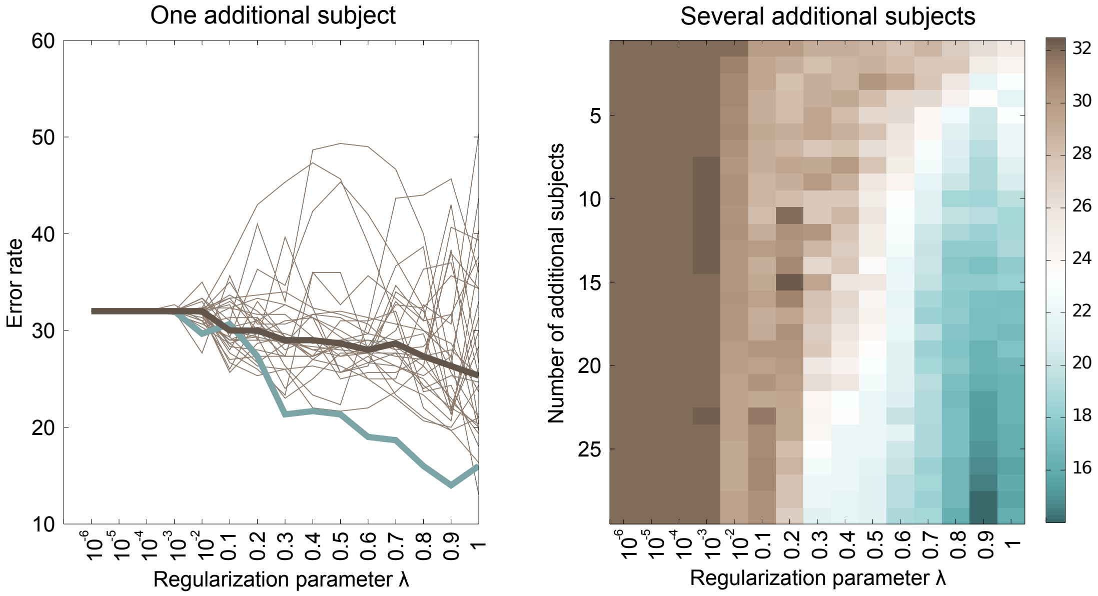

Figure 25.: *Left*: Error rates of covCSP when regularizing the covariance matrices of subject 37 towards the average covariance matrix of the 29 additional subjects (cyan line) or the individual covariance matrices (thin lines). The thick brown line is the median of the thin lines. *Right*: Error rates of covCSP when incorporating information from $k = 1 \dots 29$ additional users.

Lessons learned in this chapter

- Transferring subspaces between subjects is challenging but may improve performance and reduce nonstationarity.
- Complex relationship between stationarity and performance.
- ssCSP is robust because removing a small random subspace from data does not significantly increase error rate.
- Performance of covCSP scales with the number of additional users; the estimation of the nonstationary subspace does not necessarily improve when using many subjects.
- Nonstationary directions are neurophysiologically interpretable.

Future Work

- Investigation of other similarity measures between nonstationary subspaces; statistical test for presence or absence of common nonstationarities.
- Parameter selection scheme that does not require labeled data and is not based on the performance of the other users.
- Clustering algorithm for computation of common nonstationary subspace from subject-specific nonstationary subspaces.
- Evaluation of soft regularization strategies for ssCSP; comparison to complete removal of nonstationary subspace.
- Normalization and coregistration of data sets.

Part III

INFORMATION GEOMETRIC METHODS

6 ROBUST COVARIANCE MATRIX ESTIMATION

In the final part of this thesis we use methods and concepts from the field of *information geometry* (Amari and Nagaoka, 2000) for robustifying the feature extraction process in BCI. These methods tackle the nonstationarity and robustness problem in a principled manner and are grounded in a solid mathematical framework. This chapter introduces a novel robust covariance matrix estimator which is particularly tailored to BCI application because it downweights the influence of artifactual trials in the estimation process. Note that in contrast to classical robust covariance estimators we propose to measure "outlierness" on the time scale of trials, not with respect to individual samples. This is a very natural time scale for robustification in BCI, e.g., when whole trials are affected by an artifact. After introducing the tools needed to derive the estimator we show how to use it in order to compute robust CSP filters. Finally, we evaluate its performance using simulations and real EEG recordings.

6.1 A BRIEF INTRO INTO INFORMATION GEOMETRY

Information geometry (Amari and Nagaoka, 2000; Amari, 2010) is a branch of mathematics which studies questions of probability theory by means of differential geometry. It represents probability distributions as points on a manifold $\mathcal{M}$ and uses divergence functions to quantify the discrepancy between them. A divergence function induces a geometry on the manifold and has specific properties; we will utilize the properties of a particular divergence for robust spatial filter computation. A very simple example of $\mathcal{M}$ is the manifold of the one-dimensional Gaussian distribution with mean μ and variance σ^2. This particular class of probability distributions forms a two-dimensional manifold with coordinates μ and σ. In other words any Gaussian distribution can be represented by a point $\mathbf{z} = (\mu, \sigma)$ on this manifold. A function $D(\mathbf{z}_1 \parallel \mathbf{z}_2)$ between two points $\mathbf{z}_1$ and $\mathbf{z}_2$ on the manifold is termed *divergence function* (Amari and Cichocki, 2010; Cichocki and Amari, 2010) when

(1) $D(\mathbf{z}_1 \parallel \mathbf{z}_2) \geqslant 0$ for all $\mathbf{z}_1, \mathbf{z}_2$.

(2) $D(\mathbf{z}_1 \parallel \mathbf{z}_2) = 0$ if and only if $\mathbf{z}_1 = \mathbf{z}_2$.

(3) For infinitesimally small differences between $\mathbf{z}$ and $\mathbf{w} = \mathbf{z} + d\mathbf{z}$ the Taylor expansion $D(\mathbf{z} \parallel \mathbf{w}) = \mathbf{z}^\top \mathbf{G}(\mathbf{z})\mathbf{z}$ is a positive definite quadratic form with

$$g_{ij}(\mathbf{z}) = \frac{\partial^2}{\partial z_i \partial z_j} D(\mathbf{z} \parallel \mathbf{w})_{|\mathbf{w}=\mathbf{z}}$$

being the i, jth element of $\mathbf{G}(\mathbf{z})$.

Note that in general a divergence is neither symmetric, i.e., $D(\mathbf{z}_1 \parallel \mathbf{z}_2) \neq D(\mathbf{z}_2 \parallel \mathbf{z}_1)$, nor does it satisfy the triangular inequality, i.e., $D(\mathbf{z}_1 \parallel \mathbf{z}_2) \nleqslant D(\mathbf{z}_1 \parallel \mathbf{z}_3) + D(\mathbf{z}_3 \parallel \mathbf{z}_2)$. Depending on the divergence function a specific Riemannian metric $g_{ij}(\mathbf{z})$ is induced on the manifold $\mathcal{M}$.

An important quantity in differential geometry is the geodesic. A *geodesic* is a curve connecting two points on the manifold by preserving the tangent vector along the curve. Thus, a geodesic can be regarded as a generalization of the notion of a straight line. In order to define a geodesic we need to connect nearby tangent spaces by *affine* *connections*. The author of (Eguchi, 1983) proposed (dually coupled[1]) affine connections which are computed from the divergence function

$$\Gamma_{ijk}(\mathbf{z}) = -\frac{\partial^3}{\partial z_i \partial z_j \partial w_k} D(\mathbf{z} \parallel \mathbf{w})_{|\mathbf{w}=\mathbf{z}} \tag{29}$$

$$\Gamma^*_{ijk}(\mathbf{z}) = -\frac{\partial^3}{\partial w_i \partial w_j \partial z_k} D(\mathbf{z} \parallel \mathbf{w})_{|\mathbf{w}=\mathbf{z}} \tag{30}$$

If both connectors vanish, i.e., $\Gamma_{ijk} = \Gamma^*_{ijk} = 0$, then the induced manifold is termed *dually flat*. The class of *Bregman divergences* (Bregman, 1967; Murata et al., 2004) induces dually flat manifolds. A Bregman divergence between probability distributions $p(x)$ and $q(x)$ is defined as

$$D_\varphi(p \parallel q) = \int (\varphi(p) - \varphi(q) - (p - q)\varphi'(q))dx, \tag{31}$$

with φ being a strictly convex differentiable function and φ' being its derivative. Note that we abuse the notation at this point as we abbreviate $p(x)$ by p and ignore that the input parameter x may be multidimensional. Bregman divergences have several useful properties, e.g., they satisfy the generalized Pythagorean theorem and projection theorem (see Amari, 2010). The *Kullback-Leibler divergence* or KL divergence is an example of a Bregman divergence

$$D_{kl}(p \parallel q) = \int p \log \frac{p}{q} dx, \tag{32}$$

with $\varphi(x) = x\log(x)$. This divergence is used in many applications, has an information theoretic interpretation (MacKay, 2002) and is related to maximum likelihood estimation in statistics (see next section). Another popular divergence is the *beta divergence* proposed in (Basu et al., 1998; Eguchi and Kano, 2001). The beta divergence between distributions $p(x)$ and $q(x)$ is defined (for $\beta > 0$) as

$$D_\beta(p \parallel q) = \frac{1}{\beta}\int (p^\beta - q^\beta)p dx - \frac{1}{\beta+1}\int (p^{\beta+1} - q^{\beta+1})dx. \tag{33}$$

The next theorem shows that beta divergences belong to the class of Bregman divergences.

Theorem 3. *The beta divergence defined in Eq. (33) is a Bregman divergence with*

$$\varphi_\beta(x) = \frac{1}{\beta(\beta+1)}x^{\beta+1} - \frac{1}{\beta}x + \frac{1}{\beta+1}.$$

[1] Duality is a concept from information geometry that goes beyond the scope of this thesis. For more information we refer to (Amari and Nagaoka, 2000).

Proof. This theorem is a well-known result. We provide the following proof for it.

$$\begin{aligned}
D_{\varphi_\beta}(p \parallel q) &= \int (\varphi_\beta(p) - \varphi_\beta(q) - (p-q)\varphi'_\beta(q))dx \\
&= \int \Bigg(\frac{1}{\beta(\beta+1)}p^{\beta+1} - \frac{1}{\beta}p + \frac{1}{\beta+1} - \frac{1}{\beta(\beta+1)}q^{\beta+1} + \frac{1}{\beta}q \\
&\quad - \frac{1}{\beta+1} - (p-q)\left(\frac{1}{\beta}q^\beta - \frac{1}{\beta}\right)\Bigg) dx \\
&= \int \left(\frac{1}{\beta(\beta+1)}p^{\beta+1} - \frac{1}{\beta(\beta+1)}q^{\beta+1} - \frac{1}{\beta}pq^\beta + \frac{1}{\beta}q^{\beta+1} \right) dx \\
&= \int \left(\frac{1}{\beta}p^{\beta+1} - \frac{1}{\beta+1}p^{\beta+1} + \frac{1}{\beta+1}q^{\beta+1} - \frac{1}{\beta}pq^\beta \right) dx \\
&= \frac{1}{\beta}\int (p^\beta - q^\beta)p dx - \frac{1}{\beta+1}\int (p^{\beta+1} - q^{\beta+1})dx = D_\beta(p \parallel q)
\end{aligned}$$

□

The following theorem shows that beta divergence and Kullback-Leibler divergence coincide as β approaches zero.

Theorem 4. *The beta divergence defined in Eq. (33) converges to the Kullback-Leibler divergence defined in Eq. (32) when $\beta \longrightarrow 0$, i.e.,*

$$\lim_{\beta \to 0} D_\beta(p \parallel q) = D_{kl}(p \parallel q)$$

Proof. This theorem is a well-known result. We provide the following proof for it.

$$\begin{aligned}
\lim_{\beta \to 0} D_\beta(p \parallel q) &= \lim_{\beta \to 0} \frac{\int \left((\beta+1)(p^\beta - q^\beta)p - \beta(p^{\beta+1} - q^{\beta+1})\right) dx}{\beta(\beta+1)} \\
&\overset{*}{=} \lim_{\beta \to 0} \Bigg[\frac{\int \left(((p^\beta - q^\beta)p + (\beta+1)(p^\beta \log(p) - q^\beta \log(q))p\right) dx}{2\beta+1} \\
&\quad - \frac{\int \left((p^{\beta+1} - q^{\beta+1}) + \beta(p^{\beta+1}\log(p) - q^{\beta+1}\log(q))\right) dx}{2\beta+1} \Bigg] \\
&= \int (p\log(p) - p\log(q) + q - p)\, dx = D_{kl}(p \parallel q)
\end{aligned}$$

Note that the equality $(*)$ holds due to the l'Hôpital's rule and $\int(q-p)dx = 0$ because $p(x)$ and $q(x)$ are probability distributions which sum to 1. □

From this perspective beta divergence can be regarded as a generalization of the KL divergence. It has a robustness property which will be utilized by the covariance estimator proposed in this chapter and by the novel divergence-based CSP framework introduced in the next chapter. In previous work beta divergence has been used for robustifying algorithms such as Independent Component Analysis (ICA) (Mihoko and Eguchi, 2002) and Non-negative Matrix Factorization (NMF) (Févotte and Idier, 2011). Another well known special case of beta divergence emerges for $\beta = 1$; for this β value beta divergence coincides with the *Euclidean divergence* which is defined as

$$D_{euc}(p \parallel q) = \frac{1}{2}\int (p-q)^2 dx. \tag{34}$$

The following theorem shows the relation.

Theorem 5. *The beta divergence defined in Eq. (33) converges to the Euclidean divergence defined in Eq. (34) for* $\beta = 1$*, i.e.,*

$$D_{\beta=1}(p \parallel q) = D_{euc}(p \parallel q).$$

Proof. This theorem is a well-known result. We provide the following proof for it.

$$\begin{aligned} D_{\beta=1}(p \parallel q) &= \int (p - q)p dx - \frac{1}{2}\int (p^2 - q^2) dx \\ &= \frac{1}{2}\int (p^2 - 2pq + q^2) dx \\ &= \frac{1}{2}\int (p - q)^2 dx = D_{euc}(p \parallel q) \end{aligned}$$

□

Note that D_{euc} induces an Euclidean space, i.e., $g_{ij} = \delta_{ij}$ where δ_{ij} is the Kronecker delta. Thus, this divergence is symmetric. We will use symmetric variants of Kullback-Leiber divergence and beta divergence extensively in the next chapter. A common way to symmetrize a divergence is to use

$$\tilde{D}(p \parallel q) = D(p \parallel q) + D(q \parallel p). \tag{35}$$

In the following we denote symmetric divergences by using the tilde symbol. Note that besides Bregman divergences there are other classes of divergences (e.g. f-divergences, α-divergences) which induce a specific geometrical structure, but since we do not use them in this work we skip the discussion here (see Cichocki and Amari, 2010).

6.2 Ψ-LIKELIHOOD PRINCIPLE

Reliable computation of covariance matrices is of crucial importance in BCI. The problem can be formulated as estimation of a parameter $\mathbf{\Sigma}$ of a statistical model, e.g., zero-mean Gaussian distributions $p(\mathbf{x}; \mathbf{\Sigma})$, given observations $\mathcal{D} = \{\mathbf{x}_i : i = 1 \dots n\}$. A standard procedure to estimate this parameter is to maximize the log-likelihood function $\mathcal{L}(\mathbf{\Sigma} \mid \mathcal{D})$ of the parameter given observations

$$\mathcal{L}(\mathbf{\Sigma} \mid \mathcal{D}) = \log\left(\prod_{i=1}^{n} p(\mathbf{x}_i; \mathbf{\Sigma})\right) = \sum_{i=1}^{n} \ell(\mathbf{x}_i; \mathbf{\Sigma}). \tag{36}$$

Note that for the zero-mean Gaussian case the log-likelihood function is

$$\ell(\mathbf{x}; \mathbf{\Sigma}) = -\frac{1}{2}(\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}) - \frac{1}{2} D \log(2\pi) - \frac{1}{2}\log(|\mathbf{\Sigma}|), \tag{37}$$

where D is the dimensionality of the data. It can be shown (Bishop, 2006) that the maximum likelihood estimate of a covariance matrix under the Gaussian model is

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma}}{\operatorname{argmax}} \mathcal{L}(\mathbf{\Sigma} \mid \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top. \tag{38}$$

In practice the maximum likelihood (ML) estimator may provide suboptimal solutions because it is not robust (Huber, 1981) in the sense that single outlier observations $\mathbf{x}_i$ may dominate the estimation. The left panel in Figure 26 illustrates a situation in which a single outlier sample significantly alters the estimate of the covariance matrix. Assume there is an outlier sample $\mathbf{x}_i$ so that $\mathbf{x}_i^\top \mathbf{\Sigma}^{-1}\mathbf{x}_i$ is very large for the true parameter $\mathbf{\Sigma}$. Then, the likelihood term of $\mathbf{x}_i$ is much smaller than the other likelihood terms, i.e., $\ell(\mathbf{x}_i;\mathbf{\Sigma}) << \sum_{j=1;j\neq i}^{n}\ell(\mathbf{x}_j;\mathbf{\Sigma})$. Let $\hat{\mathbf{\Sigma}}$ be another parameter with $\sum_{j=1;j\neq i}^{n}\ell(\mathbf{x}_j;\hat{\mathbf{\Sigma}}) < \sum_{j=1;j\neq i}^{n}\ell(\mathbf{x}_j;\mathbf{\Sigma})$ and $\ell(\mathbf{x}_i;\hat{\mathbf{\Sigma}}) \approx 0$. Then, the maximum likelihood estimator will prefer parameter $\hat{\mathbf{\Sigma}}$ over $\mathbf{\Sigma}$ because

$$\sum_{j=1,j\neq i}^{n}\ell(\mathbf{x}_j;\mathbf{\Sigma}) + \ell(\mathbf{x}_i;\mathbf{\Sigma}) \quad << \quad \sum_{j=1,j\neq i}^{n}\ell(\mathbf{x}_j;\hat{\mathbf{\Sigma}}) + \ell(\mathbf{x}_i;\hat{\mathbf{\Sigma}}) \qquad (39)$$

Thus, rather than learning a parameter that fits most of the data points the maximum likelihood estimator will provide an estimate which minimizes $\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i$. In other words the maximum likelihood estimator will focus primarily on the outlier sample. One can see in the right panel of Figure 26 that the function

$$f_{ML}(\mathbf{x}_i) \quad = \quad \frac{\ell(\mathbf{x}_i;\mathbf{\Sigma})}{\sum_{j=1;j\neq i}^{n}\ell(\mathbf{x}_j;\mathbf{\Sigma})} \qquad (40)$$

grows with $\|\mathbf{x}_i\|$, i.e., the influence of the outlier sample is not bounded.

The authors of (Basu et al., 1998; Eguchi and Kano, 2001) introduced the concept of Ψ-likelihood to perform robust estimation. The Ψ-likelihood function is defined as

$$\mathcal{L}_\Psi(\mathbf{\Sigma} \mid \mathcal{D}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\Psi(\ell(\mathbf{x}_i;\mathbf{\Sigma})) \; - \; b_\Psi(\mathbf{\Sigma}), \qquad (41)$$

$$\text{with} \quad b_\Psi(\mathbf{\Sigma}) \quad = \quad \int \Psi^*(\ell(\mathbf{x};\mathbf{\Sigma}))d\mathbf{x} \quad \text{and} \quad \Psi^*(\mathbf{z}) \quad = \quad \int_0^{\mathbf{z}} \exp(s)\frac{\partial}{\partial s}\Psi(s)ds.$$

Intuitively the method provides robust estimates as it limits the influence of each observation by applying the function Ψ to the likelihood values $\ell(\mathbf{x}_i;\mathbf{\Sigma})$. Note that $b_\Psi(\mathbf{\Sigma})$ denotes the normalization constant. If $\Psi(\mathbf{z}) = \mathbf{z}$ is the identity function then the maximum Ψ-likelihood estimator reduces to the ML estimator. When applying the Ψ function defined in Eq. (44) to the above example, then the influence of the outlier $\mathbf{x}_i$ is limited, i.e., the function $f_\Psi(\mathbf{x}_i)$ is bounded by a constant C

$$f_\Psi(\mathbf{x}_i) \quad = \quad \frac{\Psi(\ell(\mathbf{x}_i;\mathbf{\Sigma}))}{\sum_{j=1;j\neq i}^{n}\Psi(\ell(\mathbf{x}_j;\mathbf{\Sigma}))} \quad < \quad C \qquad (42)$$

The authors of (Eguchi and Kano, 2001) proved the equality between Ψ-likelihood maximization and Ψ-divergence minimization. The Ψ-*divergence* between distributions $p(x)$ and $q(x)$ is defined as

$$\begin{aligned} D_\Psi(p \parallel q) \quad &= \quad \int p\Psi(\log(p))dx - \int \Psi^*(\log(p))dx \qquad (43) \\ &\quad - \int p\Psi(\log(q))dx + \int \Psi^*(\log(q))dx \end{aligned}$$

The following theorem relates the minimization of the Ψ-divergence between the empirical distribution $p(x)$ estimated from the data at hand and a model distribution $q(x)$ with parameter $\mathbf{\Sigma}$ to the maximization of the Ψ-likelihood.

Theorem 6. *The maximization of the Ψ-likelihood is equivalent to the minimization of Ψ-divergence between the empirical and the model distribution*

$$\operatorname*{argmax}_{\mathbf{\Sigma}} \mathcal{L}_\Psi(\mathbf{\Sigma} \mid \mathcal{D}) \quad = \quad \operatorname*{argmin}_{\mathbf{\Sigma}} D_\Psi(p, q(\cdot,\mathbf{\Sigma}))$$

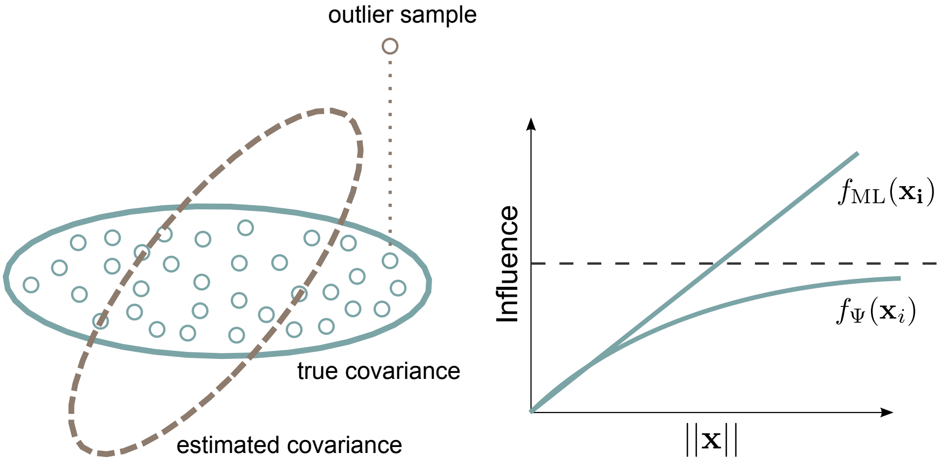

Figure 26.: *Left*: A single outlier sample may change the estimated covariance matrix drastically when relying on the maximum likelihood estimator. *Right*: The influence function of the maximum likelihood estimator is unbounded, i.e., a single term $\ell(\mathbf{x}_i; \mathbf{\Sigma})$ may dominate the sum in Eq. (36). In the case of maximum Ψ-likelihood the impact of a single term $\psi(\ell(\mathbf{x}_i; \mathbf{\Sigma}))$ is bounded, hence it provides robust estimates in the presence of outliers.

Proof. A similar proof can be found in (Eguchi and Kano, 2001). Note that the first two terms of Eq. (43) are constant in q, therefore the right side reduces to

$$\underset{\mathbf{\Sigma}}{\operatorname{argmin}} \left(- \int p\Psi(\log(q))dx + \int \Psi^*(\log(q))dx \right) = \underset{\mathbf{\Sigma}}{\operatorname{argmax}} \left(E_p[\Psi(\ell(\mathbf{x}; \mathbf{\Sigma}))] - b_\Psi(\mathbf{\Sigma}) \right)$$

This is the Ψ-likelihood function when the empirical expectation is taken over the data set. Thus, the maximization of $\mathcal{L}_\Psi$ is equivalent to Ψ-divergence minimization. Note that for $\Psi(\mathbf{x}) = \mathbf{x}$ the Ψ-divergence reduces to Kullback-Leibler divergence. Thus, the standard likelihood maximization reduces to KL divergence minimization which is a well-known result (Bishop, 2006). □

In this thesis we use a special choice of Ψ, namely

$$\Psi_\beta(z) = \frac{\exp(\beta z) - 1}{\beta}, \tag{44}$$

with a parameter β. It can be shown easily that for this choice of Ψ the Ψ-divergence reduces to β-divergence as defined in Eq. (33).

Theorem 7. *For $\Psi = \Psi_\beta$ the Ψ-divergence defined in Eq. (43) reduces to beta divergence defined in Eq. (33).*

Proof. This theorem is a well-known result. We provide the following proof for it.

$$\begin{aligned}
&D_\Psi(p \,||\, q)\\
&= \int p\Psi_\beta(\log(p))dx - \int \Psi^*_\beta(\log(p))dx - \int p\Psi_\beta(\log(q))dx + \int \Psi^*_\beta(\log(q))dx\\
&= \frac{1}{\beta}\int p(p^\beta - 1)dx - \frac{1}{\beta+1}\int p^{\beta+1}dx - \frac{1}{\beta}\int p(q^\beta - 1)dx + \frac{1}{\beta+1}\int q^{\beta+1}dx\\
&= \int \left(\frac{1}{\beta}p^{\beta+1} - \frac{1}{\beta+1}p^{\beta+1} + \frac{1}{\beta+1}q^{\beta+1} - \frac{1}{\beta}pq^\beta \right) dx = D_\beta(p \,||\, q)
\end{aligned}$$

□

Thus, minimization of β-divergence provides robust parameter estimates due to the limited influence of outlier samples.

6.3 Minimum beta divergence estimator

The authors of (Eguchi and Kano, 2001) showed that by using an iteratively reweighted algorithm the Ψ-divergence can be minimized, i.e., the Ψ-likelihood can be maximized. The estimate of the $\boldsymbol{\Sigma}$ parameter in the $(k+1)$th step is given by the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\psi(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}^{(k)}))S(\mathbf{x}_i;\boldsymbol{\Sigma}) \quad = \quad E[\psi(\ell(\mathbf{x};\boldsymbol{\Sigma}))S(\mathbf{x};\boldsymbol{\Sigma})], \tag{45}$$

where $\psi(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}\Psi(\mathbf{x})$ and $S(\mathbf{x};\boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\Sigma}}\ell(\mathbf{x};\boldsymbol{\Sigma})$. Note that $E[\cdot]$ denotes the expectation over the whole input space. For the Gaussian distribution and for Ψ as defined in Eq. (44), i.e., the beta divergence case, one can compute the covariance matrix parameter $\boldsymbol{\Sigma}$ by using the following iteration

Theorem 8. *Under the zero-mean Gaussian model the parameter $\boldsymbol{\Sigma}$ can be computed iteratively (see Eguchi and Kano, 2001) as*

$$\boldsymbol{\Sigma}^{(k+1)} \quad = \quad \frac{\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}^{(k)}))\mathbf{x}_i\mathbf{x}_i^\top}{\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}^{(k)})) \; - \; \beta/(\beta+1)^{D/2+1}}. \tag{46}$$

Note that $\boldsymbol{\Sigma}^{(k)}$ denotes the estimate of the parameter in kth step and

$$\psi_\beta(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}^{(k)})) \quad = \quad \frac{\partial}{\partial \mathbf{x}_i}\Psi_\beta(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}^{(k)})) \quad = \quad e^{-\frac{1}{2}\beta\mathbf{x}_i^\top(\boldsymbol{\Sigma}^{(k)})^{-1}\mathbf{x}_i} \tag{47}$$

is a factor downweighting the influence of outlier samples $\mathbf{x}_i$.

Proof. See appendix A.1 □

From the formula we can see that if the sample $\mathbf{x}_i$ is an outlier, i.e., it is very unlikely that it has been generated by a Gaussian with parameter $\boldsymbol{\Sigma}^{(k)}$, then its influence on the update of the parameter is very small due to vanishing $\psi_\beta(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}^{(k)}))$. Since $\psi_\beta(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}^{(k)}))$ is a monotonically decreasing function (for $\beta > 0$), it limits the influence of extreme outliers (see Figure 26). Note that for $\beta = 0$ this estimator reduces to the standard maximum likelihood estimator $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top$, i.e., all samples have uniform weight and f_Ψ reduces to a line i.e., is unbounded (see Figure 26).

Since BCI data do not consist of a time series of consecutive samples but have trial structure, the weights should not be applied to individual samples but rather to groups of samples (or trials). Robustness with respect to trials is more natural as whole trials are usually affected by artifacts, e.g., if the subject fails to properly imagine a movement then the whole trial is affected. In this work we propose a novel estimator that does not downweight individual EEG samples but rather reduces the influence of whole outlier trials. We use the Ψ-likelihood framework for robustly estimating the covariance matrix, however, in contrast to the example discussed in (Eguchi and Kano, 2001) we do not use the Gaussian distribution model but propose to minimize the beta divergence between the empirical distribution of scatter matrices and a model Wishart distribution with parameter $\boldsymbol{\Sigma}$. A *Wishart distribution* q is defined as

$$q(\mathbf{S};\boldsymbol{\Sigma},\nu) \quad = \quad \frac{1}{2^{\frac{\nu D}{2}}|\boldsymbol{\Sigma}|^{\frac{\nu}{2}}\Gamma_D\left(\frac{\nu}{2}\right)}|\mathbf{S}|^{\frac{\nu-D-1}{2}}\exp\left\{-\mathrm{tr}\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right\}, \tag{48}$$

where $\mathbf{S} = \sum_{t=1}^{N} \mathbf{x}_t \mathbf{x}_t^\top$ is the scatter matrix and Γ_D is the multivariate gamma function defined as

$$\Gamma_D\left(\frac{\nu}{2}\right) = \pi^{\frac{D(D-1)}{4}} \prod_{j=1}^{D} \Gamma\left[\frac{\nu}{2} + \frac{(1-j)}{2}\right] \tag{49}$$

$$\text{with} \quad \Gamma[t] = \int_0^\infty y^{t-1} e^{-y} dy. \tag{50}$$

From the estimated trial covariance matrices $\left\{\boldsymbol{\Sigma}_i \in \mathbb{R}^{D \times D} : i = 1 \dots n\right\}$ we compute the scatter matrices $\mathbf{S}_i$ and treat them as samples of an unknown Wishart distribution with parameters $\boldsymbol{\Sigma}$ and ν. Note that $\boldsymbol{\Sigma}$ denotes the true covariance matrix which we want to estimate and ν (under the assumptions that the samples are i.i.d.) equals the number of samples within a trial N (which is fixed). The maximum likelihood estimator for the Wishart distribution is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{\nu n} \sum_{i=1}^{n} \mathbf{S}_i, \tag{51}$$

or equivalently it is the average covariance matrix. We robustly estimate a covariance matrix $\hat{\boldsymbol{\Sigma}}$ from the scatter matrices $\mathbf{S}_i$ of trials $i = 1 \dots n$ by minimizing beta divergence using the following iterative algorithm.

Theorem 9. *Under the Wishart model the parameter $\boldsymbol{\Sigma}$ can be computed iteratively as*

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\sum_{i=1}^{n} \psi_\beta\left(\ell\left(\mathbf{S}_i; \boldsymbol{\Sigma}^{(k)}, \nu\right)\right) \mathbf{S}_i}{\nu \sum_{i=1}^{n} \psi_\beta\left(\ell\left(\mathbf{S}_i; \boldsymbol{\Sigma}^{(k)}, \nu\right)\right) - \gamma |\boldsymbol{\Sigma}^{(k)}|^{\frac{(\nu-D-1)\beta}{2}}}$$

where

$$\psi_\beta\left(\ell(\mathbf{S}; \boldsymbol{\Sigma}, \nu)\right) = |\mathbf{S}|^{\frac{(\nu-D-1)\beta}{2}} \exp\left\{-\mathrm{tr}\left(\frac{\beta}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right\}$$

is a factor downweighting the influence of outlier trials and

$$\gamma = \frac{n\beta(D+1)\Gamma_D\left(\frac{\nu(\beta+1)}{2} - \frac{(D+1)\beta}{2}\right)}{2^{\frac{\nu D}{2}} \Gamma_D\left(\frac{\nu}{2}\right)(\beta+1)} \left(\frac{2}{\beta+1}\right)^{\frac{\nu D(\beta+1)}{2} - \frac{D(D+1)\beta}{2}}$$

Proof. See appendix A.2 □

Note that for $\beta = 0$ this estimator gives the maximum likelihood solution in Eq. (51). The robustly estimated covariance matrices are used for spatial filter computation with CSP (see Algorithm 3). We denote the algorithm as *beta divergence Wishart CSP* (β-WishartCSP) method. When using the covariance estimator with a Gaussian model, we term the algorithm as *beta divergence Gaussian CSP* (β-GaussCSP). The input parameters of the β-WishartCSP algorithm are a set of trial covariance matrices $\{\boldsymbol{\Sigma}_c^i\}$, the number of spatial filters to return d, the robustness parameter β and a parameter ν capturing the uncertainty in the estimation of the trial covariance matrices. Note that ν equals the number of samples in a trial (under i.i.d. assumption), thus is fixed. After estimating the class covariance matrices (see Theorem 9) with our novel estimator, we use CSP to compute the spatial filters. Of course other CSP variants such as sCSP may also be used with the robustly estimated covariance matrix in order to combine regularization with robust estimation. Furthermore, note that we may change the time scale of robustness very easily by adapting ν and using different scatter matrices as samples. For instance, instead of computing robust estimates on the trial-by-trial basis we could apply the same algorithm to scatter matrices computed on larger time scales, e.g., multiple trials, thus obtain robust estimates on a larger group of samples. Note that the β-GaussCSP algorithm has the same structure as Algorithm 3, but applies Theorem 8 for estimation of the class covariance matrices.

Algorithm 3 Beta Divergence Wishart Common Spatial Patterns

```
function β-WishartCSP($\{\boldsymbol{\Sigma}_c^i\}$, d, β, ν)
    Compute a set of scatter matrices $\{\mathbf{S}_c^i\}$ from $\{\boldsymbol{\Sigma}_c^i\}$.
    Compute average class covariance matrices $\boldsymbol{\Sigma}_c^{(1)}$.
    for $k = 1 \dots k_{max}$ do
        Compute new estimate $\boldsymbol{\Sigma}_c^{(k+1)}$ by using Theorem 9.
    end for
    Normalize class covariance matrices.
    Compute eigenvectors $\mathbf{V} = \text{eig}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$
    Select d columns $\mathbf{W} \in \mathbb{R}^{D \times d}$ from $\mathbf{V}$.
    return W
end function
```

6.4 EXPERIMENTAL EVALUATION

In this section we evaluate the performance of the proposed β-WishartCSP algorithm and investigate its advantages and limitations using simulations and real EEG recordings from the Vital BCI and BCI Competition data set. We use CSP, shrinkCSP and MCDE+CSP as baselines and also compare the results to β-GaussCSP. Note that although we use β-GaussCSP as baseline, the method is novel in the sense that (to the best of our knowledge) it has not been used for spatial filter computation before. Since beta divergence estimators (with a Gaussian or Wishart model) have principled advantages and disadvantages over estimators such as the minimum covariance determinant estimator (MCDE) used by MCDE+CSP, we will not limit our analysis to β-WishartCSP but also investigate the properties of β-GaussCSP.

6.4.1 *Simulations*

Before we evaluate the covariance matrix estimators on real EEG recordings, we investigate their advantages and limitations using simulations. For this purpose we generate data $\mathbf{x}(t)$ using the following mixture model

$$\mathbf{x}(t) = \mathbf{A} \begin{bmatrix} \mathbf{s}^{dis}(t) \\ \mathbf{s}^{ndis}(t) \end{bmatrix} + \xi, \qquad (52)$$

where $\mathbf{A} \in \mathbb{R}^{10 \times 10}$ is a random orthogonal mixing matrix, $\mathbf{s}^{dis}$ is a discriminative source sampled from a zero mean Gaussian with variance 1.8 in one condition and 0.2 in the other condition, $\mathbf{s}^{ndis}$ are 9 sources with variance 1 in both conditions and ξ is a noise variable sampled from an isotropic Gaussian with standard deviation 2. In order to investigate the influence of artifacts we add, with varying probability, outlier trials to the data. Note that the whole trial is either affected by artifacts or not; we do not make this decision on sample level. We investigate the following two scenarios:

(1) *Large variance artifacts*: We contaminate the data with artifactual trials which are generated by the model in Eq. (52) but with the noise ξ not being isotropic (i.e. some directions have large variance). Furthermore, the distribution of the noise changes from trial to trial. More precisely, the noise ξ is sampled from a Gaussian with a diagonal covariance matrix. The diagonal elements of this matrix are resampled in every trial; with probability $1-p$ the standard deviation of a dimension is 2 and with probability p it is 10. Thus, the probability that a trial is not affected by the large variance artifacts is $(1-p)^{10}$.

(2) *Small variance artifacts*: We contaminate the data with artifactual trials which consist of Gaussian noise with standard deviation 0.1, i.e., they are not generated by the model in Eq. (52). Note that artifactual trials do not contain task-related information and have small variance. With probability $1-p$ a trial is sampled from the model in Eq. (52) and with probability p it is an artifactual trial.

For each of the two scenarios we generate 100 trials per condition, each trial contains 200 ten-dimensional samples, and repeat the experiment 100 times. Since the goal is to find a spatial filter that recovers the discriminative source $\mathbf{s}^{\mathrm{dis}}$, we quantify the quality of the solution by using the angle to the true projection. A small angle represents a good solution. Note that in the first scenario we aim to simulate artifacts which are due to loose electrodes (large changes in individual channels), whereas in the second scenario the outlier trials lack task-related information, i.e., we aim to simulate situations in which the subject fails to perform a given task (no class information in signal).

Figure 27 displays the median (over 100 repetitions) angles between the true projection and the filter computed by CSP (purple line), MCDE+CSP (green line), β-GaussCSP (brown line) and β-WishartCSP (cyan line). The values on the x-axis represent the probability of artifact p. Note that we did not include shrinkCSP as this method is not designed to be robust against artifacts but rather to perform well in small-sample settings. Since the number of trials and the dimensionality of the signal are not the critical parameters in our simulation study, shrinkCSP has a very similar angle error curve as CSP. Note that the free parameters of MCDE+CSP, β-GaussCSP and β-WishartCSP are chosen a posteriori, i.e., we compare the methods using the best parameters in order to reveal the full potential of each method.

The left panel of Figure 27 displays the results of the first simulation study (scenario 1). One can see that the performance of CSP quickly degrades as the probability of artifact increases. Since the artifactual trials dominate the estimated covariance matrices (because they have much larger variance), CSP focuses on the between-class differences of these artifactual events rather than extracting the true discriminative activity (which has much smaller variance). The MCDE+CSP method on the other hand is very robust until the point when the probability of artifact exceeds 7.5%. Note that the probability that a trial is not affected by an artifact is $(1-p)^{10}$, thus for $p = 7.5\%$ this is approximately 46%. In other words 54% of the trials (on average) contain at least in one dimension an artifact. The MCDE implementation[2] used in this thesis assumes that the proportion of outliers in the data is less than 50%. Therefore, it is not very surprising that the MCDE+CSP performance decreases when the proportion of outliers exceeds this 50% limit. We did not investigate whether it is possible and mathematically reasonable to relax this assumption. The estimators based on beta divergence minimization perform much better in this simulation study. Note that these estimators do not have such a 50% outlier limit, but (theoretically) always extract the correct parameter, even when the outlier ratio is very high, as long as the initialization is “close enough” to the true parameter and there is enough support for this parameter in the data. In our opinion the local optimally property of the beta divergence estimators (Mollah et al., 2010) is a clear advantage over the heuristic used by MCDE

[2] http://users.jyu.fi/~samiayr/DM/demot/LIBRA/mcdcov.m

as it (1) does not enforce a strict limit on the proportion of outliers in the data and (2) is more flexible and adaptive, e.g., we may run the algorithm with various initializations and chose the solution with desired properties. In this thesis we always initialize β-GaussCSP and β-WishartCSP with the average covariance matrices, do not consider multiple initializations and set k_{max} to 50. Although the average covariance matrices are not very informative when the probability of artifact exceeds 0.05 (see CSP performance), the β-GaussCSP and β-WishartCSP perform very well in our simulation study. This indicates that initialization is not critical for the performance of these methods.

The angle error curve of β-WishartCSP has an interesting shape, namely it is below the curves of all other methods until $p = 0.2$ (i.e., β-WishartCSP outperforms the baselines until this point) and then it quickly grows until it reaches CSP level. In the following we explain this behaviour. Note that at $p = 0.2$ the probability that a trial is not affected by artifacts is $(1-p)^{10} \approx 10\%$, thus on average 10 trials are artifact-free in our simulation study. We conjecture that these trials are "close enough" to the average covariance matrix which is used for initialization and therefore our estimator has enough support to converge from the initialized point to a good estimate of the true covariance matrix. Our estimator does not converge to the 90% outlier covariance matrices because (1) they are "too far away" to be generated by the model using the average covariance matrix (initialization matrix), thus are downweighted, and (2) they are too different to be generated by one model. However, if we set $p > 0.2$ then the support of the "good trials" vanishes and the estimate converges to an artifact solution. At the $p = 0.2$ point the support changes from being large enough to being too small. Note that the position of this turning point largely depends on the dimensionality of the signal, the number of trials and other parameters of the data generation process.

In the case of β-GaussCSP the transition from small to large error is much smoother than for β-WishartCSP because the estimation is based on more data (sample level vs. trial level). Thus, even if the probability that a trial is not affected by artifacts is lower than 10%, there is still enough support (number of "good samples") to compute a relatively good estimate of the true covariance matrix. Note also that the artifactual trials contain samples (extreme outliers) which largely dominate the covariance matrix estimation as well as samples which could have been generated by the true model. The extreme outlier samples will mostly be downweighted by β-GaussCSP because the probability density function of a Gaussian decreases towards the tails, thus the samples coming from the artifact-free trials will gain more influence in comparison to the standard covariance estimator. Therefore, β-GaussCSP performs better than all other approaches for $p > 0.2$. Note that the β-GaussCSP estimator works on the level of samples, thus it will never downweight all samples coming from artifactual trials but rather only downweight the outlier samples. Furthermore, it may remove (outlier) samples from trials not affected by an artifact (see discussion below and Figure 28). In contrast, β-WishartCSP works on trial level, thus it (at least for $p < 0.2$) downweights all samples of an artifact trial and does not downweight any sample of an artifact-free trial. We believe that this property of β-WishartCSP is the explanation for its superiority over β-GaussCSP in the range $p = 0.1 - 0.2$. In other words it is better to downweight the outlier trials and not to downweight the artifact-free trials than to downweight parts of both. We are convinced that the trial-level robustness concept introduced in this thesis is a very natural approach for robust parameter estimation when the data has trial structure and may be applied to many real world data sets.

The right panel of Figure 27 depicts the results of the second simulation study (scenario 2). One can see that all methods except MCDE+CSP perform very well until the point where p exceeds 90%. The good performance can be explained by the fact that due to the low variance of the artifactual trials these trials have very limited influence on the estimation of the average class covariance matrices, thus CSP does not extract between-class differences between these artifactual trial but rather the true discriminative activity (which has higher variance). Our beta divergence estimators downweight the artifactual trials because they are far away from the initialized model. The angle error of MCDE+CSP increases if more than 60% of trials are artifacts. The explanation for this strange behaviour of MCDE+CSP is as follows. Since the covariance estimator used in MCDE+CSP prefers covariance matrix estimates with small determinants, it prefers the artifactual trials as they have small variance. Thus, when the proportion of artifacts exceeds 60%, then MCDE+CSP computes spatial filters by using class covariance matrices which are less discriminative than the matrices estimated by the standard covariance estimator. Therefore, the angle error of MCDE+CSP increases. Note that we have selected the best MCDE+CSP parameters a posteriori; for other MCDE+CSP parameters the performance degrades much earlier. Thus, the heuristic used by the minimum covariance determinant estimator fails completely in this simulation study as it prefers the noninformative artifact trials over the artifact-free trials. Since our beta divergence-based estimators do not rely on such a heuristic and are initialized with the average covariance matrices, they perform well in this example.

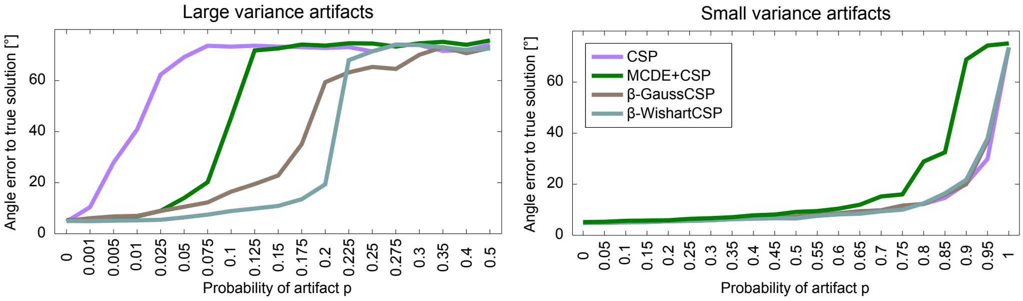

Figure 27.: *Left*: Large variance artifacts are added to a trial and dimension with varying probability. Our β-WishartCSP method clearly outperforms the baselines if $p < 0.2$. *Right*: Small variance trials are added to the data. The MCDE+CSP algorithm has a relatively large angle error as it prefers these small variance trials due to its minimum determinant heuristic.

In the following we show an example where the application of the β-WishartCSP estimator is much more reasonable than the application of β-GaussCSP. The first row Figure 28 displays a signal which consists of four trials. The signal within a trial represents a response to a stimulus. This response is lacking in Trial 3, thus this trial can be seen as an outlier in this example. The second row depicts the weights assigned to the samples of the signal when using the β-GaussCSP algorithm. One can see that the weights assigned to the samples representing the response to the stimulus are much lower than the weights assigned to other samples, i.e., the response to the stimulus is downweighted by the β-GaussCSP estimator. Note that the algorithm does not downweight Trial 3 because it robustifies the estimation on the sample-level and the samples of Trial 3 are not very different from most of the samples in the other trials, therefore they are not treated as outliers. When applying β-WishartCSP to the signal we obtain much more reasonable weights (last row). One can see that Trials 1, 2 and 4 have more or less the same weight whereas the weight assigned to Trial 3 is significantly lower. Thus only our algorithm considers the trial structure and "realizes" that Trial 3 is different from the other trials. The MCDE+CSP method (third row) also does not treat Trial 3 as outlier because it robustifies the parameter estimate on sample-level. Only a robust estimation on trial-level provides the desired solution in this example.

6.4.2 *Performance Results*

Before we compare the performance of β-WishartCSP to the baseline methods on real EEG data, we would like to comment on the ν parameter of β-WishartCSP. This pa-

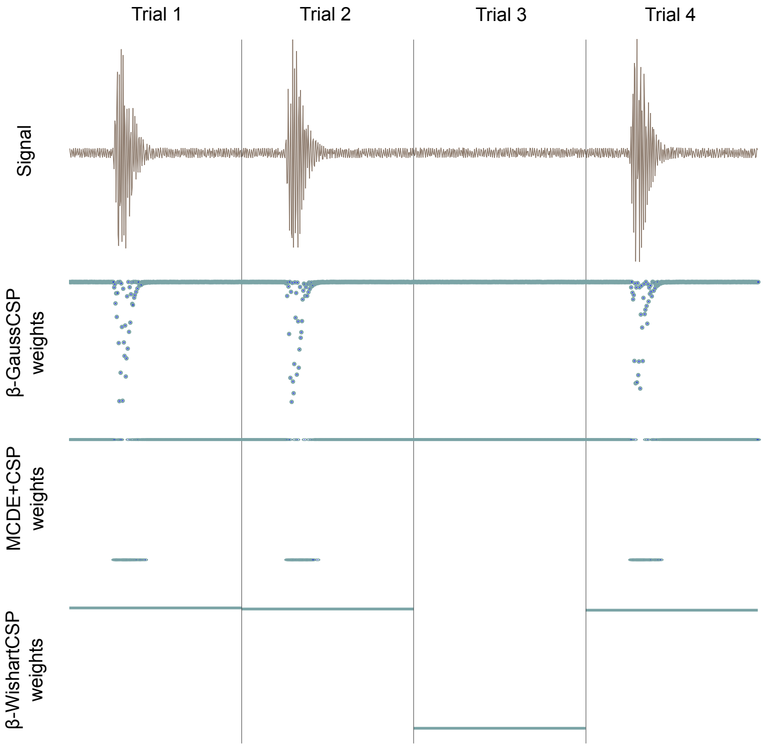

Figure 28.: *First row*: Signal divided into four trials. Trial 3 is an outlier trial because it does not contain the response. *Second row*: Weights of the β-GaussCSP estimator are small for the responses and not for the samples in Trial 3 because robustness is measured on sample-level. *Third row*: Weights of MCDE+CSP are small for the responses and not for the samples in Trial 3 because robustness is measured on sample-level. *Fourth row*: The β-WishartCSP estimator measures robustness on trial-level and assigns a small weight to Trial 3.

rameter of the Wishart distribution captures the uncertainty in estimating the scatter matrices (see Section 6.3), thus should be set to the number of samples within a trial. Note that this rule only holds if the samples within the trial are i.i.d. Since EEG samples are strongly correlated, ν must be set to a much smaller value. Optimally, the parameter should be set to the *effective sample size* (Thiébaux and Zwiers, 1984). In our experiments we do not determine the effective sample size by using advanced methods from statistics but we set ν by hand to be a fraction of the number of samples in a trial. We believe that using more advanced and subject-specific techniques to determine the ν parameter could further improve the performance results. In the following β-WishartCSP denotes the method where ν is set to the number of samples in a trial and βν-WishartCSP represents our algorithm with the reduced ν. We select the β parameter for β-WishartCSP and βν-WishartCSP from $\{0, 2^{-20}, \ldots, 0\}$ whereas in the case of β-GaussCSP we use a smaller range of parameters (as smaller values have less influence), namely $\{0, 2^{-10}, \ldots, 0\}$.

Figure 29 compares our novel beta divergence estimators to the baselines on the Vital BCI data set. Each circle represents the error rate of one subject. The three panels in the first row display the results for β-GaussCSP, the second row depict the results of β-WishartCSP and the last row compares the baselines to βν-WishartCSP. The two panels in the last column compare β-GaussCSP to β-WishartCSP and to βν-WishartCSP. The overall best performance is obtained for βν-WishartCSP, i.e., our novel estimator with the reduced ν parameter. This algorithm provides significantly lower error rates than CSP; the p-value of the one-sided Wilcoxon sign rank test is 0.0045. The performance advantage over MCDE+CSP is almost significant with a p-value of 0.0632. This result is promising as the minimum determinant covariance estimator is a state-of-the-art robust estimator that is very popular in the signal processing community (see e.g. Yong et al., 2008). Our estimator clearly outperforms β-GaussCSP for few subjects, but the overall difference in performance is not significant. This indicates that the advantage of robustifying the estimation on the trial-level rather than on the sample-level is limited in practice. Our estimator does not significantly outperform shrinkCSP, however, to be fair one should note that shrinkCSP solves a different problem than β-WishartCSP, namely the estimation problem in small-sample/high dimensionality settings. Thus, the relative advantage/disadvantage of β-WishartCSP over shrinkCSP is strongly data set dependent. Potentially, one can combine the advantages of analytic shrinkage with the robustness property of our estimator in order to obtain superior results, but we have not tried to do it in this thesis.

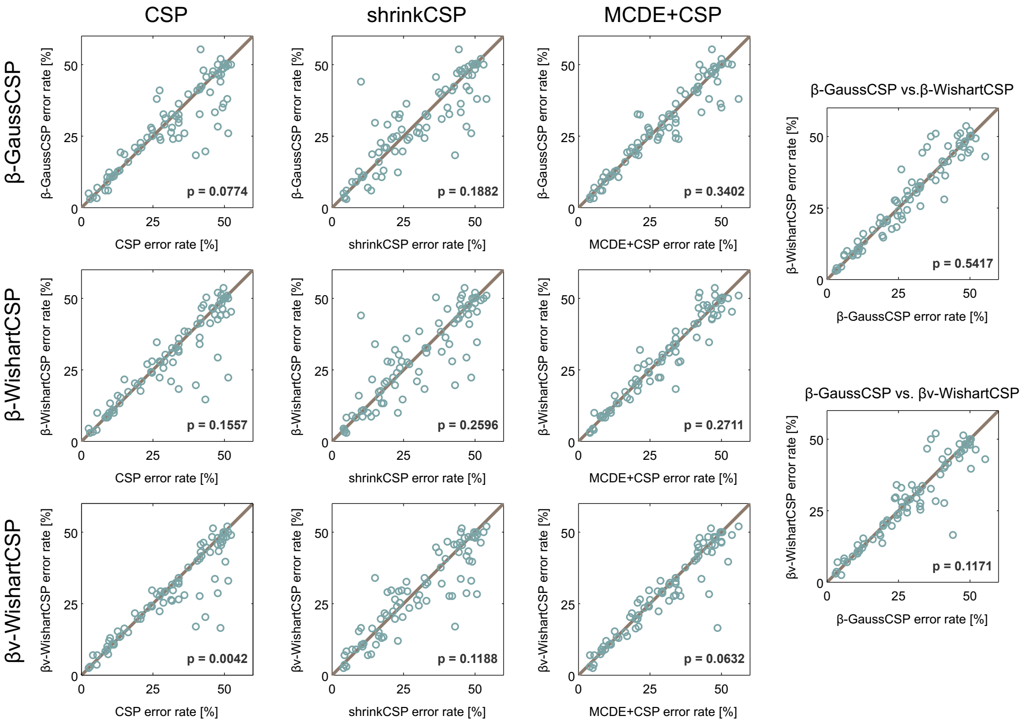

Figure 29.: Scatter plots showing error rates of β-GaussCSP, β-WishartCSP and βν-WishartCSP and the baseline methods. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner.

The performance results on the second data set are displayed in Table 7. Although this data set is less contaminated with artifacts than the Vital BCI data set, the application of our βν-WishartCSP method leads to the highest average classification accuracy. Especially, for subjects A2 and A3 this method significantly outperforms the baseline approaches. Subject A1 on the other hand seems to only benefit from shrinkage, i.e., no robust estimator improves the classification accuracy over 70% for this subject.

6.4.3 *Further Analysis*

In the following we analyse the results of our novel estimators in more detail. The left panel of Figure 30 displays the median error rate of βν-WishartCSP over all 80

Table 7.: Comparison of classification accuracies for β-GaussCSP, β-WishartCSP, βν-WishartCSP and different baselines.

	BCI Competition III					Overall		
Methods	A1	A2	A3	A4	A5	Mean	Median	Std
CSP	66.1	94.6	58.2	87.9	**90.5**	79.5	87.9	16.3
shrinkCSP	**74.1**	94.6	62.8	82.6	**90.5**	80.9	82.6	12.8
MCDE+CSP	68.8	94.6	59.2	**89.7**	88.5	80.2	**88.5**	15.3
β-GaussCSP	66.1	96.4	62.8	89.3	88.9	80.7	88.9	15.2
β-WishartCSP	66.1	96.4	58.2	87.5	**90.5**	79.7	87.5	16.6
βν-WishartCSP	66.1	**98.2**	**66.3**	87.5	**90.5**	**81.7**	87.5	14.7

subjects of the Vital BCI data set for different β parameters. Note that we visualize the error rates for a smaller and finer range of parameter than used in the experiments as the effects of β can be best seen in this range. One can see that the error rate decreases for small β values and it is significantly larger than the error rate of the initial baseline when using $\beta > 0.0039$. This U-shape curve demonstrates that the non-robust estimator (small beta) does not perform well, but also that excessively large values of β may negatively affect performance. The optimal parameter lies between these two extremes. Since the signals of different subjects are usually affected by various amounts of artifacts, the beta value should be optimizes for each subject individually. Note that the experimental results presented in last section may be improved by using a finely tuned range of β parameters (e.g. as in Figure 30).

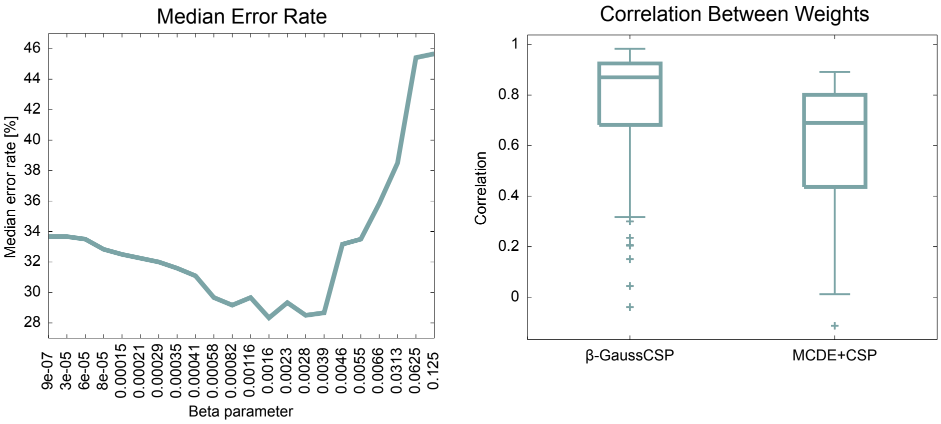

Figure 30.: *Left*: Median error rate of βν-WishartCSP over all 80 subjects of the Vital BCI data set for various β parameters. The error rate curve has an U-shape. *Right*: Correlations between trial weights of βν-WishartCSP and the baselines β-GaussCSP and MCDE+CSP.

In the following we investigate the correlation between the weights computed by βν-WishartCSP and β-GaussCSP and MCDE+CSP. The right panel of Figure 30 displays the distribution of the average (over both classes) correlations between the trial weights. We average the sample weights of β-GaussCSP and MCDE+CSP in order to obtain trial-level weights. The overall correlation over the 80 Vital BCI subjects is quite high for both estimators. This indicates that both estimators find the same outliers as our βν-WishartCSP estimator. The weights of β-GaussCSP are significantly more correlated to βν-WishartCSP as both estimators are based on the same concept. The fact that for some subjects the correlations are relatively low, indicates that there exists principled differences between the estimators. As discussed earlier the estimators robustify the solution on different scales (sample vs. trial). Interestingly, for one subject the correlation between MCDE+CSP and β-WishartCSP is negative with a value of -0.1129. We think that this result is due to random fluctuations (correlations below 0.3 or above -0.3 are usually not seen as being significant) because there is no reason to believe that MCDE+CSP and βν-WishartCSP downweight trials in an opposing manner.

Figure 31 displays the weights computed by βν-WishartCSP and β-GaussCSP for subject 21. The CSP error rate of this subject is 40% whereas the two algorithms provide error rates of 17% and 18.3%, respectively. Note that we display the average sample weights of a trial in the case of β-GaussCSP. One clearly sees that the weights of most of the trials lie in the same range whereas the weights of the outlier trials are almost one order of magnitude smaller. This indicates that some trials are outliers that are downweighted in the estimation process. Although the β-GaussCSP (left panel) and βν-WishartCSP (right panel) weights do not coincide they have the same tendency in the sense that they downweight the same trials. The bottom panel of Figure 31 displays the signal at electrode FFC5 of the trial with smallest weight in βν-WishartCSP (brown circle). The β-GaussCSP method also assigns a very low weight to this trial. This trial contains large artifactual amplitude activity at the beginning. This activity has an amplitude that is almost one order larger than the amplitude of the artifact-free signal, thus this trial is an outlier trial and would have a large impact on the estimated class covariance matrix if it was not downweighted by our method.

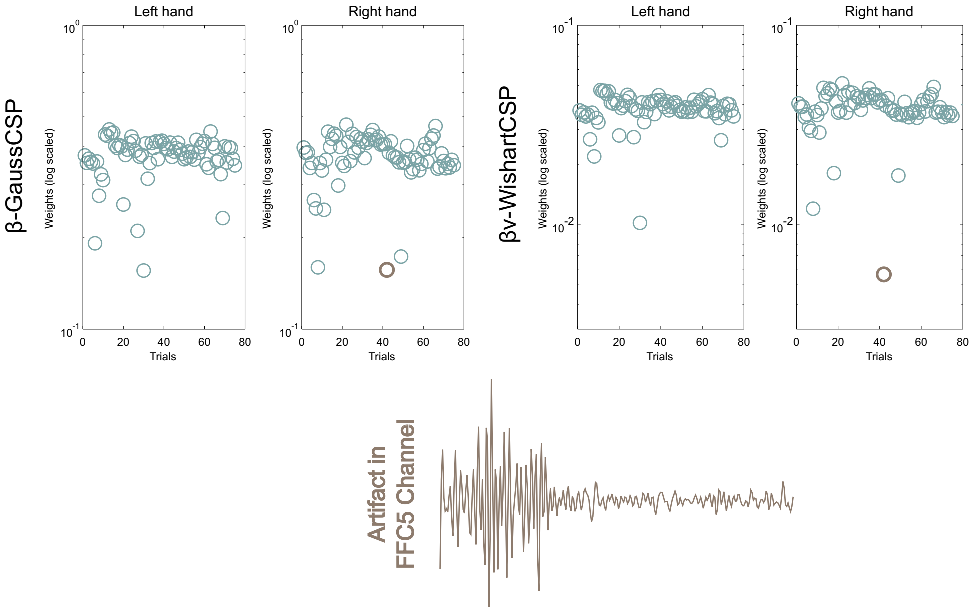

Figure 31.: *Top row*: Trial weights for subject 21 computed by the β-GaussCSP (left) and βν-WishartCSP method (right). *Bottom row*: The signal at electrode FFC5 of the trial with smallest weight (brown circle).

Finally, we would like to visualize the effects of robust estimation in terms of activation patterns. Figure 32 displays the patterns of subject 22. This subject has an CSP error rate of 43.3% and an error rate of 20.3% when applying our βν-WishartCSP algorithm. We visualize the results for a parameter with an error rate of 15%. One clearly sees that the fifth and sixth patterns of βν-WishartCSP (lower row) represent motor imagery related activity, thus they have a very high AUC value. In the case of CSP (top row) all patterns are affected by artifacts and do not capture the true BCI activity.

This is the reason for the high CSP error rate and the relatively good performance of βν-WishartCSP.

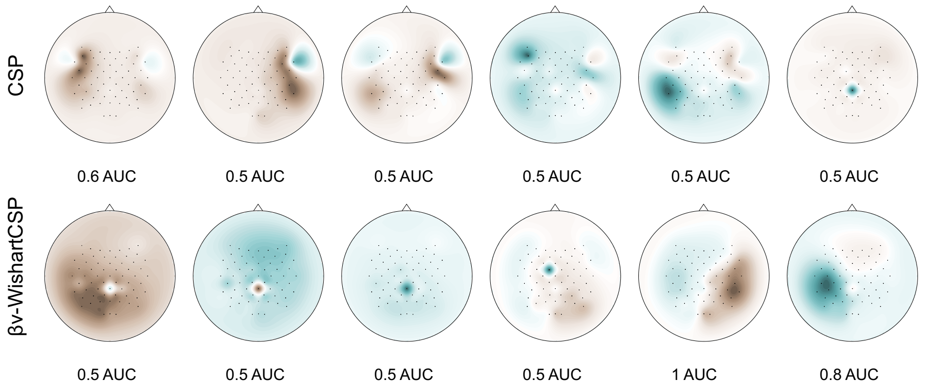

Figure 32.: *Top row*: Activation patterns of CSP. *Bottom row*: Activation patterns of βν-WishartCSP.

Lessons learned in this chapter

- Trial-level robustification is more effective than robust estimation on sample-level if whole trials are affected by artifacts.
- Estimators maximizing Ψ-likelihood neither rely on heuristics nor assume a maximum fraction of outliers in the data.
- The ν parameter of the β-WishartCSP method must be set to the effective dimensionality when data is correlated.
- Few artifactual trials suffice to significantly affect results.
- Initialization of β-WishartCSP is not a critical factor; using the sample covariance matrix works well in practice.

Future Work

- Application to other data sets, e.g., response to a stimulus.
- Computation of effective dimensionality.
- Criterion for parameter selection; Bayesian framework.
- Application to exploratory analysis, e.g., detection of “outlier subjects”.

7 DIVERGENCE-BASED CSP FRAMEWORK

Many machine learning algorithms can be cast into the framework of information geometry and formulated as divergence optimization problems. Prominent examples are algorithms such as Independent Component Analysis (Hyvärinen, 1999a), Non-negative Matrix Factorization (NMF) (Févotte and Idier, 2011) and Stationary Subspace Analysis (Kawanabe et al., 2011). Once an algorithm has been formulated in terms of a particular divergence, a whole class of novel algorithms can be obtained by using the "divergence trick"[1], i.e. by keeping the mathematical formulation of the problem but using other divergences with different properties. A divergence formulation has the advantage that it embeds the algorithm in a profoundly understood mathematical framework and provides an information geometric interpretation for it. This chapter introduces a divergence formulation of Common Spatial Patterns and proposes several novel CSP variants by utilizing the robustness property of beta divergence and by adding a divergence-based regularization term to the objective function. In contrast to the approaches presented in earlier chapters this divergence-based penalty term captures nonstationarity in a principled manner.

7.1 CSP AS DIVERGENCE MAXIMIZATION PROBLEM

As mentioned earlier in this thesis (and also shown in (Lemm et al., 2011)) the CSP projection matrix $\mathbf{W} = \tilde{\mathbf{R}}\mathbf{P}$ can be decomposed into a whitening $\mathbf{P}$ and an orthogonal projection part $\tilde{\mathbf{R}}$. The whitening transformation ensures that the extracted sources are uncorrelated (this maximizes the informativity of the features) whereas the orthogonal projection optimizes the CSP criterion. In this section we propose a novel *divergence-based Common Spatial Patterns* (divCSP) method which computes spatial filters by optimizing a divergence criterion. The following theorem relates CSP to divergence maximization.

Theorem 10. *Let* $\mathbf{W} \in \mathbb{R}^{D\times d}$ *be the top* d *(sorted by* α_i*, see Section 3.1) spatial filters computed by CSP and let* $\mathbf{V}^\top = \tilde{\mathbf{R}}\mathbf{P} \in \mathbb{R}^{d\times D}$ *be a matrix that can be decomposed into a whitening projection* $\mathbf{P} \in \mathbb{R}^{D\times D}$ *with* $(\mathbf{P}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\mathbf{P}^\top = \mathbf{I})$ *and an orthogonal projection* $\tilde{\mathbf{R}} = \mathbf{I}_d\mathbf{R} \in \mathbb{R}^{d\times D}$ *with* $\mathbf{I}_d \in \mathbb{R}^{d\times D}$ *being the identity matrix truncated to the first* d *rows and* $\mathbf{R}^\top\mathbf{R} = \mathbf{I} \in \mathbb{R}^{D\times D}$*. Then*

$$\mathrm{span}(\mathbf{W}) \;=\; \mathrm{span}(\mathbf{V}^*) \tag{53}$$

$$\text{with } \mathbf{V}^* \;=\; \underset{\mathbf{V}}{\operatorname{argmax}}\, \tilde{D}_{kl}\left(\mathcal{N}(0, \mathbf{V}^\top\mathbf{\Sigma}_1\mathbf{V}) \;||\; \mathcal{N}(0, \mathbf{V}^\top\mathbf{\Sigma}_2\mathbf{V})\right). \tag{54}$$

Proof. See appendix B.3 □

This theorem states that the CSP filters $\mathbf{W}$ project the data to a subspace with maximum discrepancy, measured by symmetric Kullback-Leibler divergence, between the d-dimensional Gaussian distributions $\mathcal{N}\left(0, \mathbf{W}^\top\mathbf{\Sigma}_1\mathbf{W}\right)$ and $\mathcal{N}\left(0, \mathbf{W}^\top\mathbf{\Sigma}_2\mathbf{W}\right)$. Thus, instead of computing spatial filters with CSP we obtain an equivalent solution when maximizing Eq. (54). Note that there exists a technical difference between this divergence maximization problem and the divergence minimization described previously. In the last chapter we have minimized the divergence between a fixed distribution,

[1] Note the analogy to the kernel trick for Support Vector Machines (Müller et al., 2001)

namely the empirical distribution, and a model distribution. In other words we optimize the divergence with respect to only one argument. In the case of divergence-based CSP we aim to find the spatial filters projecting the data to a discriminative subspace, i.e., we optimize over both divergence arguments as both arguments are dependent on the projection. In the following we present two approaches for finding the projections which maximize the divergence between class distributions. Note that our optimization algorithm is based on the Lie algebra (Plumbley, 2005) optimization techniques used by the Stationary Subspace Analysis (SSA) method (von Bünau et al., 2009; Király et al., 2012; von Bünau, 2012).

7.2 OPTIMIZATION ALGORITHMS

We present two algorithms maximizing the divCSP objective function. All divCSP variants proposed in this chapter use these two algorithms. Note that since the optimization is based on gradient descent, thus gives only local optima, we may need to restart the algorithms multiple times or initialize them with parameters which are close to the global solution. Our goal is to maximize the divergence term in Eq. (54), i.e., to find a subspace that maximizes the symmetric Kullback-Leibler divergence between two Gaussian distributions. Note that in subsequent sections we will optimize more generic objective functions consisting of sums of (beta or KL) divergences; these objective functions can also be maximized by the following algorithms.

Subspace Method

Let us first describe the *subspace* approach (see Algorithm 4). This method aims to extract the whole subspace in one run, i.e., to directly optimize for the projection matrix $\mathbf{V}$. Its input parameters are the average class covariance matrices $\{\mathbf{\Sigma}_c\}$ and the dimensionality of the subspace d, i.e., the number of spatial filters to return. The first step of the method consists of the computation of a whitening matrix $\mathbf{P} \in \mathbb{R}^{D \times D}$ which projects the data onto the unit sphere, i.e., $\mathbf{P}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\mathbf{P}^\top = \mathbf{I}$. This whitening transformation is applied to the class covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ followed by a (random) rotation with $\mathbf{R}_0 \in \mathbb{R}^{D \times D}$. Note that the rotation matrix satisfies $\mathbf{R}_0^\top \mathbf{R}_0 = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. The optimization process then consists of finding a rotation matrix $\mathbf{R} \in \mathbb{R}^{D \times D}$ which maximizes the symmetric KL divergence in the first d sources. The following theorem derives the objective function and the gradient when using Kullback-Leibler divergence.

Theorem 11. *The objective function in Eq. (54) can be represented in explicit form as*

$$\mathcal{L}_{kl}(\mathbf{R}) = \tilde{D}_{kl}\left(\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_d^\top \parallel \mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_d^\top\right) \tag{55}$$

$$= \frac{1}{2}\mathrm{tr}\left((\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_d^\top)^{-1}(\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_d^\top) + (\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_2\mathbf{R}^\top\mathbf{I}_d^\top)^{-1}(\mathbf{I}_d\mathbf{R}\tilde{\mathbf{\Sigma}}_1\mathbf{R}^\top\mathbf{I}_d^\top)\right) - d,$$

with gradient

$$\begin{aligned}\nabla_{\mathbf{R}}\mathcal{L}_{kl}(\mathbf{R}) = \mathbf{I}_d^\top\Big((\bar{\mathbf{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_2 - (\bar{\mathbf{\Sigma}}_1)^{-1}\bar{\mathbf{\Sigma}}_2(\bar{\mathbf{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_1 \\ + (\bar{\mathbf{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_1 - (\bar{\mathbf{\Sigma}}_2)^{-1}\bar{\mathbf{\Sigma}}_1(\bar{\mathbf{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_2\Big)\mathbf{R}\end{aligned} \tag{56}$$

Proof. See appendix B.1 □

Note that $\tilde{\mathbf{\Sigma}}_1$ and $\tilde{\mathbf{\Sigma}}_2$ denote the whitened covariance matrices, $\bar{\mathbf{\Sigma}}_1$ and $\bar{\mathbf{\Sigma}}_2$ are the projected covariance matrices and $\mathbf{I}_d \in \mathbb{R}^{d \times D}$ is the identity matrix truncated to the first d rows. Although $\mathbf{R}$ is a $D \times D$ rotation matrix, we only evaluate the first d rows, i.e., we only evaluate the divergence in the d-dimensional subspace.

The goal of the optimization algorithm is to maximize $\mathcal{L}_{kl}(\mathbf{R})$ under the orthogonality constraint $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$. This can be achieved by using Lie Group Methods (Plumbley, 2005). The optimization is performed by gradient descent on the manifold of orthogonal matrices. More precisely, we start with an orthogonal matrix $\mathbf{R}_0$ and find an orthogonal update $\mathbf{U}$ in the kth step such that $\mathbf{R}_{k+1} = \mathbf{U}\mathbf{R}_k$. This ensures that we stay on the manifold of orthogonal matrices at each step. Note that the update matrix may be written as a matrix exponential of a skew-symmetric matrix $\mathbf{M} = -\mathbf{M}^\top$. We find a search direction $\mathbf{H} = -\mathbf{H}^\top$ in the set of skew symmetric matrices by computing the gradient of the loss function w.r.t. $\mathbf{M}$ at $\mathbf{M} = 0$ and determine the optimal step size t along this gradient by line search (see von Bünau, 2012). Finally, we represent the update matrix as $\mathbf{U} = e^{t\mathbf{H}}$. In other words we search over the Lie group SO(n) of orthogonal matrices by computing the gradient in the corresponding Lie algebra $\mathfrak{so}(n)$. The gradient in $\mathfrak{so}(n)$ can be calculated as (see Plumbley, 2005)

$$\nabla \mathcal{L} \;=\; (\nabla_{\mathbf{R}} \mathcal{L})\,\mathbf{R}^\top \;-\; \mathbf{R}\,(\nabla_{\mathbf{R}} \mathcal{L})^\top .$$

The objective function in Eq. (55) is invariant to rotations within the d-dimensional subspace, i.e., for $|\mathbf{G}| \neq 0$ the following equality holds

$$\tilde{D}_{kl}\left(\mathbf{V}^\top \boldsymbol{\Sigma}_1 \mathbf{V} \;||\; \mathbf{V}^\top \boldsymbol{\Sigma}_2 \mathbf{V}\right) \;=\; \tilde{D}_{kl}\left(\mathbf{G}^\top \mathbf{V}^\top \boldsymbol{\Sigma}_1 \mathbf{V}\mathbf{G} \;||\; \mathbf{G}^\top \mathbf{V}^\top \boldsymbol{\Sigma}_2 \mathbf{V}\mathbf{G}\right). \quad (57)$$

Therefore, we rotate the projection matrix $\mathbf{V}$ in the last step of the algorithm, so that it maximally separate the classes along the projection directions (as is the case with CSP). The spatial filters can be rearranged so that they capture the class differences with decreasing strength (α_i sorting).

Algorithm 4 Subspace divCSP

```
function SUB-DIVCSP($\{\boldsymbol{\Sigma}_c\}$, d)
    Compute the whitening matrix $\mathbf{P} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-\frac{1}{2}}$
    Initialize $\mathbf{R}_0$ with a (random) rotation matrix
    Whiten and rotate $\boldsymbol{\Sigma}_c = (\mathbf{R}_0\mathbf{P})\boldsymbol{\Sigma}_c(\mathbf{R}_0\mathbf{P})^\top$ with $c \in \{1, 2\}$
    repeat
        Compute the gradient matrix and determine the optimal step size
        Update the rotation matrix $\mathbf{R}_{k+1} = \mathbf{U}\mathbf{R}_k$
        Apply the rotation to the data $\boldsymbol{\Sigma}_c = \mathbf{U}\boldsymbol{\Sigma}_c\mathbf{U}^\top$
    until convergence
    Let $\mathbf{V}^\top = \mathbf{I}_d\mathbf{R}_{k+1}\mathbf{P}$
    Compute the eigenvectors $\mathbf{G} \in \mathbb{R}^{d \times d}$ of $\mathbf{V}^\top\boldsymbol{\Sigma}_1\mathbf{V}$
    Let $\mathbf{V}^* = \mathbf{V}\mathbf{G}$ and rearrange the filters if necessary ($\alpha_i$ sorting)
    return $\mathbf{V}^*$
end function
```

Deflation Method

The *deflation* algorithm does not extract the whole subspace at once, but performs the optimization in a sequential manner (see also deflation FastICA (Hyvärinen, 1999b)). More precisely, we reduce the dimensionality of the data space by one in each step. This provides a sorting of the spatial filters which is analogous to CSP, i.e., the first solution is the most discriminative filter and so on. The steps of the method are described in Algorithm 5. In the first steps of the algorithm we apply the whitening transformation $\mathbf{P}$ to the class covariance matrices and initialize a matrix $\mathbf{B}$ that represents the basis of the subspace in which the spatial filters are computed. Then we repeat the following procedure d times. We calculate the spatial filter by applying the subspace divCSP algorithm described in Algorithm 4 with parameter $d = 1$. Note that we skip the whitening step as it has already been performed. After obtaining the spatial filter $\mathbf{v}$ we compute its corresponding orthogonal complement $\mathbf{V}^\perp$ and project the class covariance matrices to this subspace. This step ensures that the spatial filters computed in subsequent steps will be orthogonal to the current ones. Since the ith spatial filter $\mathbf{v}$ has been computed in the subspace with basis $\mathbf{B}$ its representation in the original coordinate system is $\mathbf{v}_i = \mathbf{B}\mathbf{v}$. In the last step of the loop we update the basis matrix $\mathbf{B}$. The final solution consists of the spatial filters $\mathbf{v}_i$ with $i = 1 \dots d$ and is already sorted by the ability to capture the class differences (α_i sorting, , see Section 3.3).

The deflation algorithm optimizes the following objective function in ith step

$$\tilde{D}_{kl}\left(\mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{v}_i \parallel \mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{v}_i\right) = \frac{\mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{v}_i}{\mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{v}_i} + \frac{\mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{v}_i}{\mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{v}_i} \tag{58}$$

$$\text{s.t.} \quad \mathbf{v}_i^\top \mathbf{v}_j = 0 \quad \forall j \in 1 \dots i-1, \tag{59}$$

where as before we denote the whitened matrices with a tilde. Note that this objective function can be written as $f(z) = z + \frac{1}{z}$ with $z = \frac{\mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_1 \mathbf{v}_i}{\mathbf{v}_i^\top \tilde{\boldsymbol{\Sigma}}_2 \mathbf{v}_i}$ and one can easily prove that this function is maximized at the border. Thus, it is maximized either for the largest z or for the smallest one (largest $\frac{1}{z}$). This solution corresponds to the ith CSP spatial filter (sorted by α_i). Thus, both methods, subspace and deflation, provide the same spatial filters, namely the CSP ones, when optimizing the objective function in Eq. (55). For the extensions of divCSP (described in next sections) the solutions of the subspace and deflation method will not coincide. We will discuss the differences of both optimization schemes in the experimental section using simulations.

Algorithm 5 Deflation divCSP

```
function DEF-DIVCSP($\{\boldsymbol{\Sigma}_c\}$, d)
    Compute the whitening matrix $\mathbf{P} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-\frac{1}{2}}$
    Apply the whitening transformation $\tilde{\boldsymbol{\Sigma}}_c = \mathbf{P}\boldsymbol{\Sigma}_c\mathbf{P}^\top$ with $c \in \{1, 2\}$
    Initialize basis $\mathbf{B} = \mathbf{I} \in \mathbb{R}^{D \times D}$
    for $i = 1 \dots d$ do
        Compute $\mathbf{v} \in \mathbb{R}^{(D-i+1) \times 1}$ by sub-divCSP($\tilde{\boldsymbol{\Sigma}}_1, \tilde{\boldsymbol{\Sigma}}_2, 1$) (no whitening)
        Compute $\mathbf{V}^\perp \in \mathbb{R}^{(D-i+1) \times (D-i)}$ the orthogonal complement of $\mathbf{v}$
        Project $\tilde{\boldsymbol{\Sigma}}_1$ and $\tilde{\boldsymbol{\Sigma}}_2$ to $(D-i)$-dimensional subspace by $\mathbf{V}^\perp$
        Backproject spatial filter $\mathbf{v}$ to original space by $\mathbf{v}_i = \mathbf{B}\mathbf{v} \in \mathbb{R}^{D \times 1}$
        Update basis $\mathbf{B} = \mathbf{B}\mathbf{V}^\perp \in \mathbb{R}^{D \times (D-i)}$
    end for
    Let $\mathbf{V}^* = \mathbf{P}[\mathbf{v}_1 \dots \mathbf{v}_d]$
    return $\mathbf{V}^*$
end function
```

7.3 ROBUSTNESS THROUGH BETA DIVERGENCE

In this section we introduce a divCSP method based on symmetric beta divergence which downweights the influence of outlier trials. Thus, our idea is to keep the mathematical formulation of the CSP problem (see Eq. (54)) but use beta divergence as it is known to be robust to outliers. However, the direct application of beta divergence to the divCSP objective function has no robustness effect because the average class covariance matrices may be affected by trial artifacts (if not estimated with the method presented in last chapter). The downweighting of beta divergence does not have any effect when applied to only one divergence term.

In order to deal with artifacts on a trial-by-trial basis we need to reformulate the above objective function. Instead of maximizing the divergence between the average class distributions we propose to optimize the sum of trial-wise divergences

$$\mathcal{L}_{\text{sumkl}}(\mathbf{V}) \;=\; \sum_{i=1}^{n} \tilde{D}_{kl}\left(\mathbf{V}^\top \boldsymbol{\Sigma}_1^i \mathbf{V} \;\|\; \mathbf{V}^\top \boldsymbol{\Sigma}_2^i \mathbf{V}\right), \tag{60}$$

where $\boldsymbol{\Sigma}_1^i$ and $\boldsymbol{\Sigma}_2^i$ denote the covariance matrices estimated from the ith trial of class 1 and class 2, respectively, and n is the number of trials per class (which is assumed to be the same for both classes). Note that the reformulated problem is not equivalent to CSP; in Eq. (54) averaging is performed w.r.t. the covariance matrices (divergence is computed on average covariance matrices), whereas in Eq. (60) averaging is performed w.r.t. the divergences. We denote the former approach by kl-*divCSP* and the latter one by *sum*kl-*divCSP*. The following theorem relates both approaches in the asymptotic case.

Theorem 12. *Suppose that the number of discriminative sources is one; then let* c *be such that* $D/N \to c$ *as* $D, N \to \infty$ *(*D *dimensions,* N *data points per trial). Then, if there exists* $\gamma(c)$ *with* $n/D \to \gamma(c)$ *for* $n \to \infty$ *(*n *the number of trials) then the empirical maximizer of* $\mathcal{L}_{\text{sumkl}}(\mathbf{v})$ *(and of course also of* $\mathcal{L}_{kl}(\mathbf{v})$*) converges almost surely to the true solution.*

Sketched proof. Since there is only one discriminative direction we may perform analysis in a basis whereby the covariances of both classes have the form $\text{diag}(a, 1, \ldots, 1)$ and $\text{diag}(b, 1, \ldots, 1)$. If we show in this basis that consistency holds then it is a simple matter to prove consistency in the original basis. We want to show that as the number of trials n increases the filter provided by sumkl-divCSP converges to the true solution. If the support of the density of the eigenvalues includes a region around 0, then there is no hope of showing that the matrix inversion is stable. However, it has been shown in the random matrix theory literature (Baik and Silverstein, 2006) that if D and N tend to ∞ in a ratio $c = \frac{D}{N}$ then all of the eigenvalues apart from the largest lie between $(1-\sqrt{c})^2$ and $(1+\sqrt{c})^2$ whereas the largest sample eigenvalue (α denotes the true non-unit eigenvalue) converges almost surely to $\alpha + c\frac{\alpha}{\alpha-1}$ provided $\alpha > 1 + \sqrt{c}$, independently of the distribution of the data; a similar result applies if one true eigenvalue is smaller than the rest. This implies that for sufficient discriminability in the true distribution and sufficiently many data points per trial, each filter maximizing each term in the sum has non-zero dot-product with the true maximizing filter. But since the trials are independent, this implies that in the limit of n trials the maximizing filter corresponds to the true filter. □

Theorem 2 states that both divergence-based CSP variants, kl-divCSP and sumkl-divCSP, almost surely converge to the same (true) solution in the asymptotic case. Note that both of the above approaches, kl-divCSP and sumkl-divCSP, are not robust w.r.t. artifacts as they both perform simple (non-robust) averaging of the covariance matrices and of the divergence terms, respectively. In the following we show that by using symmetric beta divergence we robustify the averaging of the divergence terms in sumkl-divCSP in the same manner as done in last chapter (i.e. beta divergence will downweight the influence of outlier divergence terms). This downweighting has the effect that our *beta divergence CSP* (β-divCSP) algorithm will not try to extract features that have a high average discriminativity, but features that are discriminative across trials (stable solution). We propose to use symmetric beta divergence for the objective function in Eq. (60). The relations between CSP, kl-divCSP, sumkl-divCSP and β-divCSP are depicted in Figure 33.

The objective function of our β-divCSP approach can be written explicitly and both algorithms, subspace and deflation method, can be used for maximizing it.

Theorem 13. *The objective function of β-divCSP can be represented in explicit form as*

$$
\begin{aligned}
\mathcal{L}_{\beta}(\mathbf{V}) &= \sum_{i} \tilde{D}_{\beta}\left(\mathbf{V}^{\top}\boldsymbol{\Sigma}_1^i\mathbf{V} \;\|\; \mathbf{V}^{\top}\boldsymbol{\Sigma}_2^i\mathbf{V}\right) \qquad (61)\\
&= \gamma \sum_{i} \Big(|\bar{\boldsymbol{\Sigma}}_1^i|^{-\frac{\beta}{2}} + |\bar{\boldsymbol{\Sigma}}_2^i|^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}}\Big(|\bar{\boldsymbol{\Sigma}}_2^i|^{\frac{1-\beta}{2}}|\beta\bar{\boldsymbol{\Sigma}}_1^i + \bar{\boldsymbol{\Sigma}}_2^i|^{-\frac{1}{2}}\\
&\quad + |\bar{\boldsymbol{\Sigma}}_1^i|^{\frac{1-\beta}{2}}|\beta\bar{\boldsymbol{\Sigma}}_2^i + \bar{\boldsymbol{\Sigma}}_1^i|^{-\frac{1}{2}}\Big)\Big),
\end{aligned}
$$

with $\gamma = \frac{1}{\beta}\sqrt{\frac{1}{(2\pi)^{\beta d}(\beta+1)^{d}}}$ *and* $\bar{\boldsymbol{\Sigma}}_c^i = \mathbf{V}^{\top}\boldsymbol{\Sigma}_c^i\mathbf{V}$. *The gradient is*

$$
\begin{aligned}
\nabla_{\mathbf{R}}\mathcal{L}_{kl}(\mathbf{R}) &= \gamma\mathbf{I}_d^{\top}\Big(\beta|\bar{\boldsymbol{\Sigma}}_1|^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 + \beta|\bar{\boldsymbol{\Sigma}}_2|^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2 \qquad (62)\\
&- (\beta+1)^{\frac{d}{2}}|\bar{\boldsymbol{\Sigma}}_2|^{\frac{1-\beta}{2}}\cdot|\beta\bar{\boldsymbol{\Sigma}}_1 + \bar{\boldsymbol{\Sigma}}_2|^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2\\
&+ (\beta\bar{\boldsymbol{\Sigma}}_1 + \bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_1 + \tilde{\boldsymbol{\Sigma}}_2)\Big] - (\beta+1)^{\frac{d}{2}}|\bar{\boldsymbol{\Sigma}}_1|^{\frac{1-\beta}{2}}\cdot|\beta\bar{\boldsymbol{\Sigma}}_2 + \bar{\boldsymbol{\Sigma}}_1|^{-\frac{1}{2}}\\
&\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 + (\beta\bar{\boldsymbol{\Sigma}}_2 + \bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_2 + \tilde{\boldsymbol{\Sigma}}_1)\Big]\Big)\mathbf{R}
\end{aligned}
$$

Proof. See appendix B.2 □

In the following we show that the robustness property of β-divCSP can be directly understood from inspection of its objective function. Assume $\bar{\boldsymbol{\Sigma}}_1^i$ and $\bar{\boldsymbol{\Sigma}}_2^i$ are full rank $d \times d$ covariance matrices. We investigate the behaviour of the objective functions of β-divCSP and sumkl-divCSP when $\bar{\boldsymbol{\Sigma}}_1^i$ is constant and $\bar{\boldsymbol{\Sigma}}_2^i$ (i.e. one of its eigenvalues) becomes very large, e.g., because the trial is affected by artifacts. It is not hard to see that for $\beta > 0$ the objective function $\mathcal{L}_{\beta}$ does not go to infinity but is constant as $\bar{\boldsymbol{\Sigma}}_2^i$ becomes arbitrarily large. The first term of the objective function $|\bar{\boldsymbol{\Sigma}}_1^i|^{-\frac{\beta}{2}}$ is constant with respect to changes of $\bar{\boldsymbol{\Sigma}}_2^i$ and all the other terms go to zero as $\bar{\boldsymbol{\Sigma}}_2^i$ increases. Thus, the influence function of the β-divCSP estimator is bounded w.r.t. changes in $\bar{\boldsymbol{\Sigma}}_2^i$ (the same argument holds for changes of $\bar{\boldsymbol{\Sigma}}_1^i$). Note that this robustness property increases with β and vanishes when applying Kullback-Leibler divergences defined in Eq. (55) as the trace term $\operatorname{tr}\left((\bar{\boldsymbol{\Sigma}}_1^i)^{-1}\bar{\boldsymbol{\Sigma}}_2^i\right)$ is not bounded when $\bar{\boldsymbol{\Sigma}}_2^i$ becomes arbitrarily large, thus this artifactual term will dominate the solution.

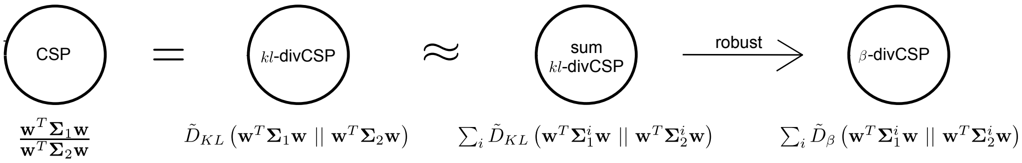
CSP
=
kl-divCSP
≈
sum
kl-divCSP
robust
β-divCSP
$\frac{\mathbf{w}^T\boldsymbol{\Sigma}_1\mathbf{w}}{\mathbf{w}^T\boldsymbol{\Sigma}_2\mathbf{w}}$
$\tilde{D}_{KL}\left(\mathbf{w}^T\boldsymbol{\Sigma}_1\mathbf{w} \,\|\, \mathbf{w}^T\boldsymbol{\Sigma}_2\mathbf{w}\right)$
$\sum_i \tilde{D}_{KL}\left(\mathbf{w}^T\boldsymbol{\Sigma}_1^i\mathbf{w} \,\|\, \mathbf{w}^T\boldsymbol{\Sigma}_2^i\mathbf{w}\right)$
$\sum_i \tilde{D}_{\beta}\left(\mathbf{w}^T\boldsymbol{\Sigma}_1^i\mathbf{w} \,\|\, \mathbf{w}^T\boldsymbol{\Sigma}_2^i\mathbf{w}\right)$

Figure 33.: The relations between CSP, kl-divCSP, sumkl-divCSP and β-divCSP.

7.4 UNIFIED VIEW ON REGULARIZATION

We have proven that CSP can be formulated in a divergence maximization framework and have derived a robust version of the algorithm based on beta divergence. However, maximizing the band power ratios may not be the only objective for feature extraction. For instance, imposing stationarity on the extracted features is also of high interest in Brain-Computer Interfacing (see e.g. sCSP and ssCSP algorithms presented in this thesis). A natural way of regularizing the extracted spatial filters towards stationarity is to combine the objective function of divCSP with a divergence term which accounts for the stationarity of the features. In this section we present four such regularization terms which tackle various nonstationarity problems. Since the optimization process is not affected by changing the way how stationarity is measured (as long as it is a divergence), our framework integrates several stationary CSP variants in a principled manner, moreover, it allows one to utilize information from additional subjects.
The objective function of the proposed regularized divCSP method can be written as

$$\mathcal{L}(\mathbf{V}) = \underbrace{(1-\lambda)\tilde{D}_{kl}\left(\mathbf{V}^\top\mathbf{\Sigma}_1\mathbf{V} \parallel \mathbf{V}^\top\mathbf{\Sigma}_2\mathbf{V}\right)}_{\text{CSP Term}} - \underbrace{\lambda\mathbf{\Delta}}_{\text{Regularization Term}} \qquad (63)$$

where $\mathbf{\Delta}$ is the regularization term that can be arbitrarily defined, depending on the type of invariance we want to achieve, and λ is a regularization parameter trading-off the influence of the CSP objective function and the regularization term. Note that the objective functions of all algorithms presented in this section can be written as weighted sum of divergences and the goal is to find a projection to a d-dimensional subspace which maximizes this sum (by using Algorithm 4 or 5). In the following we discuss four different regularization terms.

Within-Session Stationarity (divCSP-WS): In order to reduce the influence of artifacts or shifts that are present in the training session we divide the data into a set of smaller epochs. The epochs consist of concatenated recordings of one or several subsequent trials of the same class. The nonstationarity of the extracted features is measured as average divergence between the data distribution of the epochs and the whole data distribution for each class separately (analogous to sCSP). More precisely, we compute

$$\mathbf{\Delta} = \frac{1}{2n}\sum_{c=1}^{2}\sum_{i=1}^{n} D_{KL}\left(\mathbf{V}^\top\mathbf{\Sigma}_c^i\mathbf{V} \parallel \mathbf{V}^\top\mathbf{\Sigma}_c\mathbf{V}\right), \qquad (64)$$

where n denotes the number of trials and $\mathbf{\Sigma}_c^i$ stands for the estimated covariance matrix of class c and epoch i. Note that we use the Kullback-Leibler divergence (and not its symmetric version) for capturing the changes; the reasons for that will be explained in the Section 7.5.5. Subtracting the regularization term $\mathbf{\Delta}$ from the divCSP objective function (as in Eq. (63)) reduces the within-class variability of the extracted training features.

Between-Session Stationarity (divCSP-BS): The purpose of the following regularization term is to reduce the shift between the data distribution in calibration and test phase. Since we may assume that test data is not available at the time of computing the spatial filters we utilize information from additional subjects to estimate these changes. Note that this approach implicitly assumes that the between-session nonstationarities are similar among different users, e.g., because they are induced by a change in experimental paradigm (e.g., no feedback vs. visual feedback), thus this regularization term is based on the idea of transferring nonstationary information between subjects introduced in this thesis. For our experiments we consider the following regularization term

$$\Delta \;=\; \frac{1}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K} \tilde{D}_{KL}\left(\mathbf{V}_n^\top \boldsymbol{\Sigma}_{tr,c}^{k}\mathbf{V} \;||\; \mathbf{V}^\top \boldsymbol{\Sigma}_{te,c}^{k}\mathbf{V}\right), \tag{65}$$

where K stands for the number of additional subjects and $\boldsymbol{\Sigma}_{tr,c}^{k}$ and $\boldsymbol{\Sigma}_{te,c}^{k}$ denotes subject's k class covariance matrix estimated on training and test data, respectively. Note that in contrast to ssCSP this regularization term measures the shift for each class separately.

Across-Subject Stationarity (divCSP-AS): If the goal is to reduce differences between subjects, e.g., because one assumes that the underlying processes governing motor imagery are very similar between users, then one may use the class-unrelated changes between the average data of the subject of interest ℓ and the data of additional subjects k as regularization term

$$\Delta \;=\; \frac{1}{K}\sum_{k=1}^{K} \tilde{D}_{KL}\left(\mathbf{V}^\top \boldsymbol{\Sigma}_{tr}^{\ell}\mathbf{V} \;||\; \mathbf{V}^\top \boldsymbol{\Sigma}_{tr}^{k}\mathbf{V}\right). \tag{66}$$

Multi-Subject (divCSP-MS): Rather than combining the discriminativity term with a regularization term which captures nonstationarity, we may also combine it with the divCSP objective functions of other subjects. This allows us to extract a more subject-independent feature space. In this case we need to invert the sign of $\boldsymbol{\Delta}$ as we aim to maximize this regularization term

$$\Delta \;=\; -\frac{1}{K}\sum_{k=1}^{K} \tilde{D}_{KL}\left(\mathbf{V}^\top \boldsymbol{\Sigma}_{1}^{k}\mathbf{V} \;||\; \mathbf{V}^\top \boldsymbol{\Sigma}_{2}^{k}\mathbf{V}\right). \tag{67}$$

Many other forms of regularization, e.g., considering multiple classes or containing priori information, can be easily integrated into our framework. Note that β-divCSP can also be formulated in this framework with a regularization target

$$\Delta \;=\; -\sum_{i=1}^{n} \tilde{D}_{kl}\left(\mathbf{V}^\top \boldsymbol{\Sigma}_{1}^{i}\mathbf{V} \;||\; \mathbf{V}^\top \boldsymbol{\Sigma}_{2}^{i}\mathbf{V}\right) \tag{68}$$

and regularization parameter $\lambda = 1$.

The presented regularized divCSP algorithms can be computed by using KL divergence or beta divergence. As mentioned before the effect of using beta divergence is that it robustly averages the terms in Eq. (63) as it implicitly downweights the influence of outlier terms. This allows us to control (by changing the β parameter) the type of nonstationarity we want to become invariant to, for instance, using beta divergence with small β in divCSP-WS penalizes single extreme events with large deviation from the average activity (as they are not downweighted thus will dominate $\boldsymbol{\Delta}$), e.g., electrode artifacts, whereas larger β parameters penalize more stable variations that occur throughout the experiment. We will discuss this property of beta divergence in the simulation section in more detail. Note that also the multi-subject algorithm divCSP-MS may profit from using beta divergence because integrating data from several subjects usually requires subject selection as different users may have very different signal properties due to differences in head size, electrode montage, state of mind etc. With increasing β parameter we implicitly perform this type of subject selection as the influence of "outlier subjects" with very different signal characteristics will be reduced. On the other hand in some applications we may be interested in the similarity between subjects and may want to identify these “outlier subjects”. Using our framework with a small β parameter enhances the differences between the activity of multiple subjects, thus can be very helpful for exploratory analysis. This property makes beta divergence a very valuable tool for our divergence-based CSP framework.

7.5 SIMULATION STUDIES

This section evaluates the divergence-based framework in several simulation studies. The understanding of the advantages and limitations of the proposed divCSP algorithms helps to select the best method (or the optimal set of parameters) in practice.

7.5.1 *One Step vs. Two Step Methods*

Two-step methods such as SSA+CSP remove information from the data prior to CSP computation. However, if relevant information is removed in the first step these methods are bound to fail (Tomioka and Müller, 2010). The following example shows a situation in which the recently proposed two step method, *SSA+CSP* (von Bünau et al., 2010; Samek et al., 2011), fails to extract the true spatial filters.

Consider the observed signal $\mathbf{x}(t) \in \mathbb{R}^{10}$ generated as mixture of 10 sources $\mathbf{s}(t) = [s_1(t) \dots s_{10}(t)]^\top$ with a random orthogonal mixing matrix $\mathbf{A} \in \mathbb{R}^{10\times 10}$

$$\mathbf{x}(t) \quad = \quad \mathbf{A}\mathbf{s}(t).$$

Assume sources s_1 and s_2 are nonstationary. The signal of the first source is sampled from $\mathcal{N}\left(0, \sigma_1^2\right)$ for class 1 and $\mathcal{N}\left(0, \sigma_2^2\right)$ for class 2, whereas the signal of source s_2 is sampled from $\mathcal{N}\left(0, \sigma_3^2\right)$ irrespectively of class. All the other sources generate normally distributed data with zero mean and unit variance. Now let $\sigma_1^2 = 1.2 + \epsilon_1$, $\sigma_2^2 = 0.8 + \epsilon_1$ and $\sigma_3^2 = 1 + \epsilon_2$ with $\epsilon_1 \sim \mathcal{N}\left(0, \frac{1}{2^2}\right)$ and $\epsilon_2 \sim \mathcal{N}\left(0, \frac{1}{3^2}\right)$ be the variance parameters which are resampled for each trial. In summary, we have constructed a data set with one discriminative and nonstationary source and nine nondiscriminative sources from which one source is also nonstationary. We sample 100 trials per condition, each trial contains 200 ten-dimensional samples, and repeat the experiment 100 times. The goal is to find a spatial filter that recovers the discriminative source s_1.

Figure 34 depicts the angle between the spatial filter computed by divCSP-WS (one-step method) or SSA+CSP (two-step method) and the true projection to the discriminative source s_1. One clearly sees that the two-step method removes the discriminative information in the first step, i.e., the median angle is over 50° when projecting out one or more dimensions. In other words the two-step method projects out (parts of) the activity generated by source s_1 simply because this activity is nonstationary. Thus, SSA+CSP relies on the assumption that the discriminative sources are stationary. If this assumption does not hold this method may fail. On the other hand when applying divCSP-WS we control the strength of regularization. That means we can trade-off stationarity and discriminativity; in real applications some amount of variation will always be present even when projecting to the sources that represent motor imagery related activity. Consequently the simultaneous optimization of stationarity and discriminativity is not only more natural but also allows one to fine tune the amount of stationarity and discriminativity in the features.

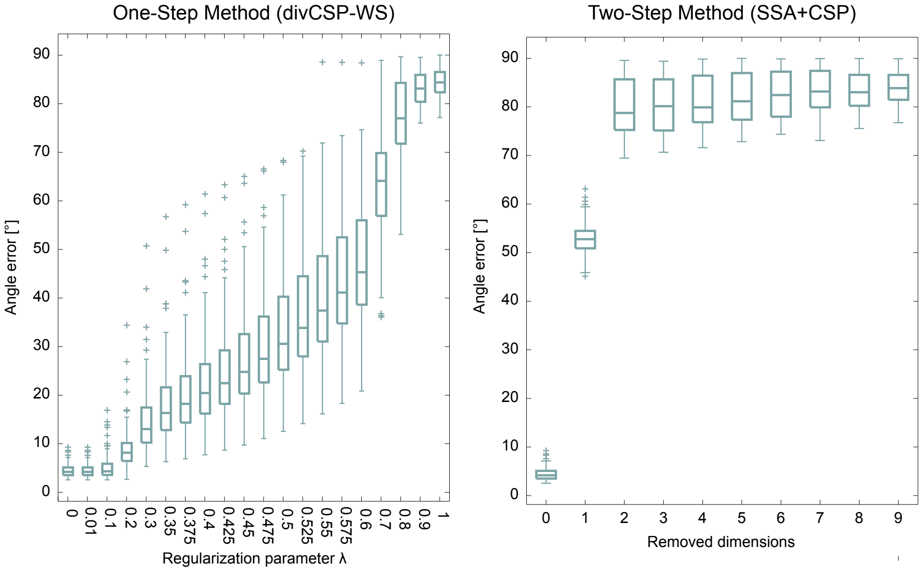

Figure 34.: *Left*: Angle between the true projection and the projection computed by divCSP-WS for various regularization parameters. *Right*: Same quantity for the case of SSA+CSP.

7.5.2 *Joint Diagonalizability*

In Section 4.3 we have shown that sCSP may fail to extract the spatial filter $\mathbf{w} = [w_1 \;\; w_2]^\top = [1 \;\; 0]^\top$ which maximizes the variance ratio between classes while minimizing the nonstationarity of the features. Using the same example

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.9 & 0.15 \\ 0.15 & 0.1 \end{bmatrix}, \; \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.9 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_1^1 = \begin{bmatrix} 0.9 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, \; \boldsymbol{\Sigma}_1^2 = \begin{bmatrix} 0.9 & 0.25 \\ 0.25 & 0.1 \end{bmatrix}$$

we demonstrate that our divergence-based approach (as well as SSA+CSP and KLCSP) penalizes the off-diagonal terms because it does not rely (in contrast to sCSP) on the assumption that the covariance matrices are jointly diagonalizable but rather evaluate nonstationarity in a principled manner using KL divergence. The divCSP-WS method uses the following regularization term $\boldsymbol{\Delta}$

$$\sum_{i=1}^{2} D_{\mathrm{KL}}\left(\mathbf{w}^\top \boldsymbol{\Sigma}_1^i \mathbf{w} \;\|\; \mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}\right)$$

$$= \frac{1}{2}\left(\frac{\mathbf{w}^\top \boldsymbol{\Sigma}_1^1 \mathbf{w}}{\mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}} + \log\left(\frac{\mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \boldsymbol{\Sigma}_1^1 \mathbf{w}}\right) + \frac{\mathbf{w}^\top \boldsymbol{\Sigma}_1^2 \mathbf{w}}{\mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}} + \log\left(\frac{\mathbf{w}^\top \boldsymbol{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \boldsymbol{\Sigma}_1^2 \mathbf{w}}\right)\right) - 1$$

$$= \frac{1}{2}\left(2 + \log\left(\frac{0.9w_1^2 + 0.3w_1w_2 + 0.1w_2^2}{0.9w_1^2 + 0.1w_1w_2 + 0.1w_2^2}\right) + \log\left(\frac{0.9w_1^2 + 0.3w_1w_2 + 0.1w_2^2}{0.9w_1^2 + 0.5w_1w_2 + 0.1w_2^2}\right)\right) - 1.$$

This divCSP-WS penalty term is zero if and only if $w_1 w_2 = 0$, i.e., when disregarding the off-diagonal terms. Thus, divCSP-WS finds the optimal trade-off between stationarity and discriminativity.

7.5.3 *Deflation vs. Subspace Algorithm*

In the following let us apply the multi-subject algorithm divCSP-MS to data of five simulated subjects. As before we use a mixture model with random orthogonal mixing matrix $\mathbf{A}$ to generate the data $\mathbf{x}^j(t)$ of each subject j

$$\mathbf{x}_c^j(t) \sim \mathcal{N}\left(0, \mathbf{A}^\top \mathbf{\Sigma}_c^j \mathbf{A}\right).$$

Let $\mathbf{\Sigma}_c^j = \begin{bmatrix} \mathbf{\Gamma}_c^j & 0 \\ 0 & \mathbf{\Delta}_c^j \end{bmatrix} \in \mathbb{R}^{12\times 12}$ denote the source covariance matrix of class c and subject j with $\mathbf{\Gamma}_c^j \in \mathbb{R}^{2\times 2}$ being the covariance matrix of discriminative sources common to all subjects and $\mathbf{\Delta}_c^j \in \mathbb{R}^{10\times 10}$ the corresponding subject specific matrix. Let the first two sources of all subjects be discriminative but have different correlations. In other words we simulate the case where the projections which reconstruct the two (independent) discriminative sources of subject i will reconstruct a linear mixture of the discriminative sources of subject j. Thus, the discriminative sources of subject i and j lie in the same subspace but have different correlations. This may happen when e.g. the mixing matrix of subject i is a rotated version of the mixing matrix of subject j, e.g., because of tiny differences in electrode position or head size. For simplicity let us assume the mixing matrix is fixed for all subjects, but the correlations between the sources differ. The goal of a multi-subject algorithm is to extract discriminative activity common to all subjects, i.e., to extract the first two sources.

Let the first two sources of subject 1 be generated by a zero mean Gaussian with variance 1.5 and 0.5 for condition one and variance 0.5 and 1.5 for condition two, i.e., $\mathbf{\Gamma}_1^1 = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix}$ and $\mathbf{\Gamma}_2^1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1.5 \end{bmatrix}$. The covariance matrices $\mathbf{\Gamma}_c^j$ for the other subjects have same structure as for subject one, but are rotated by a (random) rotation matrix with angle $\alpha \in [-90^\circ\ 90^\circ]$. The first row of Figure 35 visualizes a possible data distribution of the first two sources for three subjects.

Note that the first two sources are discriminative for all subjects, thus, they should be recovered by multi-subject CSP algorithms. However, when applying divCSP-MS in deflation mode the (single) filter which separates class one and two for subject 1 may not separate the classes for the other subjects as their source activity is rotated (see first row of Figure 35). Only when extracting the whole subspace, i.e., sources one and two, the algorithm "realizes" that these subspaces are equivalently discriminative for all subjects. Thus, only a subspace method helps for these kind of data integration problems. Note that the constructed example is equivalent to the well-known XOR problem in feature selection literature (Guyon and Elisseeff, 2003).

Now let us assume that every subject has two other discriminative sources with variance 1.6 / 0.4 and 0.4 / 1.6 in condition one and two, respectively. However, these sources are at random positions in $\mathbf{\Delta}^j$, i.e., they are not necessarily at the same position for all subjects. All other sources in $\mathbf{\Delta}^j$ are nondiscriminative, i.e., are sampled from a Gaussian with variance 1 irrespectively of condition. The second row of Figure 35 illustrates a case where the sources are discriminative for subject 1, but the subspace is not discriminative for the other subjects.

The plot at the bottom of Figure 35 displays the results of applying divCSP-MS (100 repetitions) in deflation (brown line) and in subspace (cyan line) mode to the data of subject 1. With increasing regularization parameter the algorithms utilize information from the four additional subjects. The plot depicts the median of the largest principal angles between the true filters capturing the activity of the first two sources and the filters computed by divCSP-MS. One clearly sees that for small regularization parameters λ (i.e. when only using data from subject 1) both methods do not reconstruct the activity of the common subspace (first row). This is because the subject specific activity (second row) is simply more discriminative, the subject specific sources have a variance ratio of 1.6 / 0.4 compared to 1.5 / 0.5. However, with increasing regularization, i.e., when taking into account other subjects' data the subspace method "realizes" that there is a subspace which is discriminative for all users, thus, this subspace is preferred and the angle error decreases to $0°$. On the other hand the deflation method does not reconstruct the common subspace because it is not able to utilize the XOR-like structure.

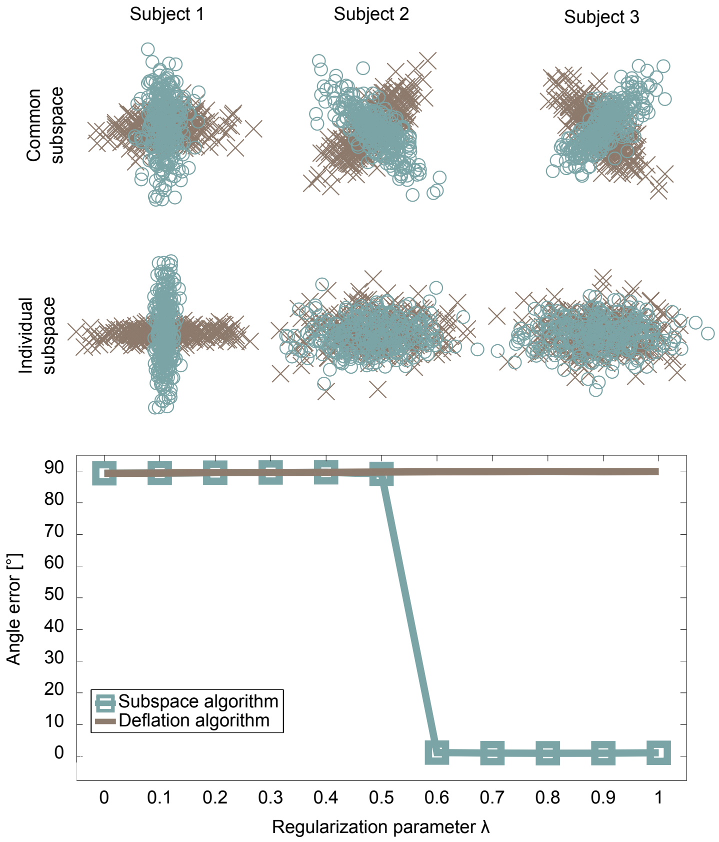

Figure 35.: *First row*: Example of a distribution which is discriminative for all three subjects when considering the whole subspace, but is not discriminative when considering individual directions. *Second row*: Example of a distribution which is only discriminative for the first subject. *Third row*: Angle between the projection to the common discriminative subspace and the projections computed by the subspace and deflation divCSP-MS algorithm.

7.5.4 *Effects of Beta Divergence*

In the following we investigate the influence of the β parameter on the type of stationarity enforced by divCSP-WS. Let us consider the two types of changes displayed in Figure 36, namely gradual changes and abrupt changes. The first row of Figure 36 visualizes the data distributions of five epochs that change gradually. We denote the covariance matrix of the ith epoch as $\boldsymbol{\Sigma}_1^i$. The second row of Figure 36 displays four relatively stable (stationary) distributions and one extreme change. We denote the covariance matrix of the ith epoch as $\mathbf{\Sigma}_2^i$. We measure both types of variations as average divergence between the data distribution in the first epoch (reference) and the four subsequent distributions. The bottom row of Figure 36 plots the ratio of the divergence terms computed on the examples in the second row and the first row for various β parameters, i.e.,

$$r = \frac{\sum_{i=2}^{5} \tilde{D}_\beta \left(\mathbf{\Sigma}_2^i \parallel \mathbf{\Sigma}_2^1\right)}{\sum_{i=2}^{5} \tilde{D}_\beta \left(\mathbf{\Sigma}_1^i \parallel \mathbf{\Sigma}_1^1\right)}$$

Note that we change the scale of the x-axis at zero (the reason for this sudden drop) as setting β to lower values than -0.0115 results in numerical problems (see derivation of beta divergence in appendix B.2). Note that if the ratio of the divergences is above 1 then the abrupt change is regarded as more nonstationary than the gradual change by the algorithm; the opposite holds if the value is below 1. Thus, by using beta divergence we have an additional degree of freedom, namely we can shift the focus from gradual changes which are relatively stable over the data set to strong abrupt events such as electrode artifacts. Thus, we can easily match the types of nonstationarities which are present in the data and compute invariant features. This flexibility can also be utilized for exploratory analysis, i.e., identification of gradual changes.

In the next experiment we investigate the impact of strong artifactual trials on CSP and demonstrate the robustness property of β-divCSP. For that we generate data $\mathbf{x}(t)$ using the following mixture model

$$\mathbf{x}(t) = \mathbf{A} \begin{bmatrix} \mathbf{s}^{dis}(t) \\ \mathbf{s}^{ndis}(t) \end{bmatrix} + \xi \tag{69}$$

where $\mathbf{A} \in \mathbb{R}^{10\times10}$ is a random orthogonal mixing matrix, $\mathbf{s}^{dis}$ is a discriminative source sampled from a zero mean Gaussian with variance 1.8 in one condition and 0.2 in the other condition, $\mathbf{s}^{ndis}$ are 9 sources with variance 1 in both conditions and ξ is a noise variable with standard deviation 2. We generate 100 trials per condition, each consisting of 200 data points. Furthermore, we randomly add artifacts with standard deviation 10 independently to each data dimension (i.e. virtual electrode) and trial with varying probability and evaluate the angle between the true filter extracting the source activity of $\mathbf{s}^{dis}$ and the spatial filter computed by CSP and β-divCSP. The median angles computed from 100 repetitions of the simulation experiment are visualized in Figure 37. One clearly sees that the angle error between the spatial filter extracted by CSP and the true filter increases with higher artifact probability (brown line). Furthermore, one can see from the figure that using very small β values does not attenuate the artifact problem, but it rather increases the error by adding up trial-wise divergences without downweighting outliers. However, as the β value increases the artifactual trials are downweighted and a robust average is computed over the trial-wise divergence terms. This increased robustness significantly reduces the angle error.

7.5.5 *KL Divergence vs. Symmetric KL Divergence*

In the following we want to touch upon the difference between the symmetric KL divergence and the KL divergence. The KL divergence between two zero-mean Gaussians with covariances $\mathbf{A}$ and $\mathbf{B}$ can be written in explicit form as

$$D_{kl}\left(\mathbf{A} \parallel \mathbf{B}\right) = \log\left|\mathbf{A}^{-1}\mathbf{B}\right| + \mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{A}\right), \tag{70}$$

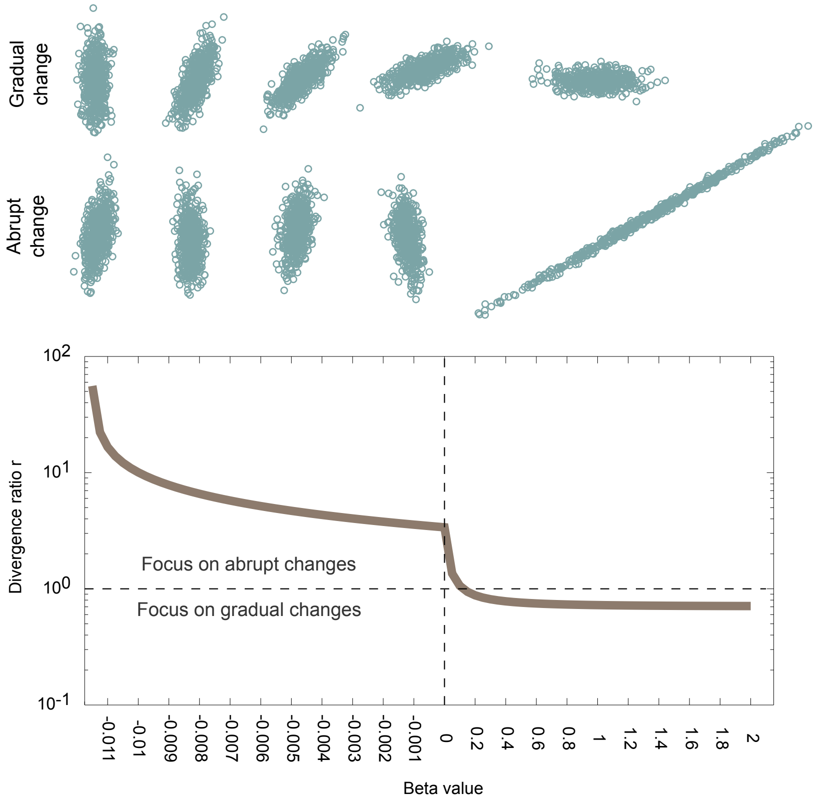

Figure 36.: *First row*: Example of distributions which change gradually. *Second row*: Example of four relatively stable distributions and one distribution which is very different from the rest. *Third row*: Ratio of the divergence terms computed from the distributions in the second and first row. Each term is the average symmetric KL divergence between the distribution in the first epoch and the distributions in the other epochs. Thus, if the curve is above 1 (= 10^0) then abrupt changes are preferred, i.e., the divergence term computed from the distributions in the second row is higher than the one computed from distributions in the first row, whereas if it is below 1 we give higher regularization to the gradual change. Thus, by changing the beta parameter we can shift the focus from abrupt changes to gradual changes.

whereas its symmetric counterpart is

$$\tilde{\mathrm{D}}_{\mathrm{kl}}\left(\mathbf{A} \,\|\, \mathbf{B}\right) = \mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{B}\right) + \mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{A}\right). \tag{71}$$

From linear algebra (Bhatia, 1997) the following relation is known

$$\log|\mathbf{A}| = \mathrm{tr}(\log(\mathbf{A})). \tag{72}$$

Using this relation we can rewrite the KL divergence objective function as

$$\mathrm{D}_{\mathrm{kl}}\left(\mathbf{A} \,\|\, \mathbf{B}\right) = \mathrm{tr}\left(\log\left(\mathbf{A}^{-1}\mathbf{B}\right)\right) + \mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{A}\right) \tag{73}$$

Thus, the difference between both divergences is the log operator inside the first trace term. This log operation downweights the influence of the $\mathrm{tr}\left(\log\left(\mathbf{A}^{-1}\mathbf{B}\right)\right)$ term compared to $\mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{B}\right)$ when the eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$ are very large. The question is when does such an operation make sense ?

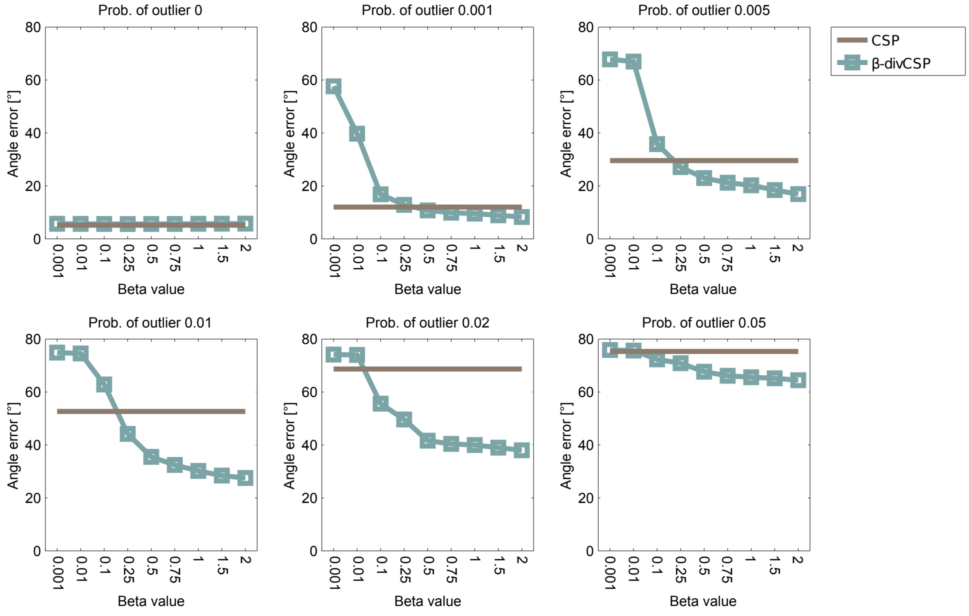

Figure 37.: Angle between the true spatial filter and the filter computed by CSP and β-divCSP for different probabilities of artifacts. The robustness of our approach increases with the β value and significantly outperforms the CSP solution.

When $\mathbf{A}$ is ill-conditioned it may have eigenvalues close to zero. In this case the term $\operatorname{tr}\left(\mathbf{A}^{-1}\mathbf{B}\right)$ becomes very large, consequently it will dominate the solution. Using the (non-symmetric) KL divergence significantly reduces the influence of the ill-conditioned matrix $\mathbf{A}$. Thus, using the log operator makes perfectly sense in the divCSP-WS algorithm as it operates on trial-wise covariance matrices that may be poorly estimated. In this case the KL divergence should be preferred. On the other hand when using average matrices as in divCSP-BS, divCSP-AS or divCSP-MS there is no reason to downweight one term of the divergence, thus the symmetric divergence should be applied.

7.5.6 *Matrix Regularization vs. Divergence Regularization*

The covCSP algorithm (see Section 3.5) applies regularization to the covariance matrices prior to CSP computation. For the single filter case one can represent the covCSP objective as

$$\tilde{D}_{kl}\left(\mathbf{v}^\top\left((1-\lambda)\mathbf{\Sigma}_1^\ell + \lambda\tilde{\mathbf{\Sigma}}_1\right)\mathbf{v} \;\|\; \mathbf{v}^\top\left((1-\lambda)\mathbf{\Sigma}_2^\ell + \lambda\tilde{\mathbf{\Sigma}}_2\right)\mathbf{v}\right) \qquad (74)$$
$$= \frac{\mathbf{v}^\top((1-\lambda)\mathbf{\Sigma}_1^\ell + \lambda\tilde{\mathbf{\Sigma}}_1)\mathbf{v}}{\mathbf{v}^\top((1-\lambda)\mathbf{\Sigma}_2^\ell + \lambda\tilde{\mathbf{\Sigma}}_2)\mathbf{v}} + \frac{\mathbf{v}^\top((1-\lambda)\mathbf{\Sigma}_2^\ell + \lambda\tilde{\mathbf{\Sigma}}_2)\mathbf{v}}{\mathbf{v}^\top((1-\lambda)\mathbf{\Sigma}_1^\ell + \lambda\tilde{\mathbf{\Sigma}}_1)\mathbf{v}},$$

with $\tilde{\mathbf{\Sigma}}_c = \frac{1}{K}\sum_{k=1}^K \mathbf{\Sigma}_c^k$ being the average covariance matrix computed on other subjects' data. In the case of divCSP-AS regularization is applied to the divergences (i.e.

after the nonlinear divergence function has been applied). For the single filter case the divCSP-AS method maximizes the following term

$$\begin{aligned}
&(1-\lambda)\tilde{D}_{kl}\left(\mathbf{v}^\top\boldsymbol{\Sigma}_1^\ell\mathbf{v} \,\|\, \mathbf{v}^\top\boldsymbol{\Sigma}_2^\ell\mathbf{v}\right) - \lambda\frac{1}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\tilde{D}_{kl}\left(\mathbf{v}^\top\boldsymbol{\Sigma}_c^k\mathbf{v} \,\|\, \mathbf{v}^\top\boldsymbol{\Sigma}_c^\ell\mathbf{v}\right) \qquad (75)\\
&= (1-\lambda)\left[\frac{\mathbf{v}^\top\boldsymbol{\Sigma}_1^\ell\mathbf{v}}{\mathbf{v}^\top\boldsymbol{\Sigma}_2^\ell\mathbf{v}} + \frac{\mathbf{v}^\top\boldsymbol{\Sigma}_2^\ell\mathbf{v}}{\mathbf{v}^\top\boldsymbol{\Sigma}_1^\ell\mathbf{v}}\right] - \lambda\frac{1}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\left[\frac{\mathbf{v}^\top\boldsymbol{\Sigma}_c^\ell\mathbf{v}}{\mathbf{v}^\top\boldsymbol{\Sigma}_c^k\mathbf{v}} + \frac{\mathbf{v}^\top\boldsymbol{\Sigma}_c^k\mathbf{v}}{\mathbf{v}^\top\boldsymbol{\Sigma}_c^\ell\mathbf{v}}\right].
\end{aligned}$$

We compare the behaviour of both objective functions in a simulation experiment for $\mathbf{v} \in [-1 \;\; 1] \times [-1 \;\; 1]$ and various λ parameters. Let $\ell = 1, K = 2$ and

$$\begin{aligned}
\boldsymbol{\Sigma}_1^1 &= \begin{bmatrix} 1.3 & 0 \\ 0 & 1.0 \end{bmatrix}, & \boldsymbol{\Sigma}_2^1 &= \begin{bmatrix} 0.7 & 0 \\ 0 & 1.0 \end{bmatrix}\\
\boldsymbol{\Sigma}_1^2 &= \begin{bmatrix} 1.3 & 0 \\ 0 & 1.5 \end{bmatrix}, & \boldsymbol{\Sigma}_2^2 &= \begin{bmatrix} 0.7 & 0 \\ 0 & 0.5 \end{bmatrix}
\end{aligned}$$

One can see in Figure 38 that both objective functions are not the same. Note that the objective value increases with the intensity of the blue color. The covCSP method prefers (i.e. assigns larger values to) the spatial filter $\mathbf{v}^\top = [1 \;\; 0]$ for small λ parameters but if the regularization of the covariance matrices $\boldsymbol{\Sigma}_1^1$ and $\boldsymbol{\Sigma}_2^1$ increases (large λ) the preference switches to $\mathbf{v}^\top = [0 \;\; 1]$ because the second source of subject 2 is more discriminative than the first source (ratio 1.5/0.5 compared to 1.3/0.7). The divCSP-AS regularization on the other hand aims to find discriminative sources which are similar between both subjects, thus it always prefers the filter $\mathbf{v}^\top = [1 \;\; 0]$ that extracts source 1 which is discriminative and common to both users. Depending on the particular application scenario, i.e., whether we want to find the most discriminative source or a source that is discriminative and common among subjects, the covCSP regularization or the divCSP-AS regularization scheme may be the method of choice.

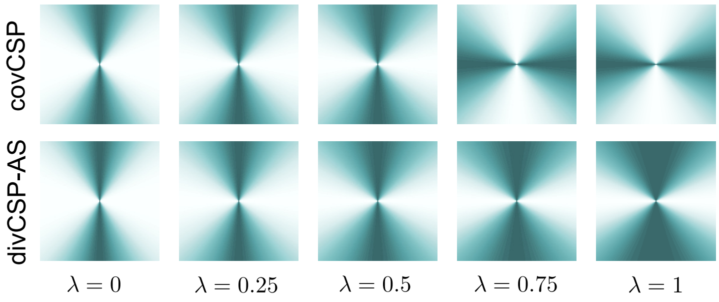

Figure 38.: *Top row*: Values of covCSP objective function in Eq. (74) for $\mathbf{v} \in [-1 \;\; 1] \times [-1 \;\; 1]$ and various λ parameters. The objective value increases with the intensity of the blue color. *Bottom row*: An analogous plot for the divCSP objective function in Eq. (75) and various λ parameters.

7.6 experimental evaluation

This section evaluates the divergence-based CSP approaches introduced in this chapter and compares them to several state-of-the-art baselines. Due to the large number of experiments we restrict the evaluation to the Vital BCI data set.

7.6.1 *Robustness to Artifacts*

We compare the β-divCSP method with three baseline approaches, namely CSP, shrinkCSP and MCDE+CSP. The parameter β is selected from the set of 15 candidates {0, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1, 1.5, 2, 5} by 5-fold cross-validation on the calibration data using minimal training error rate as selection criterion. For faster convergence we use the rotation part of the CSP solution as initial rotation matrix. The results are displayed in Figure 39. Each circle represents the error rate of a subject. One can see that the β-divCSP method outperforms the baselines as most circles are below the solid line. Furthermore, the performance increases are significant according to the one-sided Wilcoxon sign rank test as the p-values are smaller than 0.05.

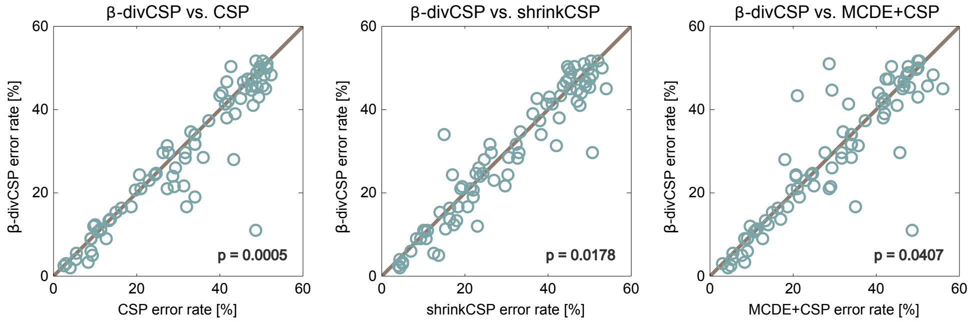

Figure 39.: Comparison of β-divCSP to three baselines. Each circle represents the error rate of one subject. Our method outperforms the baselines for circles that are below the solid line. The p-values of the one-sided Wilcoxon sign rank test are displayed in the lower right corner.

We made an interesting observation when analysing the subject with largest improvement over the CSP baseline; the error rates were 48.6% (CSP), 48.6% (MCDE+CSP) and 11.0% (β-divCSP). Over all ranges of MCDE+CSP parameters this subject has an error rate higher than 48% i.e. MCDE+CSP was not able to properly estimate the covariance matrices. This example demonstrates that β-divCSP and MCDE+CSP are not equivalent. Enforcing robustness on the CSP algorithm may in some cases be better than enforcing robustness when estimating the covariance matrices.

In the following we study the robustness property of the β-divCSP method on subject 74, the user with the largest improvement. The left panel of Figure 40 displays the activity pattern associated with the first CSP filter $\mathbf{w}_1$ of subject 74. One clearly sees that the pattern does not encode neurophysiologically relevant activity, but focuses on a single electrode, namely FFC6. When analysing the (filtered) EEG signal of this electrode one can identify a strong artifact in one of the trials. Since neither the empirical covariance estimator nor the CSP algorithm is robust to this type of outliers, the artifact dominates the solution. However, the resulting pattern is meaningless as it does not represent motor imaginary related activity. The right panel of Figure 40 displays the relative importance of the divergence term of the artifactual trial i^* with respect to the average divergence terms of the other trials $i = 1 \dots n, i \neq i^*$. This ratio is

$$r = \frac{\tilde{D}_\beta\left(\mathbf{w}_1^\top \boldsymbol{\Sigma}_1^{i^*} \mathbf{w}_1 \;\|\; \mathbf{w}_1^\top \boldsymbol{\Sigma}_2^{i^*} \mathbf{w}_1\right)}{\frac{1}{n-1}\sum_{i=1\dots n, i\neq i^*}^{n} \tilde{D}_\beta\left(\mathbf{w}_1^\top \boldsymbol{\Sigma}_1^{i} \mathbf{w}_1 \;\|\; \mathbf{w}_1^\top \boldsymbol{\Sigma}_2^{i} \mathbf{w}_1\right)}$$

One can see that the divergence term computed from the artifactual trial is over 1800 times larger than the average of the other trials. This ratio decreases rapidly for larger β values, i.e., the influence of the artifact decreases. This increased robustness is the reason for the superior performance of β-divCSP.

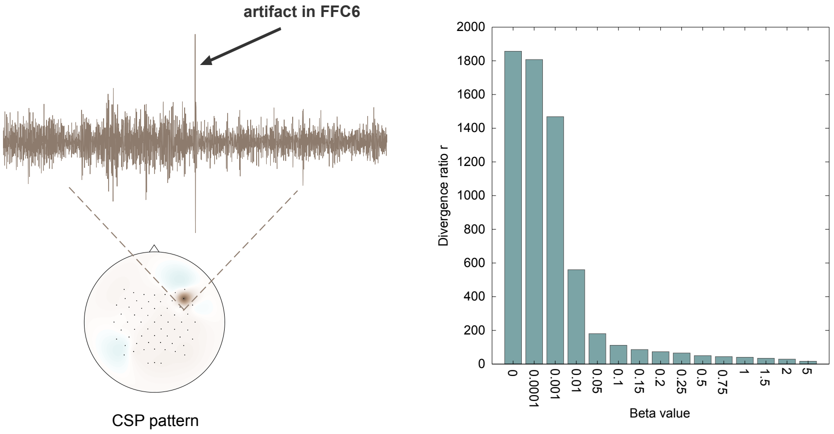

Figure 40.: *Left*: The CSP pattern of subject 74 does not reflect neurophysiological activity but it represents the artifact in electrode FFC6. *Right*: The relative importance of the artifactual trial decreases with the β parameters. The relative importance is measured as ratio between the divergence term of the artifactual trial and the average divergence terms of the other trials.

7.6.2 *Reducing Within-Session Nonstationarity*

In the next experiment we aim to increase the stationarity of the training features by applying divCSP-WS. In order to capture different types of variations, both single extreme events and common slow changes, we test our algorithm with various beta parameters. We use $\beta = 0, 0.5, 1$ and the minimal possible negative value from $-0.0005, -0.0010, -0.0015, \ldots$. We select the best of these four β values for each subject by applying cross validation. Figure 41 displays the error rates of all subjects for the subspace and deflation divCSP-WS method and compares them to CSP (first column), SSA+CSP (second column) and KLCSP (third column). Note that we did not reimplement the original KLCSP algorithm, but use the deflation divCSP-WS algorithm with $\beta = 0$ as both algorithms optimize the same objective. Each circle in the scatter plot represents the error rate of one subject and the number in the lower right corner denotes the p-value when applying the one-sided Wilcoxon sign-rank test. The error rate of our approach is represented on the y-axis, i.e., if the circle is below the solid line then our method outperforms the baseline for this subject. The null hypothesis of the Wilcoxon test states that the median of the error rate differences (our method (y-axis) - baseline method (x-axis)) is greater than or equal to zero. For $p < 0.05$ we reject this null hypothesis, thus we say that our method significantly outperforms the baseline.

One can see from the plot that the deflation divCSP-WS outperforms the subspace method. It significantly decreases classification error rates in comparison to CSP ($p = 0.0481$); the subspace approach does not show any improvement. The subspace method performs poorly as it considers changes in correlations between different spatial filters. These correlations are ignored in the feature extraction and classification process, thus should not be considered when computing the spatial filters. One can also see from the results that the simultaneous optimization of two objectives, discriminativity and stationarity in this case, is superior to the sequential optimization as done by the two-step SSA+CSP approach. The improvement of the deflation divCSP-WS algorithm over SSA+CSP is very close to being significant ($p = 0.0526$). The observation that one step methods outperform two-step approach is in line with the simulations performed in the last section. The fact that two-step methods may remove information in the first step which is important for the second step is a significant disadvantage of these approaches. We will comment on this in the next paragraph. The scatter plots in the last column demonstrate the advantage of using the beta divergence version of our algorithms. The results show that the Kullback-Leibler divergence algorithm (as in (Arvaneh et al., 2013a)) performs worse than our deflation divCSP-WS method and the difference between both algorithms is close to being significant ($p = 0.0750$). The improvement is due to the additional flexibility of beta divergence; it may capture a whole range of different nonstationarities. On the other hand one can see that KLCSP significantly outperforms the subspace divCSP-WS method ($p = 0.9814$). Thus, the additional flexibility of using different beta values does not compensate for the disadvantage of considering nonstationarities in correlation.

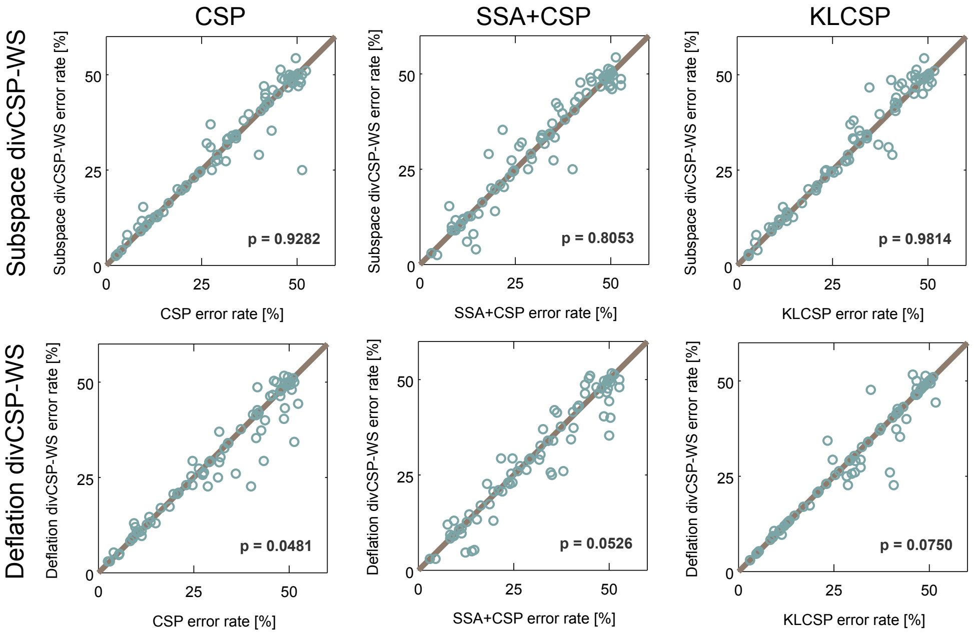

Figure 41.: Scatter plots showing error rates of subspace and deflation divCSP-WS and three baseline methods. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner.

Above we discussed an example where two-step methods provide suboptimal performance. Figure 42 depicts the boxplot of the difference in error rate between SSA+CSP (Samek et al., 2011) and CSP. One can see from the figure that the classification performance of SSA+CSP drops with increasing number of removed dimensions. This means that the directions removed in the first step of SSA+CSP contain increasing amount of discriminative information (which is required for the second step). The two scalp plots visualize the activity patterns corresponding to the removed directions for two subjects. One clearly sees that the upper scalp plot shows activity over the left motor and temporal cortex. Since such activity contains motor imagery related information (right hand class) it is not advisable to remove it. Since SSA only evaluates the amount of nonstationarity and does not take into account the information content it removes this activity, thus the corresponding subject shows a significant increase in error rate, namely from 9.3 % to 18.3 %. The lower scalp plot corresponds to a subject which improves classification accuracy (from 40 % to 18 %) by applying SSA preprocessing. One can see that some temporal activity is removed from this subjects data.

Since this information is not motor imagery related it may be safely removed. This example demonstrates that two-step methods may fail in practice. Although the authors of (Samek et al., 2012a) propose to trade-off nonstationarity and discriminativity when using SSA, we emphasize the limits of applying two step approaches for feature extraction in BCI.

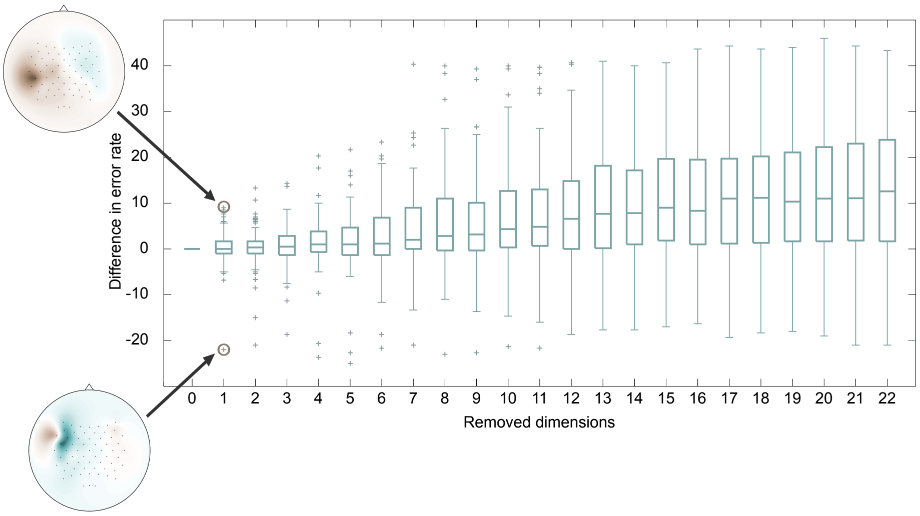

Figure 42.: Change in error rate when removing $0 \ldots 22$ dimensions from the data by applying SSA. One can see that the error rates significantly increase as more dimensions are removed. The two scalp plots visualize the activity patterns corresponding to the removed direction. One clearly sees that for the subject with increasing error rate the (upper) scalp plot shows activity related to motor imagery. Since this information should not be removed, SSA+CSP increases the error rate for this subject.

The effects of various beta parameters are studied in Figure 43. The upper panel displays a subject's EEG signal with a strong artifact in electrode FFC6. The three scalp plots at the bottom panel visualize the activity patterns of the first spatial filter extracted by divCSP-WS with $\lambda = 0.5$ and various beta values. One clearly sees that for $\beta = 0$ (left scalp plot) there is no regularization towards stationarity, therefore, the pattern focuses on the activity in FFC6 (due to the strong artifact) and does not represent motor imagery related information. Thus, the area under the ROC curve (AUC) value is low (0.55). If using a beta value of 1 (right scalp plot) there is an improvement, i.e., a right hand motor imagery pattern emerges, however, the focus on the electrode FFC6 is still present. This is because larger β values downweight the influence of the artifactual trial in the penalty term $\boldsymbol{\Delta}$, thus the regularization does not penalize strong extreme events such as the artifact in FFC6. The situation changes if $\beta < 0$ as then we enhance the extreme values in the penalty term $\boldsymbol{\Delta}$ computation, i.e., the artifact dominates the penalty term thus is substantially more strongly penalized in the optimization process. The effect of this penalty is that a true motor imagery related pattern emerges and the focus on electrode FFC6 vanishes. The AUC value of this pattern also largely increases to 0.87. We have demonstrated a similar effect in the toy simulations in last section. This additional degree of freedom renders our method(s) much more flexible than, e.g., KLCSP or ssCSP.

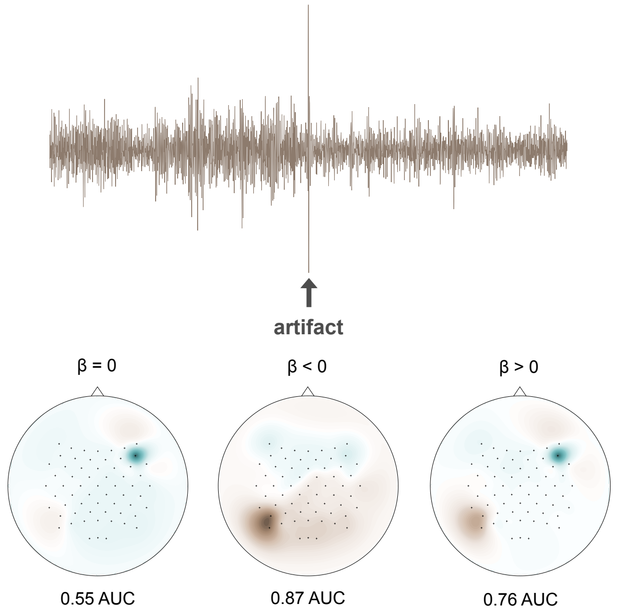

Figure 43.: *Top row*: Example of an artifact in the signal of the FFC6 electrode. *Bottom row*: Activity patterns computed by divCSP-WS with $\lambda = 0.5$. One can see that the regularization (minimizing the effect of FFC6 on the solution) only works properly if $\beta < 0$ as this enhances the artifactual activity and thus increases its relative penalty.

7.6.3 *Reducing Between-Session Shifts*

In this subsection we describe several between-session information transfer experiments using divCSP-BS. As before we apply the subspace and deflation algorithm and use the beta values $0, 0.5$ and 1. We compare the results to CSP, to the ssCSP method and to divCSP-BS with $\beta = 0$. Note that we only integrate information from additional subjects with the same motor imagery classes and select the regularization parameters by minimizing test error on the other subjects' data. The results in the first row of Figure 44 show a performance improvement of the deflation divCSP-BS method over CSP. Although there is a trend the difference is not statistically significant ($p = 0.0938$). This confirms the observation that information about shifts between sessions can be transferred across subjects. In contrast to the within-session nonstationarities presented in the last subsection we are not so much interested in single extreme nonstationarities between training and test sessions, but rather in changes which are stable over subjects. By using $\beta > 0$ we penalize these (common) changes between calibration and feedback. However, it seems that using different beta parameters does not have a large impact on the results (see last column). The second column of Figure 44 compares divCSP-BS to ssCSP. Although we compute the shift between calibration and feedback session for each class separately, our method does not outperform ssCSP which does not use class labels ($p = 0.5377$). This suggests that the nonstationarities between calibration and feedback session are not class-dependent.

The upper plot in Figure 45 displays the median (over subjects) KL divergence differences between CSP (no regularization) and deflation divCSP-BS with increasing regularization. The divergence is computed between the calibration and feedback feature distribution when applying the filters computed by divCSP-BS with increasing λ. One can see from the plot that incorporating information from other users about the shift between calibration and feedback constantly reduces this shift on the subject of interest. This confirms our observation that nonstationarities are similar between different

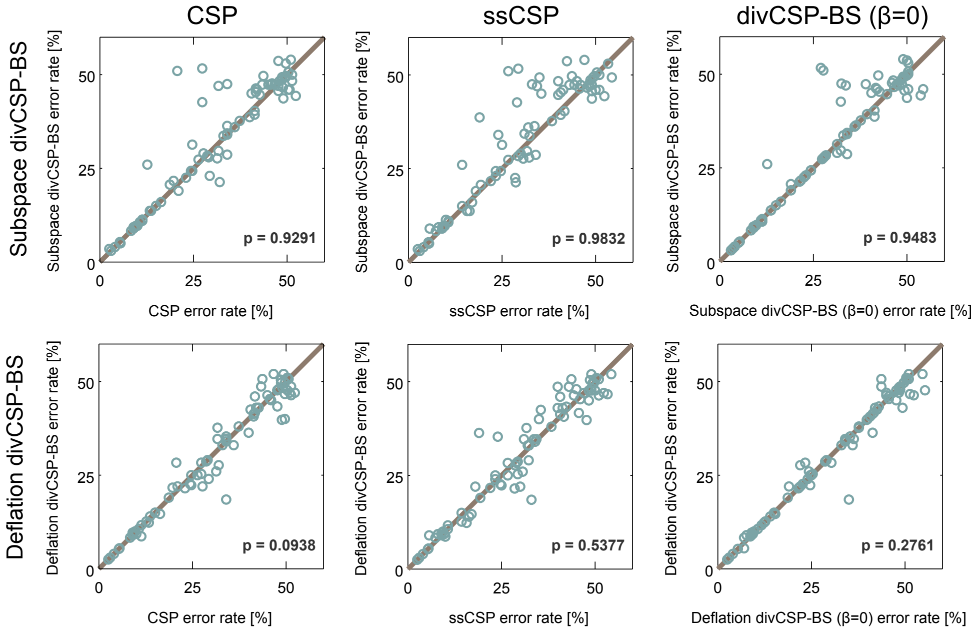

Figure 44.: Scatter plots showing error rates of deflation and subspace divCSP-BS and three baseline methods. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner.

subjects. The lower panel of the figure visualizes the effect of applying divCSP-BS. It depicts the feature distribution of the 'left hand class' train data (cyan circles) and the corresponding test data (brown crosses) of subject 13. The six dimensional feature distribution is projected to two dimensions by using the largest PCA component and the normal vector to the classification hyperplane. One can see that when applying CSP there is a large shift in the distribution between training and test. If on the other hand incorporating information from additional subjects one obtains a stationary distribution with no significant shift between training and test.

7.6.4 *Stationarity Across Subjects*

This subsection discusses the results of divCSP-AS; as before we use the beta values $0, 0.5$ and 1. Figure 46 displays the results of both the subspace and deflation algorithm and compares them to CSP, covCSP and klcovCSP. One can see (first column) that divCSP-AS significantly outperforms CSP ($p < 10^{-4}$), i.e., regularizing the feature distribution towards the feature distribution of the other subjects seems to have a strong effect on the quality of the spatial filters. This regularization effect is stronger than when regularizing the covariance matrices towards other subjects as done by covCSP ($p = 0.0626$) and klcovCSP ($p = 0.1120$).

Figure 47 evaluates the improvement of subject 74, the user with largest decrease in error rate. The lower boxplot visualizes the distribution of the KL divergence between subject 74 and the other subjects when applying the first spatial filter computed by divCSP-AS with increasing regularization parameters. One can see that there is a large gap when moving from $\lambda = 0.2$ to $\lambda = 0.3$, i.e., the feature distribution of subject 74 becomes similar to the distribution of other subjects. Above the boxplot we visualize the activation patterns of the first spatial filter computed with divCSP-AS. One clearly sees the electrode artifact in FFC6 (see also Figure 43). Since this activity is not present

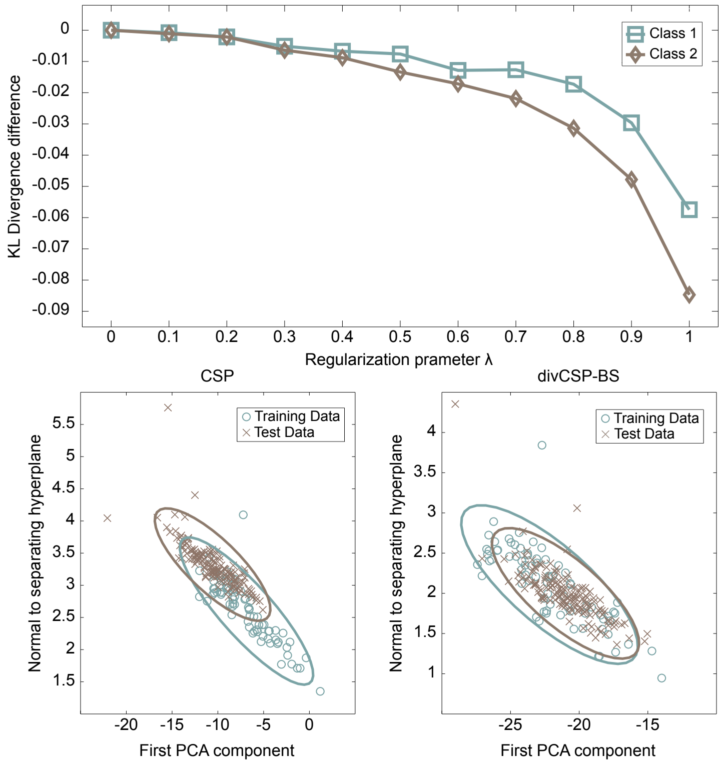

Figure 45.: *Top row*: Median KL divergence difference between deflation divCSP-BS with increasing regularization and CSP. The divergence is computed between the calibration and feedback feature distribution. One can see that the divergence decreases with increasing regularization. *Bottom row*: The 'left hand' feature distribution of training data (cyan circles) and test data (brown circles) when applying CSP and divCSP-BS. The features are projected to the largest PCA component and the normal vector to the classification hyperplane. One clearly sees that the divCSP-BS solution provides substantially more stationary feature distributions than CSP.

in the other subjects data it is penalized when applying divCSP-AS. For regularization parameters larger than $\lambda = 0.3$ it completely vanishes. In other words regularizing the feature distribution towards other subjects helps in removing this type of anomalies. Note also that for $\lambda > 0.5$ the activity patterns show strong activation in motor imagery related areas. This activation is captured by divCSP-AS as it is present in all subjects (having the same classes as subject 74).

7.6.5 *Subject-Independent Spatial Filters*

In the last subsection we perform regularization towards other subjects, here we aim to use other subjects' data to extract a subject independent feature space. Therefore, we apply divCSP-MS, covCSP and klcovCSP with $\lambda = 1$. In other words we estimate the spatial filters by using other subjects' data only. Note that we still use the calibration data to train the LDA classifier, only the spatial filters are "subject independent". As before we apply our algorithm with the three beta parameters $0, 0.5$ and 1. Figure 48 compares the error rates of the subspace and deflation divCSP-MS algorithm with three baselines. One clearly sees that both the subspace and deflation divCSP-MS provide better feature spaces than covCSP and klcovCSP. The improvement is statistically significant for the subspace algorithm with $p = 0.0004$ when comparing its

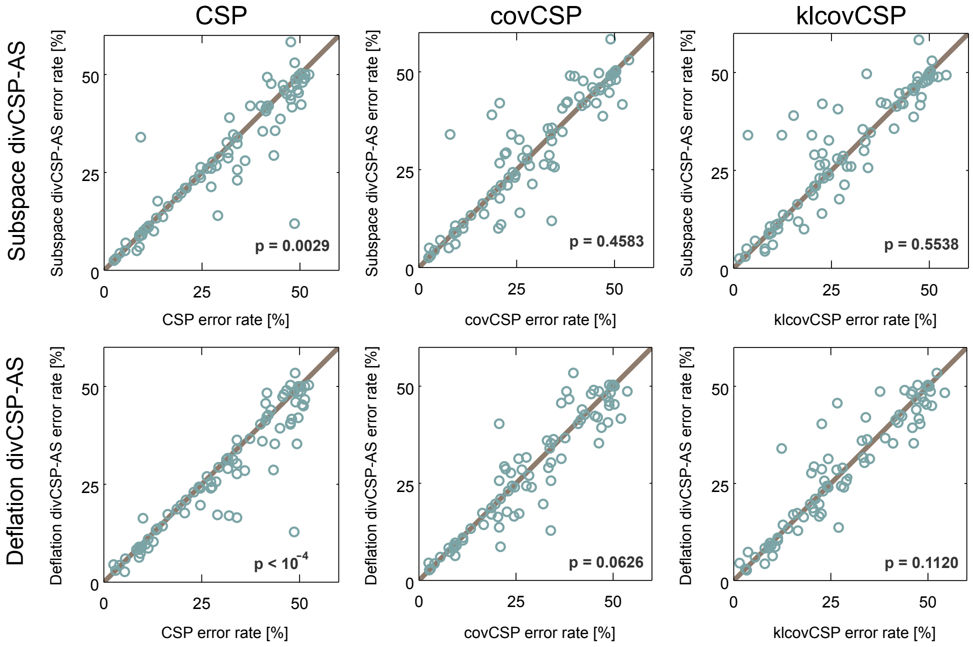

Figure 46.: Scatter plots showing error rates of deflation and subspace divCSP-AS and three baseline methods. Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner.

performance to covCSP and $p = 0.0147$ when comparing to klcovCSP. The subspace method performs significantly better than covCSP ($p = 0.0193$), the improvement over klcovCSP is not significant ($p = 0.2105$). This means that integrating information from additional subjects by combining divergence terms which measure motor imagery related activity is superior to combining the covariance matrices, i.e., fusing all information. As observed in the simulations the subspace method is (slightly) better than the deflation approach. The subspace method is not affected by changes in correlation, thus it identifies the common subspace even when differences in correlation of the sources exist between subjects. We also see (third column) that using beta divergence significantly improves the algorithm, the p-value for the subspace approach is 0.0101, for the deflation method it is smaller than 10^{-4}. It seems that beta values larger than zero have a positive effect on performance as they downweight the influence of individual subjects and help to extract common motor imagery related activity.

A direct comparison of the subspace and deflation method for the three beta values is shown in the upper panel of Figure 49. For the case of $\beta = 0$ one can see a clear advantage of the subspace method ($p = 0.0001$). As demonstrated in the simulations (see Figure 35) the deflation approach may prefer single-subject solutions as it does not capture common activity if the correlations of the sources vary between subjects. However, the relative gain of these single-subject solutions decreases with increasing beta value (because of downweighting effect), therefore the subspace and deflation algorithms perform similarly for $\beta = 1$. The lower panel of Figure 49 compares the subject independent feature spaces computed by divCSP-MS (after selecting β by cross-validation) to the CSP solution when computed on increasing number of trials. For that we randomly select $n = 2 \dots 75$ trials per class from the calibration data and compute CSP on this smaller data set. Afterwards we train the LDA classifier on the whole calibration data and apply it to the feedback data. We repeat this 50 times. In the left boxplot one can see clearly that the subject-independent spatial filters computed

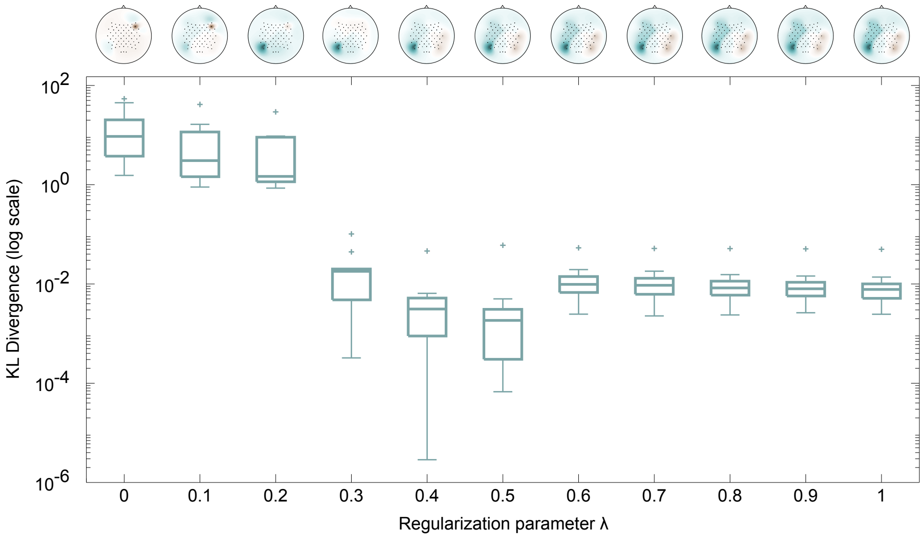

Figure 47.: Effects of applying divCSP-AS. The boxplot visualizes the similarity, measured as symmetric KL divergence, of the feature distribution of subject 74 and the other subjects when projecting the data to the first spatial filter computed with divCSP-AS. The activity patterns show that the influence of the artifact in electrode FFC6 decreases with increasing regularization.

with the subspace method (brown line) perform as well as the filters computed by CSP, even when using all 75 trials for the covariance estimation. The deflation divCSP-MS method shows a similar performance, although it has much higher variance and its 25% quantile is significantly lower than in the case of the subspace approach. Thus, for subject-independent spatial filters we strongly recommend using the subspace method and the beta divergence algorithm.

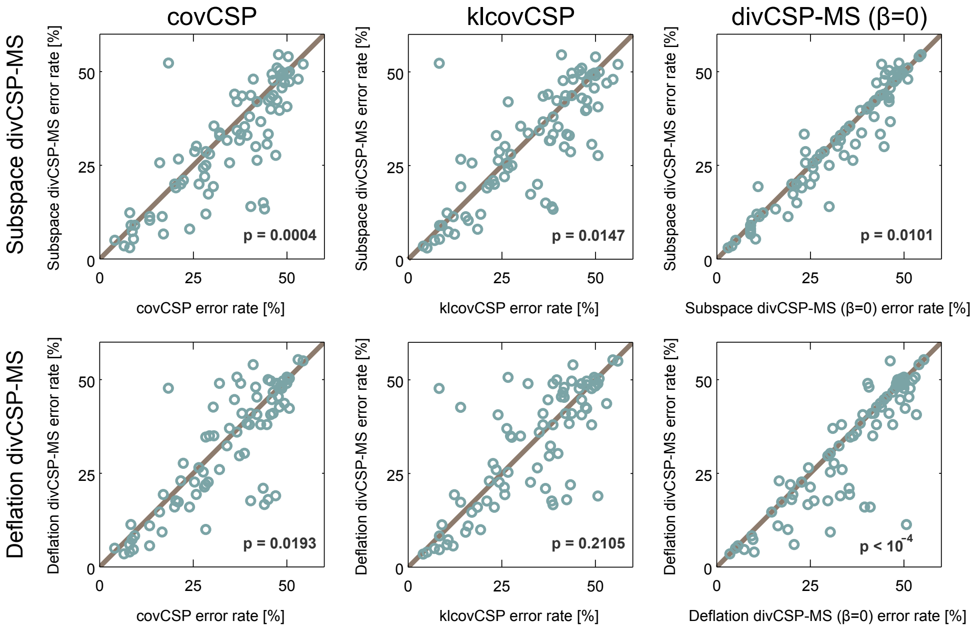

Figure 48.: Scatter plots showing error rates of deflation and subspace divCSP-MS and three baseline methods (for $\lambda = 1$). Each circle represents one subject and if the circle is below the solid line then our method outperforms the baseline for this subject. The p-value of the Wilcoxon signed rank test is displayed in the right bottom corner.

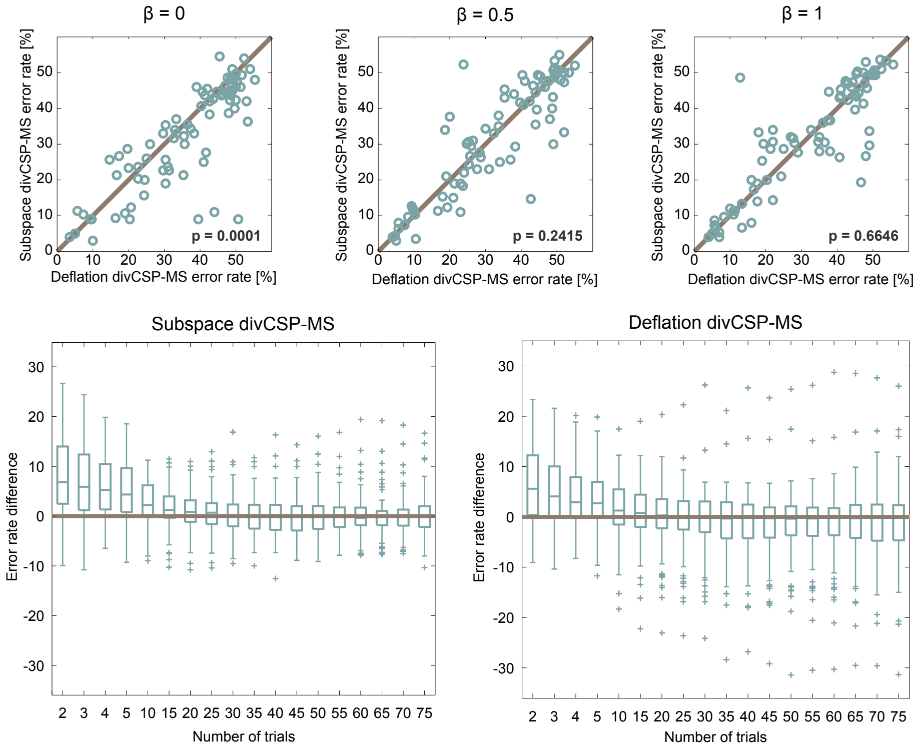

Figure 49.: *Top row*: Error rates of the deflation divCSP-MS (x-axis) and the subspace divCSP-MS (y-axis) for three beta values and $\lambda = 1$. For the case of $\beta = 0$ one sees that most of the circles representing the error rate of a particular subject are below the solid line, i.e., the subspace method perform better for these subjects. The relative advantage of the subspace method decreases constantly with increasing beta value. *Bottom row*: Distribution of error rate differences for both the subspace and deflation divCSP-MS approach with $\lambda = 1$ and CSP computed with various numbers of trials. Both divCSP-MS methods provide significantly better results than CSP when trained on less than 15 trials per class. Although divCSP-MS computes spatial filters by using other subjects' data only, its performance is on par with CSP which uses all 75 trials for covariance matrix estimation.

Lessons learned in this chapter

- Significant differences in performance between subspace and deflation algorithm.
- Only the subspace algorithm solves XOR-like problems.
- Across-subject regularization scheme provides excellent results and is superior to regularization of covariance matrices.
- Simultaneous optimization of two objectives more effective than two-step optimization.
- Beta divergence allows us to set the focus on particular types of nonstationarities.
- Beta divergence allows us to robustly incorporate data from other users by implicitly downweighting outlier subjects.
- Beta divergence robustifies the sumkl-divCSP method by downweighting outlier divergence terms.
- Divergences capture nonstationarity in a principled manner; no assumption of joint diagonalizability required.
- Flexibility of divergence-based framework allows us to tackle different robustness and nonstationarity problems in BCI.

Future Work

- Investigation of other divergences and regularization targets.
- Application of divergence-based classification algorithm; one framework for feature extraction and classification.
- Criterion for parameter selection; range of parameters.
- Theoretical analysis of relation between kl-divCSP and sumkl-divCSP.
- Theoretical analysis of robustness of β-divCSP; comparison to robust parameter estimation.
- Information geometric analysis and interpretation of the optimization; analysis of the geometry induced beta divergence.

8 SUMMARY AND OUTLOOK

We conclude this thesis with a brief summary of the presented algorithms and an outlook on future work. We also provide hints for the application of our methods in practice and comment on how to avoid common pitfalls.

8.1 WHAT METHOD TO USE IN PRACTICE ?

In this thesis we have proposed two novel spatial filter computation algorithms, sCSP and ssCSP, which extract stationary features by minimizing within-session changes and utilizing data from additional subjects in order to reduce between-session nonstationarity, respectively. Both algorithms compute spatial filters by solving a generalized eigenvalue problem, thus, are computationally efficient and neither require multiple initializations nor converge to a local optimum. If computation time is critical to an experiment, then methods such as sCSP or ssCSP should be preferred over algorithms which rely on gradient descent optimization or other nonlinear optimization techniques. In terms of efficiency sCSP has clear advantages over ssCSP as it solves a simpler problem. More precisely, sCSP estimates and reduces nonstationarity on the available training data whereas ssCSP is a transfer learning algorithm which aims to minimize between-session nonstationarity without "seeing" the required data. We think that ssCSP would benefit when signals of different users were normalized and coregistrated. Our experiments showed that sCSP can be widely applied in practice and positively affects performance in various settings. The assumption made by sCSP about the joint diagonalizability of the covariance matrices is not critical in practical application. A pitfall of the method, however, is the selection of the regularization parameter; normalization of the penalty matrix and the class covariance matrices helps to determine a meaningful parameter range.

The two other main contributions of this thesis are the derivation of a novel beta divergence-based covariance matrix estimator which is particularly tailored towards trial-structured data and the development of a divergence-based spatial filtering framework which allows one to compute robust and stationary features in a principled manner. The key advantage of the divergence-based CSP framework is its flexibility. The framework integrates many state-of-the-art spatial filter computation methods by incorporating data from additional users, robustifying the solution against artifacts or enforcing additional properties on the spatial filters by regularization. Furthermore, the framework provides an additional degree of freedom by utilizing the downweighting property of beta divergence. We discuss and extensively evaluate several variants of the divCSP algorithm in Chapter 7. The main disadvantages of the divergence-based methods are the demanding optimization and the presence of local optima in the objective function. Note that these disadvantages are not specific to our methods but are integral parts of nonconvex optimization problems. Therefore, we either have to rely on approximations as in the case of sCSP or accept the higher computational costs.

The robust covariance matrix estimator is the method of choice when outliers are present in the data. If data is scarce and of high dimension, then we recommend that the spatial filter computation algorithm applies the shrinkage estimator presented in Section 3.5. The beta divergence estimator based on the Wishart model should be the method of choice for robust parameter estimation in BCI because data have trial structure and usually whole trials can be regarded as outliers, e.g., when the subject fails to perform a given task. A pitfall in the implementation of β-WishartCSP is the

Table 8.: Main properties of methods presented in this thesis.

Method	Property
sCSP	Fast, approximates desired penalty function, reduces within-session nonstationarity, applicable to multiple time scales.
ssCSP	Fast, reduces between-session nonstationarity, requires additional data sets, assumes existence of common nonstationarities.
β-WishartCSP	Relatively fast, improves estimation in presence of outliers, solution depends on initialization, requires estimation of effective sample size.
β-divCSP	Slow, robust to artifacts, solution depends on initialization, robustifies an approximation of the CSP objective.
divCSP	Slow, flexible, principled measure of nonstationarity, solution depends on initialization, allows one to focus on specific types of nonstationarity, incorporates data from additional users, two optimization algorithms.

computation of the ratio of the two Γ-functions because each term may be very large and cause numerical problems. One avoids this numerical instability by an iterative computation of the ratios or by the application of a data type which is designed for handling very large numbers. A summary of the major properties of the proposed algorithms is displayed in Table 8.

In the following we comment on what spatial filter computation algorithm should be applied in practice. Unfortunately, we can not give a definitive answer to this question because there is no best algorithm for spatial filter computation; the performance of a method will always depend on the data. We have proposed several algorithms for tackling various challenges, e.g., robustness to artifacts or invariance to different types of nonstationarity, which occur in real-world BCI applications, thus, we have significantly enlarged the arsenal of methods which may be used in practice. Our spatial filtering framework developed in last chapter integrates these specific CSP methods and allows us to easily implement novel variants of the algorithm.

A question we can answer is "What method did perform best on the Vital BCI data set and on the BCI Competition data set ?". According to the average performance measure the four best methods for the Vital BCI data set are sCSP, klcovCSP, divCSP-AS and βν-WishartCSP and for the BCI Competition data set covCSP, klcovCSP, sCSP and βν-WishartCSP. Note that we did not compute the results of divCSP-AS for the latter data set. Thus, among the top approaches two methods incorporate data from additional subjects and two methods only use the training recordings of the subject of interest. This suggests that stationarity and robustness are of critical importance to reliable computation of spatial filters. From the above discussion our main practical recommendations can be summarized as

(1) Estimating nonstationary subspaces on other subjects' data is only advisable if these changes are very strong and stable, e.g., induced by a common effect.

(2) If data is contaminated by artifacts, then we recommend the application of a robust covariance matrix estimator or β-divCSP.

(3) The within-session nonstationarity should be minimized.

(4) If data from additional users is available, then we recommend that across-subject regularization or covariance matrix regularization is applied.

(5) Adaptation of the classifier or the spatial filters may provide further advantages.

8.2 FUTURE WORK

The selection of free parameters such as λ or β is usually performed by applying cross-validation. No mathematical criteria exist to derive the optimal parameters from data. Furthermore, the range of possible parameters is often arbitrary and not optimized for the given task. Parameter selection may also become a bottleneck in terms of computational efficiency if multiple parameters need to be determined, e.g., when combining multiple regularization strategies. Future work should provide tools and standard procedures to improve and simplify the selection of regularization parameters in practice. These parameter selection procedures may rely on heuristic criteria or be part of a Bayesian approach. We are convinced that by applying advanced parameter selection schemes and determining appropriate ranges of parameters, we may further improve the classification accuracy of most of the spatial filter computation methods presented in this thesis.

Several important topics not covered by this dissertation should be investigated in future research. This includes the optimization of spatio-temporal approaches with respect to stationarity or robustness, the integration of the robust feature extraction and robust classification steps into one common (divergence-based) framework and the theoretical analysis of the relation between robust spatial filtering and adaptation. A profound information geometric understanding of the properties of the spatial filter computation problem and a detailed analysis of the induced geometry are also interesting future topics because they allow one to utilize the properties of other divergences and construct CSP-like algorithms with specific properties. Extending the divergence-based CSP framework beyond Gaussian distributions, i.e., to other distribution classes (e.g. heavy-tailed distribution), will enable us to directly optimize for independence or stationarity of the extracted sources. Note that by using Student's t-distribution we may automatically robustify the divCSP algorithm against artifacts (in a similar manner as (Wu et al., 2009)). Furthermore, we would like to use our information geometric framework for classification purposes directly on the manifold of covariance matrices as done in (Barachant et al., 2012), in the context of kernel machines (Montavon et al., 2013) and apply our methods to multimodal data (Bießmann et al., 2011).

Spatial filter computation is not only relevant for motor imagery BCI, but also in other fields of neuroscience and beyond. For instance, myoelectric control (Hahne et al., 2012) is a promising field of application for the methods developed in this thesis. Since the EMG signal is also affected by artifacts and nonstationarity we conjecture that improvements can be achieved on these data sets. Another possible application of our methods beyond BCI is in the analysis of epilepsy data. Spatial filtering methods may be used to identify the origin of a seizure event and the techniques developed in this thesis may guide the spatial filter computation process towards desired solutions. Using beta divergence is especially useful for this task as this divergence allows one to shift the focus from gradual to abrupt changes. Finally, our methods are applicable to problems beyond neuroscience, e.g., to face recognition (Li and Savvides, 2007).

A

DERIVATION OF COVARIANCE ESTIMATOR

In the first part of the appendix we derive the iterative reweighting algorithm presented in Chapter 6 for the Gaussian and Wishart model. In order to obtain the estimate of the $\mathbf{\Sigma}$ parameter in the $(k+1)$th step we need to solve the following equation iteratively (see (Eguchi and Kano, 2001))

$$\frac{1}{n}\sum_{i=1}^{n} \psi_\beta(\ell(\mathbf{x}_i;\mathbf{\Sigma}))S(\mathbf{x}_i;\mathbf{\Sigma}) \;=\; \mathrm{E}\left[\psi_\beta(\ell(\mathbf{x};\mathbf{\Sigma}))S(\mathbf{x};\mathbf{\Sigma})\right], \qquad (76)$$

where $\psi_\beta(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}\Psi_\beta(\mathbf{x})$ and $S(\mathbf{x};\mathbf{\Sigma}) = \frac{\partial}{\partial \mathbf{\Sigma}}\ell(\mathbf{x};\mathbf{\Sigma})$. Note that $\mathrm{E}[\cdot]$ denotes the expectation over the whole input space.

A.1 ROBUST ESTIMATOR WITH GAUSSIAN MODEL

Assume we have the zero-mean multivariate D-dimensional Gaussian distribution

$$\begin{aligned}
p(\mathbf{x};\mathbf{\Sigma}) &= \underbrace{\frac{1}{(2\pi)^{\frac{D}{2}}}}_{\alpha}|\mathbf{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}}\\
\ell(p(\mathbf{x};\mathbf{\Sigma})) &= \log\alpha - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}\\
S(\mathbf{x};\mathbf{\Sigma}) &= -\frac{1}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\\
\psi_\beta(\ell(p(\mathbf{x};\mathbf{\Sigma}))) &= \alpha^\beta|\mathbf{\Sigma}|^{-\frac{\beta}{2}}e^{-\frac{\beta}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}}
\end{aligned}$$

If we put these definitions into the Eq. (76) we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{x}_i;\mathbf{\Sigma}))\left(-\frac{1}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\right)\\
&= \int\left(\alpha|\mathbf{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}}\right)\left(\alpha^\beta|\mathbf{\Sigma}|^{-\frac{\beta}{2}}e^{-\frac{\beta}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}}\right)\left(-\frac{1}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\right)d\mathbf{x},
\end{aligned}$$

After multiplication of both sides with $\sqrt{2}\mathbf{\Sigma}$ from the left and from the right we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{x}_i;\mathbf{\Sigma}))\left(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Sigma}\right) \;=\; \int\left(\alpha^{\beta+1}|\mathbf{\Sigma}|^{-\frac{\beta+1}{2}}e^{-\frac{\beta+1}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}}\right)\left(\mathbf{x}\mathbf{x}^\top - \mathbf{\Sigma}\right)d\mathbf{x},$$

Let $\tilde{\mathbf{\Sigma}} = \frac{1}{\beta+1}\mathbf{\Sigma}$. Then we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{x}_i;\mathbf{\Sigma}))\left(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Sigma}\right) \;=\; \frac{\alpha^\beta}{(\beta+1)^{\frac{D}{2}}}|\mathbf{\Sigma}|^{-\frac{\beta}{2}}\int\left(\alpha|\tilde{\mathbf{\Sigma}}|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{x}^\top\tilde{\mathbf{\Sigma}}^{-1}\mathbf{x}}\right)\left(\mathbf{x}\mathbf{x}^\top - \mathbf{\Sigma}\right)d\mathbf{x}$$

Note that if splitting the integral on the right hand side into two integrals then the first one gives the second moment of the multivariate Gaussian distribution and the second one is the zeroth moment times $\boldsymbol{\Sigma}$. Thus we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\beta}(\ell(\mathbf{x}_i;\boldsymbol{\Sigma}))\left(\mathbf{x}_i\mathbf{x}_i^{\top}-\boldsymbol{\Sigma}\right) = \frac{\alpha^{\beta}}{(\beta+1)^{\frac{D}{2}+1}}|\boldsymbol{\Sigma}|^{-\frac{\beta}{2}}(\tilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})$$

This is

$$\frac{1}{n}\sum_{i=1}^{n}\alpha^{\beta}|\boldsymbol{\Sigma}|^{-\frac{\beta}{2}}e^{-\frac{1}{2}\beta\mathbf{x}_i^{\top}(\boldsymbol{\Sigma})^{-1}\mathbf{x}_i}\left(\mathbf{x}_i\mathbf{x}_i^{\top}-\boldsymbol{\Sigma}\right) = \frac{-\alpha^{\beta}\beta}{(\beta+1)^{\frac{D}{2}+1}}|\boldsymbol{\Sigma}|^{-\frac{\beta}{2}}\boldsymbol{\Sigma}$$

This is

$$\frac{1}{n}\sum_{i=1}^{n}e^{-\frac{1}{2}\beta\mathbf{x}_i^{\top}(\boldsymbol{\Sigma})^{-1}\mathbf{x}_i}\boldsymbol{\Sigma} - \frac{\beta}{(\beta+1)^{\frac{D}{2}+1}}\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}e^{-\frac{1}{2}\beta\mathbf{x}_i^{\top}(\boldsymbol{\Sigma})^{-1}\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i$$

With this we obtain the iterative formula

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\frac{1}{n}\sum_{i=1}^{n}e^{-\frac{1}{2}\beta\mathbf{x}_i^{\top}(\boldsymbol{\Sigma}^{(k)})^{-1}\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i^{\top}}{\frac{1}{n}\sum_{i=1}^{n}e^{-\frac{1}{2}\beta\mathbf{x}_i^{\top}(\boldsymbol{\Sigma}^{(k)})^{-1}\mathbf{x}_i} - \beta/(\beta+1)^{\frac{D}{2}+1}}. \tag{77}$$

A.2 ROBUST ESTIMATOR WITH WISHART MODEL

Suppose that we have a set of scatter matrices $\{\mathbf{S}_1,\ldots,\mathbf{S}_n\}$ where

$$\mathbf{S}_i = \mathbf{X}_i\mathbf{X}_i^{\top} \tag{78}$$

and $\mathbf{X}_i \in \mathbb{R}^{D\times N}$ consists of the original D-dimensional observations of size ν in the ith trial. If the samples are i.i.d. with Gaussian distribution $\mathcal{N}(0,\boldsymbol{\Sigma})$, the scatter matrices are subject to Wishart distribution $\mathcal{W}(\boldsymbol{\Sigma},\nu)$. More generally, the density function of Wishart distribution is

$$q(\mathbf{S};\boldsymbol{\Sigma},\nu) = \frac{1}{2^{\frac{\nu D}{2}}|\boldsymbol{\Sigma}|^{\frac{\nu}{2}}\Gamma_D\left(\frac{\nu}{2}\right)}|\mathbf{S}|^{\frac{\nu-D-1}{2}}e^{-\mathrm{tr}\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)},$$

where Γ_D is the multivariate gamma function defined as

$$\Gamma_D\left(\frac{\nu}{2}\right) = \pi^{\frac{D(C-1)}{4}}\prod_{j=1}^{D}\Gamma\left[\frac{\nu}{2}+\frac{(1-j)}{2}\right]$$

$$\text{with} \quad \Gamma[t] = \int_0^{\infty}y^{t-1}e^{-y}dy.$$

We remark that the mean of Wishart random variable is $\nu\boldsymbol{\Sigma}$. The average covariance $\boldsymbol{\Sigma}$ can be determined by minimizing Eq. (76) iteratively or equivalently by minimizing the beta divergence between the empirical distribution of the observed scatter matrices and a model Wishart distribution. The following terms can be expressed explicitly

$$\begin{aligned}
\ell(\mathbf{S};\boldsymbol{\Sigma},\nu) &= \log \underbrace{\frac{1}{2^{\frac{\nu D}{2}}|\boldsymbol{\Sigma}|^{\frac{\nu}{2}}\Gamma_D\left(\frac{\nu}{2}\right)}}_{\alpha} + \frac{\nu-D-1}{2}\log|\mathbf{S}| - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\\
\psi_\beta(\ell(\mathbf{S};\boldsymbol{\Sigma},\nu)) &= \alpha^\beta \cdot |\mathbf{S}|^{\frac{\beta(\nu-D-1)}{2}} \cdot e^{-\frac{1}{2}\beta\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})}\\
S(\mathbf{S};\boldsymbol{\Sigma},\nu) &= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\nu\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

If we put these definitions into the Eq. (76) we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma},\nu))\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}_i\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\nu\boldsymbol{\Sigma}^{-1}\right)$$

$$= \int\left(\alpha|\mathbf{S}|^{\frac{(\nu-D-1)}{2}}e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})}\right)\left(\alpha^\beta|\mathbf{S}|^{\frac{\beta(\nu-D-1)}{2}}e^{-\frac{1}{2}\beta\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})}\right)\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\nu\boldsymbol{\Sigma}^{-1}\right)d\mathbf{S},$$

After multiplication of both sides with $\sqrt{2}\boldsymbol{\Sigma}$ from the left and from the right we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma},\nu))(\mathbf{S}_i - \nu\boldsymbol{\Sigma}) = \int\left(\alpha^{\beta+1}|\mathbf{S}|^{\frac{(\beta+1)(\nu-D-1)}{2}}e^{-\frac{1}{2}(\beta+1)\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})}\right)(\mathbf{S} - \nu\boldsymbol{\Sigma})\,d\mathbf{S},$$

Let $\tilde{\boldsymbol{\Sigma}} = \frac{1}{\beta+1}\boldsymbol{\Sigma}$, $\nu' = (\beta+1)\nu - \beta D - \beta$ and $\alpha' = \frac{1}{2^{\frac{\nu' D}{2}}|\tilde{\boldsymbol{\Sigma}}|^{\frac{\nu'}{2}}\Gamma_D\left(\frac{\nu'}{2}\right)}$. Then we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma},\nu))(\mathbf{S}_i - \nu\boldsymbol{\Sigma}) = \frac{\alpha^{\beta+1}}{\alpha'}\int\left(\alpha'|\mathbf{S}|^{\frac{\nu'-D-1}{2}}e^{-\frac{1}{2}\operatorname{tr}(\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{S})}\right)(\mathbf{S} - \nu\boldsymbol{\Sigma})\,d\mathbf{S},$$

Note that if splitting the integral on the right hand side into two integrals then the first one gives the first moment of the Wishart distribution and the second one is the zeroth moment times $\nu\boldsymbol{\Sigma}$. Thus, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma},\nu))(\mathbf{S}_i - \nu\boldsymbol{\Sigma}) = \frac{\alpha^{\beta+1}}{\alpha'}\left(\nu'\tilde{\boldsymbol{\Sigma}} - \nu\boldsymbol{\Sigma}\right)$$

This is

$$\frac{1}{n}\sum_{i=1}^{n}\psi_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma},\nu))(\mathbf{S}_i - \nu\boldsymbol{\Sigma}) = -\frac{\alpha^{\beta+1}(\beta D + \beta)}{\alpha'(\beta+1)}\boldsymbol{\Sigma}$$

$$\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(|\mathbf{S}_i|^{\frac{\beta(\nu-D-1)}{2}}e^{-\frac{1}{2}\beta\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_i)}\right)}_{\psi'_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma},\nu))}(\mathbf{S}_i - \nu\boldsymbol{\Sigma}) = -\frac{\alpha(\beta D + \beta)}{\alpha'(\beta+1)}\boldsymbol{\Sigma}$$

This leads to the iterative formula

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\psi'_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma}^{(k)},\nu))\mathbf{S}_i}{\frac{\nu}{n}\sum_{i=1}^{n}\psi'_\beta(\ell(\mathbf{S}_i;\boldsymbol{\Sigma}^{(k)},\nu)) - \frac{\alpha(\beta D+\beta)}{\alpha'(\beta+1)}}$$

Note that

$$\frac{\alpha(\beta D+\beta)}{\alpha'(\beta+1)} = \frac{2^{\frac{((\beta+1)\nu-\beta D-\beta)D}{2}}\left|\frac{1}{\beta+1}\boldsymbol{\Sigma}\right|^{\frac{(\beta+1)\nu-\beta D-\beta}{2}}\Gamma_D\left(\frac{\nu'}{2}\right)(\beta D+\beta)}{2^{\frac{\nu D}{2}}|\boldsymbol{\Sigma}|^{\frac{\nu}{2}}\Gamma_D\left(\frac{\nu}{2}\right)(\beta+1)}$$

$$= \frac{2^{\frac{(\beta\nu-\beta D-\beta)D}{2}}\frac{1}{\beta+1}^{\frac{((\beta+1)\nu-\beta D-\beta)D}{2}}|\boldsymbol{\Sigma}|^{\frac{\beta\nu-\beta D-\beta}{2}}\Gamma_D\left(\frac{\nu'}{2}\right)(\beta D+\beta)}{\Gamma_D\left(\frac{\nu}{2}\right)(\beta+1)}$$

$$= \frac{\beta(D+1)\Gamma_D\left(\frac{\nu(\beta+1)}{2}-\frac{(D+1)\beta}{2}\right)}{2^{\nu D}(\beta+1)\Gamma_D\left(\frac{\nu}{2}\right)}\left(\frac{2}{\beta+1}\right)^{\frac{\nu D(\beta+1)-D(D+1)\beta}{2}}|\boldsymbol{\Sigma}|^{\frac{\beta(\nu-D-1)}{2}}$$

This is the same formula as in Theorem (9) in Section 6.3.

B DIVERGENCE FRAMEWORK FOR CSP

This part of the appendix shows the derivation of the objective function and the gradient for the proposed divergence-based CSP framework introduced in Chapter 7. Note that all divCSP algorithms aim to find a projection that maximizes a sum of divergences between zero-mean Gaussian distributions. The first section shows the derivation when using Kullback-Leibler divergence, the second one when applying beta divergence. In the last section we give a full proof of Theorem 10 relating divergence maximization and Common Spatial Patterns. In the following we denote the projected covariance matrix with a bar symbol $\bar{\mathbf{\Sigma}} = (\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)$ and the whitened covariance matrix with a tilde symbol $\tilde{\mathbf{\Sigma}} = \mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top$.

B.1 DERIVATION USING KL DIVERGENCE

In Eq. (55) we have shown that the objective function and the gradient can be represented explicitly when using Kullback-Leibler divergence. From information theory (MacKay, 2002) it is well known that the KL divergence between two zero mean Gaussians $g_i \sim \mathcal{N}\left(0, \bar{\mathbf{\Sigma}}_i\right)$ and $f_j \sim \mathcal{N}\left(0, \bar{\mathbf{\Sigma}}_j\right)$ has the following explicit representation

$$D\left(\mathcal{N}\left(0, \bar{\mathbf{\Sigma}}_i\right) \;||\; \mathcal{N}\left(0, \bar{\mathbf{\Sigma}}_j\right)\right) \;=\; \frac{1}{2}\left(\log|\bar{\mathbf{\Sigma}}_j| \;-\; \log|\bar{\mathbf{\Sigma}}_i| \;+\; \mathrm{tr}\left[(\bar{\mathbf{\Sigma}}_j)^{-1}\bar{\mathbf{\Sigma}}_i\right] \;-\; d\right).$$

Note that the log terms cancel out when using the symmetric divergence, however, an additional trace term (with swapped $\bar{\mathbf{\Sigma}}_i$ and $\bar{\mathbf{\Sigma}}_j$) appears. The gradient of the divergence with respect to $\mathbf{R}$ can be computed separately for every term in the sum.

(1) Let us consider the log-determinant term. It can be rewritten as

$$\nabla_{\mathbf{R}} \log\left|(\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\right| \;=\; \mathbf{I}_d^\top\left[\nabla_{\mathbf{G}} \log\left|\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G}\right|\right]^\top$$

with $\mathbf{G} = \tilde{\mathbf{R}}^\top$ and $\tilde{\mathbf{R}}$ is the $d \times D$ matrix consisting of the first d rows of $\mathbf{R}$. According to (Petersen and Pedersen, 2012) this is

$$\mathbf{I}_d^\top\left[2\tilde{\mathbf{\Sigma}}_j\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G})^{-1}\right]^\top \quad \text{or} \quad 2\mathbf{I}_d^\top(\bar{\mathbf{\Sigma}}_j)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_j\mathbf{R}.$$

The derivative of the other log-determinant term can be computed in an analogous way and gives

$$2\mathbf{I}_d^\top(\bar{\mathbf{\Sigma}}_i)^{-1}\mathbf{I}_d\tilde{\mathbf{\Sigma}}_i\mathbf{R}.$$

(2) The derivative of the trace term can be computed as follows. Let us rewrite

$$\begin{aligned}&\nabla_{\mathbf{R}}\,\mathrm{tr}\left[((\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top))^{-1}((\mathbf{I}_d\mathbf{R}\mathbf{P})\mathbf{\Sigma}_i(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top))\right]\\ &= \mathbf{I}_d^\top\left[\nabla_{\mathbf{G}}\,\mathrm{tr}\left[\left(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G}\right)^{-1}\left(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_i\mathbf{G}\right)\right]\right]^\top,\end{aligned}$$

with $\mathbf{G}$ being defined as above. According to (Petersen and Pedersen, 2012) this is

$$\mathbf{I}_d^\top\left[-2\tilde{\mathbf{\Sigma}}_j\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G})^{-1}\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_i\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_j\mathbf{G})^{-1} \;+\; 2\tilde{\mathbf{\Sigma}}_i\mathbf{G}(\mathbf{G}^\top\tilde{\mathbf{\Sigma}}_i\mathbf{G})^{-1}\right]^\top.$$

Thus the derivative of the trace term is

$$-2\mathbf{I}_d^\top \left((\bar{\mathbf{\Sigma}}_j)^{-1} \bar{\mathbf{\Sigma}}_i (\bar{\mathbf{\Sigma}}_j)^{-1} \mathbf{I}_d \tilde{\mathbf{\Sigma}}_j \; - \; (\bar{\mathbf{\Sigma}}_i)^{-1} \mathbf{I}_d \tilde{\mathbf{\Sigma}}_i \right) \mathbf{R}.$$

Table 9 gives an overview over the gradients and objective functions of all divCSP variants proposed in this thesis.

Table 9.: Objective and gradients for the divCSP methods using KL divergence.

Method	Objective $\mathcal{L}(\mathbf{R})$ and gradient $\nabla_{\mathbf{R}}\mathcal{L}$
CSP term	$\mathcal{L}_{csp}(\mathbf{R}) = \frac{1}{2}\mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_i)^{-1}\bar{\boldsymbol{\Sigma}}_j\right] + \frac{1}{2}\mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_j)^{-1}\bar{\boldsymbol{\Sigma}}_i\right] - d$ $\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) = \mathbf{I}_d^\top\left((\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2 - (\bar{\boldsymbol{\Sigma}}_1)^{-1}\bar{\boldsymbol{\Sigma}}_2(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 + (\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 - (\bar{\boldsymbol{\Sigma}}_2)^{-1}\bar{\boldsymbol{\Sigma}}_1(\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2\right)\mathbf{R}$
divCSP-WS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{4n}\sum_{c=1}^{2}\sum_{i=1}^{n}\left(\log\left\lvert\bar{\boldsymbol{\Sigma}}_c\right\rvert - \log\left\lvert\bar{\boldsymbol{\Sigma}}_c^i\right\rvert + \mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_c)^{-1}\bar{\boldsymbol{\Sigma}}_c^i\right] - d\right)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{2n}\sum_{c=1}^{2}\sum_{i=1}^{n}\mathbf{I}_d^\top\left((\bar{\boldsymbol{\Sigma}}_c^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c^i - (\bar{\boldsymbol{\Sigma}}_c)^{-1}\bar{\boldsymbol{\Sigma}}_c^i(\bar{\boldsymbol{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c - (\bar{\boldsymbol{\Sigma}}_c^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c^i + (\bar{\boldsymbol{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c\right)\mathbf{R}$
divCSP-BS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{4K}\sum_{c=1}^{2}\sum_{k=1}^{K}\left(\mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^k\right] + \mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{te,c}^k\right] - 2d\right)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^\top\left((\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{te,c}^k + (\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k - (\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{te,c}^k(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k\right.$ $\left.\quad - (\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^k(\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{te,c}^k\right)\mathbf{R}$
divCSP-AS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{4K}\sum_{c=1}^{2}\sum_{k=1}^{K}\left(\mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell\right] + \mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^k\right] - 2d\right)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^\top\left((\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^\ell + (\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k - (\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k\right.$ $\left.\quad - (\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell)^{-1}\bar{\boldsymbol{\Sigma}}_{tr,c}^k(\bar{\boldsymbol{\Sigma}}_{tr,c}^\ell)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^\ell\right)\mathbf{R}$
divCSP-MS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{2K}\sum_{k=1}^{K}\left(\mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\bar{\boldsymbol{\Sigma}}_2^k\right] + \mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\bar{\boldsymbol{\Sigma}}_1^k\right] - 2d\right)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda}{K}\sum_{k=1}^{K}\mathbf{I}_d^\top\left((\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^k - (\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\bar{\boldsymbol{\Sigma}}_2^k(\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^k + (\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^k - (\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\bar{\boldsymbol{\Sigma}}_1^k(\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^k\right)\mathbf{R}$
sumkl-divCSP	$\mathcal{L}(\mathbf{R}) = \sum_{i=1}^{n}\left(\mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_1^i)^{-1}\bar{\boldsymbol{\Sigma}}_2^i\right] + \mathrm{tr}\left[(\bar{\boldsymbol{\Sigma}}_2^i)^{-1}\bar{\boldsymbol{\Sigma}}_1^i\right] - 2d\right)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = \sum_{i=1}^{n}\mathbf{I}_d^\top\left((\bar{\boldsymbol{\Sigma}}_2^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^i - (\bar{\boldsymbol{\Sigma}}_1^i)^{-1}\bar{\boldsymbol{\Sigma}}_2^i(\bar{\boldsymbol{\Sigma}}_1^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^i + (\bar{\boldsymbol{\Sigma}}_1^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^i - (\bar{\boldsymbol{\Sigma}}_2^i)^{-1}\bar{\boldsymbol{\Sigma}}_1^i(\bar{\boldsymbol{\Sigma}}_2^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^i\right)\mathbf{R}$

B.2 DERIVATION USING BETA DIVERGENCE

In Eq. (61) we have shown that the objective function and the gradient can be represented explicitly when using beta divergence. Beta divergence between two zero-mean Gaussians $g_i \sim \mathcal{N}(0, \bar{\boldsymbol{\Sigma}}_i)$ and $f_j \sim \mathcal{N}(0, \bar{\boldsymbol{\Sigma}}_j)$ is defined as

$$\begin{aligned} D_\beta\left(\mathcal{N}(0, \bar{\boldsymbol{\Sigma}}_i) \;\|\; \mathcal{N}(0, \bar{\boldsymbol{\Sigma}}_j)\right) &= \frac{1}{\beta(\beta+1)} \int g_i^{\beta+1}(x)dx - \frac{1}{\beta} \int f_j^\beta g_i(x)dx \\ &+ \frac{1}{\beta+1} \int f_j^{\beta+1}(x)dx. \end{aligned}$$

The integral $\int f_j^{\beta+1}(x)dx$ can be expressed as

$$\begin{aligned} \int f_j^{\beta+1}(x)dx &= \frac{1}{(2\pi)^{\frac{(\beta+1)d}{2}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{\beta+1}{2}}} \int e^{-\frac{1}{2}x^T(\beta+1)\bar{\boldsymbol{\Sigma}}_j^{-1}x} dx \\ &= \frac{1}{(2\pi)^{\frac{(\beta+1)d}{2}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{\beta+1}{2}}} \int e^{-\frac{1}{2}x^T(\frac{1}{\beta+1}\bar{\boldsymbol{\Sigma}}_j)^{-1}x} dx \\ &\overset{*}{=} \frac{1}{(2\pi)^{\frac{(\beta+1)d}{2}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{\beta+1}{2}}} (2\pi)^{\frac{d}{2}} \left(\frac{1}{\beta+1}\right)^{\frac{d}{2}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{1}{2}} \\ &= \frac{1}{(2\pi)^{\frac{\beta d}{2}} (\beta+1)^{\frac{d}{2}}} |\bar{\boldsymbol{\Sigma}}_j|^{-\frac{\beta}{2}} \end{aligned}$$

Note that step * assumes a Gaussian distribution under the integral, i.e. $\beta > -1$. The integral $\int g_i^{\beta+1}(x)dx$ can be computed in an analogous way.
The integral $\int f_j^\beta(x) g_i(x)dx$ is expressed in explicit form as

$$\begin{aligned} \int f_j^\beta(x) g_i(x)dx &= \frac{1}{(2\pi)^{\frac{\beta d}{2}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{\beta}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}} |\bar{\boldsymbol{\Sigma}}_i|^{\frac{1}{2}}} \int e^{-\frac{1}{2}x^T(\beta(\bar{\boldsymbol{\Sigma}}_j)^{-1} + (\bar{\boldsymbol{\Sigma}}_i)^{-1})x} dx \\ &\overset{*}{=} \frac{1}{(2\pi)^{\frac{\beta d}{2}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{\beta}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}} |\bar{\boldsymbol{\Sigma}}_i|^{\frac{1}{2}}} (2\pi)^{\frac{d}{2}} \left|\beta(\bar{\boldsymbol{\Sigma}}_j)^{-1} + (\bar{\boldsymbol{\Sigma}}_i)^{-1}\right|^{-\frac{1}{2}} \\ &= \frac{1}{(2\pi)^{\frac{\beta d}{2}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{\beta}{2}} |\bar{\boldsymbol{\Sigma}}_i|^{\frac{1}{2}}} \left|\beta(\bar{\boldsymbol{\Sigma}}_j)^{-1} + (\bar{\boldsymbol{\Sigma}}_i)^{-1}\right|^{-\frac{1}{2}} \\ &= \frac{1}{(2\pi)^{\frac{\beta d}{2}}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}} \left|\bar{\boldsymbol{\Sigma}}_j \left(\beta(\bar{\boldsymbol{\Sigma}}_j)^{-1} + (\bar{\boldsymbol{\Sigma}}_i)^{-1}\right) \bar{\boldsymbol{\Sigma}}_i\right|^{-\frac{1}{2}} \\ &= \frac{1}{(2\pi)^{\frac{\beta d}{2}}} |\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}} \left|\beta\bar{\boldsymbol{\Sigma}}_i + \bar{\boldsymbol{\Sigma}}_j\right|^{-\frac{1}{2}} \end{aligned}$$

Note that also here step * assumes that $\beta(\bar{\boldsymbol{\Sigma}}_j)^{-1} + (\bar{\boldsymbol{\Sigma}}_i)^{-1}$ is symmetric positive definite. This assumption is always true for $\beta \geqslant 0$, however, it is violated for $\beta < c$ with c being some negative constant. Therefore we apply very small negative β values for divCSP-WS, more precisely we select the smallest possible β from $-0.0005, -0.0010, -0.0015, \dots$

When using the symmetric beta divergence some terms cancel out and a simplified explicit representation can be derived. As before we separately compute the gradient of each term of the beta divergence objective function.

(1) The gradient of $|\bar{\boldsymbol{\Sigma}}_j|^{-\frac{\beta}{2}}$ with respect to $\mathbf{R}$ can be computed when rewriting

$$\nabla_{\mathbf{R}} \left|(\mathbf{I}_d\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)\right|^{-\frac{\beta}{2}} \;=\; \mathbf{I}_d^\top\left[\nabla_{\mathbf{G}}|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{\beta}{2}}\right]^\top$$

with $\mathbf{G} = \tilde{\mathbf{R}}^\mathsf{T}$ and $\tilde{\mathbf{R}}$ is the $d \times D$ matrix consisting of the first d rows of $\mathbf{R}$. According to matrix cookbook (Petersen and Pedersen, 2012) this is

$$-\beta\mathbf{I}_d^\top|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{\beta}{2}}\cdot\left(\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}(\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G})^{-1}\right)^\top \;=\; -\beta\mathbf{I}_d^\top|\bar{\boldsymbol{\Sigma}}_j|^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_j)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_j\mathbf{R}.$$

The gradient of $|\bar{\boldsymbol{\Sigma}}_i|^{-\frac{\beta}{2}}$ can be derived in an analogous way and gives

$$-\beta\mathbf{I}_d^\top|\bar{\boldsymbol{\Sigma}}_i|^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_i\mathbf{R}.$$

(2) Let us rewrite the gradient of $|\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}}|\beta\bar{\boldsymbol{\Sigma}}_i + \bar{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2}}$ as

$$\begin{aligned}&\nabla_{\mathbf{R}}\left[|(\mathbf{I}_d^\top\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)|^{\frac{1-\beta}{2}}\cdot|\beta(\mathbf{I}_d\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_i(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top) \;+\; (\mathbf{I}_d\mathbf{R}\mathbf{P})\boldsymbol{\Sigma}_j(\mathbf{P}^\top\mathbf{R}^\top\mathbf{I}_d^\top)|^{-\frac{1}{2}}\right]\\ &=\; \mathbf{I}_d^\top\left[\nabla_{\mathbf{G}}\left(|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{\frac{1-\beta}{2}}\cdot|\beta\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_i\mathbf{G} \;+\; \mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{1}{2}}\right)\right]^\mathsf{T}\end{aligned}$$

with $\mathbf{G}$ being defined as above. According to the product rule this is

$$\begin{aligned}&-\mathbf{I}_d^\top\left[(\beta-1)|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{\beta+1}{2}}\cdot|\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|\cdot\left(\mathbf{G}\tilde{\boldsymbol{\Sigma}}_j(\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G})^{-1}\right)\cdot|\beta\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_i\mathbf{G} \;+\; \mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{-\frac{1}{2}}\right.\\ &+\; |\mathbf{G}^\top\tilde{\boldsymbol{\Sigma}}_j\mathbf{G}|^{\frac{1-\beta}{2}}\cdot|\mathbf{G}^\top(\beta\tilde{\boldsymbol{\Sigma}}_i \;+\; \tilde{\boldsymbol{\Sigma}}_j)\mathbf{G}|^{-\frac{3}{2}}\cdot|\mathbf{G}^\top(\beta\tilde{\boldsymbol{\Sigma}}_i \;+\; \tilde{\boldsymbol{\Sigma}}_j)\mathbf{G}|\\ &\left.\cdot\left((\beta\tilde{\boldsymbol{\Sigma}}_i+\tilde{\boldsymbol{\Sigma}}_j)\mathbf{G}(\mathbf{G}^\top(\beta\tilde{\boldsymbol{\Sigma}}_i+\tilde{\boldsymbol{\Sigma}}_j)\mathbf{G})^{-1}\right)\right]^\top\end{aligned}$$

Writing it back gives

$$\begin{aligned}&-\mathbf{I}_d^\top\left[(\beta-1)|\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}}\cdot|\beta\bar{\boldsymbol{\Sigma}}_i \;+\; \bar{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2}}\cdot(\bar{\boldsymbol{\Sigma}}_j)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_j \;+\; |\bar{\boldsymbol{\Sigma}}_j|^{\frac{1-\beta}{2}}\cdot|\beta\bar{\boldsymbol{\Sigma}}_i \;+\; \bar{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2}}\right.\\ &\left.\cdot(\beta\bar{\boldsymbol{\Sigma}}_i \;+\; \bar{\boldsymbol{\Sigma}}_j)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_i \;+\; \tilde{\boldsymbol{\Sigma}}_j)\right]^\top\mathbf{R}\end{aligned}$$

Table 10 gives an overview over the gradients and objective function of all divCSP variants proposed in this thesis. The definition of the variables is same as before.

Table 10.: Objective and gradients for the divCSP methods using beta divergence.

	Objective $\mathcal{L}(\mathbf{R})$ and gradient $\nabla_{\mathbf{R}}\mathcal{L}$
CSP term	$\mathcal{L}(\mathbf{R}) = \gamma\Big(\lvert\bar{\boldsymbol{\Sigma}}_1\rvert^{-\frac{\beta}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_2\rvert^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}}\Big(\lvert\bar{\boldsymbol{\Sigma}}_2\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_1 + \bar{\boldsymbol{\Sigma}}_2\rvert^{-\frac{1}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_1\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_2 + \bar{\boldsymbol{\Sigma}}_1\rvert^{-\frac{1}{2}}\Big)\Big)$ $\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) = \gamma\mathbf{I}_d^{\top}\Big(\beta\lvert\bar{\boldsymbol{\Sigma}}_1\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 + \beta\lvert\bar{\boldsymbol{\Sigma}}_2\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2 - (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_2\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_1 + \bar{\boldsymbol{\Sigma}}_2\rvert^{-\frac{1}{2}}\cdot\big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2$ $+ (\beta\bar{\boldsymbol{\Sigma}}_1 + \bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_1 + \tilde{\boldsymbol{\Sigma}}_2)\big] - (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_1\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_2 + \bar{\boldsymbol{\Sigma}}_1\rvert^{-\frac{1}{2}}\cdot\big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 + (\beta\bar{\boldsymbol{\Sigma}}_2 + \bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_2 + \tilde{\boldsymbol{\Sigma}}_1)\big]\Big)\mathbf{R}$
divCSP-WS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda\gamma}{2n}\sum_{i=1}^{n}\Big((\beta+1)^{-1}\lvert\bar{\boldsymbol{\Sigma}}_c^i\rvert^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_c\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_c^i + \bar{\boldsymbol{\Sigma}}_c\rvert^{-\frac{1}{2}} + \beta(\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_c\rvert^{-\frac{\beta}{2}}\Big)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda\gamma}{2n}\sum_{c=1}^{2}\sum_{i=1}^{n}\mathbf{I}_d^{\top}\Big(\frac{\beta}{\beta+1}\lvert\bar{\boldsymbol{\Sigma}}_c^i\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_c^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c^i + \frac{\beta^2}{\beta+1}\lvert\bar{\boldsymbol{\Sigma}}_c\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c$ $- (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_c\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_c^i + \bar{\boldsymbol{\Sigma}}_c\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_c)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_c + (\beta\bar{\boldsymbol{\Sigma}}_c^i + \bar{\boldsymbol{\Sigma}}_c)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_c^i + \tilde{\boldsymbol{\Sigma}}_c)\Big]^{\top}\Big)\mathbf{R}$
divCSP-BS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda\gamma}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\Big(\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{\beta}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_{te,c}^k\rvert^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}}\Big(\lvert\bar{\boldsymbol{\Sigma}}_{te,c}^k\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^k + \bar{\boldsymbol{\Sigma}}_{te,c}^k\rvert^{-\frac{1}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_{te,c}^k + \bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{1}{2}}\Big)\Big)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda\gamma}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^{\top}\Big(\beta\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k + \beta\lvert\bar{\boldsymbol{\Sigma}}_{te,c}^k\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{te,c}^k$ $- (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_{te,c}^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^k + \bar{\boldsymbol{\Sigma}}_{te,c}^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_{te,c})^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{te,c} + (\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^k + \bar{\boldsymbol{\Sigma}}_{te,c}^k)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_{tr,c}^k + \tilde{\boldsymbol{\Sigma}}_{te,c}^k)\Big]$ $- (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_{te,c}^k + \bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_{tr,c})^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c} + (\beta\bar{\boldsymbol{\Sigma}}_{te,c}^k + \bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_{te,c}^k + \tilde{\boldsymbol{\Sigma}}_{tr,c}^k)\Big]\Big)\mathbf{R}$
divCSP-AS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda\gamma}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\Big(\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell}\rvert^{-\frac{\beta}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}}\Big(\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell} + \bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{1}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell}\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^k + \bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell}\rvert^{-\frac{1}{2}}\Big)\Big)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda\gamma}{2K}\sum_{c=1}^{2}\sum_{k=1}^{K}\mathbf{I}_d^{\top}\Big(\beta\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k + \beta\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell}\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell})^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^{\ell}$ $- (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell}\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^k + \bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell}\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell})^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^{\ell} + (\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^k + \bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell})^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_{tr,c}^k + \tilde{\boldsymbol{\Sigma}}_{tr,c}^{\ell})\Big]$ $- (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell} + \bar{\boldsymbol{\Sigma}}_{tr,c}^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_{tr,c}^k + (\beta\bar{\boldsymbol{\Sigma}}_{tr,c}^{\ell} + \bar{\boldsymbol{\Sigma}}_{tr,c}^k)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_{tr,c}^{\ell} + \tilde{\boldsymbol{\Sigma}}_{tr,c}^k)\Big]\Big)\mathbf{R}$
divCSP-MS	$\mathcal{L}(\mathbf{R}) = (1-\lambda)\mathcal{L}_{csp}(\mathbf{R}) - \frac{\lambda\gamma}{K}\sum_{k=1}^{K}\Big(\lvert\bar{\boldsymbol{\Sigma}}_1^k\rvert^{-\frac{\beta}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_2^k\rvert^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}}\Big(\lvert\bar{\boldsymbol{\Sigma}}_2^k\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_1^k + \bar{\boldsymbol{\Sigma}}_2^k\rvert^{-\frac{1}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_1^k\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_2^k + \bar{\boldsymbol{\Sigma}}_1^k\rvert^{-\frac{1}{2}}\Big)\Big)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = (1-\lambda)\nabla_{\mathbf{R}}\mathcal{L}_{csp}(\mathbf{R}) + \frac{\lambda\gamma}{K}\sum_{k=1}^{K}\mathbf{I}_d^{\top}\Big(\beta\lvert\bar{\boldsymbol{\Sigma}}_1^k\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_1^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^k + \beta\lvert\bar{\boldsymbol{\Sigma}}_2^k\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_2^k)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^k - (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_2^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_1^k + \bar{\boldsymbol{\Sigma}}_2^k\rvert^{-\frac{1}{2}}\cdot\big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_2)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2$ $+ (\beta\bar{\boldsymbol{\Sigma}}_1^k + \bar{\boldsymbol{\Sigma}}_2^k)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_1^k + \tilde{\boldsymbol{\Sigma}}_2^k)\Big] - (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_1^k\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_2^k + \bar{\boldsymbol{\Sigma}}_1^k\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_1)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1 + (\beta\bar{\boldsymbol{\Sigma}}_2^k + \bar{\boldsymbol{\Sigma}}_1^k)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_2^k + \tilde{\boldsymbol{\Sigma}}_1^k)\Big]\Big)\mathbf{R}$
βdivCSP	$\mathcal{L}(\mathbf{R}) = \gamma\sum_{i=1}^{n}\Big(\lvert\bar{\boldsymbol{\Sigma}}_1^i\rvert^{-\frac{\beta}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_2^i\rvert^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}}\Big(\lvert\bar{\boldsymbol{\Sigma}}_2^i\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_1^i + \bar{\boldsymbol{\Sigma}}_2^i\rvert^{-\frac{1}{2}} + \lvert\bar{\boldsymbol{\Sigma}}_1^i\rvert^{\frac{1-\beta}{2}}\lvert\beta\bar{\boldsymbol{\Sigma}}_2^i + \bar{\boldsymbol{\Sigma}}_1^i\rvert^{-\frac{1}{2}}\Big)\Big)$ $\nabla_{\mathbf{R}}\mathcal{L}(\mathbf{R}) = \gamma\sum_{i=1}^{n}\mathbf{I}_d^{\top}\Big(\beta\lvert\bar{\boldsymbol{\Sigma}}_1^i\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_1^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^i + \beta\lvert\bar{\boldsymbol{\Sigma}}_2^i\rvert^{-\frac{\beta}{2}}(\bar{\boldsymbol{\Sigma}}_2^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^i - (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_2^i\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_1^i + \bar{\boldsymbol{\Sigma}}_2^i\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_2^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_2^i$ $+ (\beta\bar{\boldsymbol{\Sigma}}_1^i + \bar{\boldsymbol{\Sigma}}_2^i)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_1^i + \tilde{\boldsymbol{\Sigma}}_2^i)\Big] - (\beta+1)^{\frac{d}{2}}\lvert\bar{\boldsymbol{\Sigma}}_1^i\rvert^{\frac{1-\beta}{2}}\cdot\lvert\beta\bar{\boldsymbol{\Sigma}}_2^i + \bar{\boldsymbol{\Sigma}}_1^i\rvert^{-\frac{1}{2}}\cdot\Big[(\beta-1)(\bar{\boldsymbol{\Sigma}}_1^i)^{-1}\mathbf{I}_d\tilde{\boldsymbol{\Sigma}}_1^i + (\beta\bar{\boldsymbol{\Sigma}}_2^i + \bar{\boldsymbol{\Sigma}}_1^i)^{-1}\mathbf{I}_d(\beta\tilde{\boldsymbol{\Sigma}}_2^i + \tilde{\boldsymbol{\Sigma}}_1^i)\Big]\Big)\mathbf{R}$

B.3 PROOF OF MAIN THEOREM

In this section we show the proof of Theorem 10 relating divergence maximization and Common Spatial Patterns. Note that (Wang, 2012) has provided a proof for the special case of one spatial filter.

Let $\tilde{\mathbf{R}} \in \mathbb{R}^{d \times D}$ denote the orthogonal projection onto a subspace of dimension d and let $\tilde{\boldsymbol{\Sigma}}_1$ and $\tilde{\boldsymbol{\Sigma}}_2$ represent the whitened covariance matrices with $\tilde{\boldsymbol{\Sigma}}_1 + \tilde{\boldsymbol{\Sigma}}_2 = \mathbf{I}$. Without loss of generality[1] we assume that $\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top = \boldsymbol{\Delta}_1$ and $\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top = \mathbf{I} - \boldsymbol{\Delta}_1$ with $\boldsymbol{\Delta}_1$ are diagonal matrices.

The KL divergence divCSP algorithm ($\lambda = 0$) optimizes the following objective function $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ (ignoring constant terms)

$$\begin{aligned}
&\tilde{D}\left((\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top)^{-1} \mid\mid (\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top)\right) \\
= \;& \operatorname{tr}\left((\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top)^{-1}(\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top)\right) + \operatorname{tr}\left((\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{R}}^\top)^{-1}(\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{R}}^\top)\right) \\
= \;& \operatorname{tr}\left(\boldsymbol{\Delta}_1^{-1}(\mathbf{I} - \boldsymbol{\Delta}_1)\right) + \operatorname{tr}\left((\mathbf{I} - \boldsymbol{\Delta}_1)^{-1}\boldsymbol{\Delta}_1\right) \\
= \;& \sum_{i=1}^{d} \frac{1-\nu_i}{\nu_i} + \sum_{i=1}^{d} \frac{\nu_i}{1-\nu_i},
\end{aligned}$$

where ν_i is the ith diagonal element of $\boldsymbol{\Delta}_1$.

Let us decompose $\tilde{\mathbf{R}} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ into two matrices $\mathbf{U} \in \mathbb{R}^{k \times D}$ and $\mathbf{V} \in \mathbb{R}^{d-k \times D}$ as follows

$$\begin{aligned}
\mathbf{U} &= \left\{\mathbf{r}_i : \frac{1-\nu_i}{\nu_i} > \frac{\nu_i}{1-\nu_i}\right\} \Longrightarrow \nu_i < 0.5 \\
\mathbf{V} &= \left\{\mathbf{r}_i : \frac{1-\nu_i}{\nu_i} \leqslant \frac{\nu_i}{1-\nu_i}\right\} \Longrightarrow \nu_i \geqslant 0.5.
\end{aligned}$$

Thus we can rewrite the objective function $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ as

$$\underbrace{\sum_{i=1}^{k} \frac{1-\nu_i}{\nu_i} + \frac{\nu_i}{1-\nu_i}}_{\mathbf{U}} + \underbrace{\sum_{i=k+1}^{d} \frac{1-\nu_i}{\nu_i} + \frac{\nu_i}{1-\nu_i}}_{\mathbf{V}}.$$

We prove that the top d CSP filters $\mathbf{W}$, i.e. the top d eigenvectors $\mathbf{v}_i$ ($i = 1 \dots d$) of $\tilde{\boldsymbol{\Sigma}}_1$ sorted by $\alpha_i = \max\{\mu_i, 1 - \mu_i\}$ where μ_i denotes the ith eigenvalue of $\tilde{\boldsymbol{\Sigma}}_1$, maximize $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$. Let us divide $\mathbf{W}$ into $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ as done above.

Case 1: Assume $\tilde{\mathbf{R}}$ maximizes $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ and it consists of eigenvectors $\mathbf{v}_i$ of $\tilde{\boldsymbol{\Sigma}}_1$, but there exist $\mathbf{v}_j \in \tilde{\mathbf{R}}$ with $j > d$ (i.e. it is not among the top (according to the above sorting) d eigenvectors). Thus $\mathbf{v}_j \notin \mathbf{W}$ and there exist $\mathbf{w}_l \in \mathbf{W}$ (which is among the top d eigenvectors) with $\mathbf{w}_l \notin \tilde{\mathbf{R}}$.

Without loss of generality assume $\mathbf{v}_j \in \mathbf{U}$. In the following we prove

$$\frac{1-\nu_j}{\nu_j} + \frac{\nu_j}{1-\nu_j} < \frac{1-\nu_l}{\nu_l} + \frac{\nu_l}{1-\nu_l},$$

[1] Because the basis in the projected subspace is arbitrary, i.e. the Kullback-Leibler divergence is invariant to right multiplication of any non-singular matrix $\mathbf{G} \in \mathbb{R}^{d \times d}$ with $\mathcal{L}_{kl}(\mathbf{V}) = \mathcal{L}_{kl}(\mathbf{VG})$.

where ν_l and ν_j denote the diagonal element when applying $\mathbf{w}_l$ and $\mathbf{v}_j$, respectively. Note that the function $f(\nu) = \frac{1-\nu}{\nu} + \frac{\nu}{1-\nu}$ is maximized at the borders (one can show this by taking the derivative).
Assume $\mathbf{w}_l \in \tilde{\mathbf{U}}$. Then $\nu_l < \nu_j < 0.5$ because $\mathbf{w}_l$ is selected before $\mathbf{v}_j$ (remember $\mathbf{v}_j \notin \mathbf{W}$) according to above sorting. Thus $f(\nu_j) < f(\nu_l)$ as $f(\nu)$ is maximized for the smallest argument ν (if $\nu < 0.5$).
Assume $\mathbf{w}_l \in \tilde{\mathbf{V}}$. Then $1-\nu_l < \nu_j < 0.5$ because $\mathbf{w}_l$ is selected before $\mathbf{v}_j$ according to above sorting. Thus $f(\nu_j) < f(1-\nu_l) = f(\nu_l)$.
Let us define $\mathbf{B}$ as $\tilde{\mathbf{R}}$, but with $\mathbf{w}_l$ instead of $\mathbf{v}_j$. Thus $\mathcal{L}_{kl}(\tilde{\mathbf{R}}) < \mathcal{L}_{kl}(\mathbf{B})$. This is a contradiction to the assumption that $\tilde{\mathbf{R}}$ is the optimal solution.

Case 2: Assume $\tilde{\mathbf{R}}$ maximizes $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$ and there exist (at least one) $\mathbf{r}_j \in \tilde{\mathbf{R}}$ with $\mathbf{r}_j$ is not an eigenvector of $\tilde{\mathbf{\Sigma}}_1$. Without loss of generality assume $\mathbf{r}_j \in \mathbf{U}$. Let us define a new solution $\mathbf{B} = \begin{bmatrix} \tilde{\mathbf{U}} \\ \tilde{\mathbf{V}} \end{bmatrix}$ as follows:

$\tilde{\mathbf{U}}$ consists of k eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ with smallest eigenvalues.
$\tilde{\mathbf{V}}$ consists of $d-k$ eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ with largest eigenvalues.

Let us denote the diagonal elements (eigenvalues) of $\mathbf{U}\tilde{\mathbf{\Sigma}}_1\mathbf{U}^\top$ as $\nu_1 < \ldots < \nu_k < 0.5$ and those obtained with $\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}_1\tilde{\mathbf{U}}^\top$ as $u_1 < \ldots < u_k < 0.5$. Note that $u_i = \mu_i$ where $\mu_1 < \ldots < \mu_D$ are the eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ (because $\tilde{\mathbf{U}}$ consists of the smallest eigenvectors of $\tilde{\mathbf{\Sigma}}_1$). Cauchy's interlacing theorem (Bhatia, 1997) establishes the following relation between ν_i and u_i, namely $u_i \leqslant \nu_i$. Note that equality only holds if $\mathbf{U}$ and $\tilde{\mathbf{U}}$ are the same, i.e. if $\mathbf{U}$ consists of the eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ (irrespectively of permutation). Cauchy's theorem implies that there are no ν_i and ν_j with $u_k < \nu_i < \nu_j < u_{k+1}$. Together with the fact that $f(\nu) = \frac{1-\nu}{\nu} + \frac{\nu}{1-\nu}$ is maximized at the borders (i.e. for smallest ν in this case) this for all i implies

$$\frac{1-\nu_i}{\nu_i} + \frac{\nu_i}{1-\nu_i} \leqslant \frac{1-u_i}{u_i} + \frac{u_i}{1-u_i},$$

Since $\exists i$ where this relation is strictly positive (because we assumed $\mathbf{r}_j \in \mathbf{U}$), we obtain $\mathcal{L}_{kl}(\mathbf{U}) < \mathcal{L}_{kl}(\tilde{\mathbf{U}})$.
Let us denote the diagonal elements (eigenvalues) of $\mathbf{V}\tilde{\mathbf{\Sigma}}_1\mathbf{V}^\top$ as $\nu_1 > \ldots > \nu_{d-k} \geqslant 0.5$ and those obtained with $\tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}_1\tilde{\mathbf{V}}^\top$ as $u_1 > \ldots > u_{d-k} \geqslant 0.5$. Note that $u_i = \mu_i$ where $\mu_1 > \ldots > \mu_D$ are the eigenvectors of $\tilde{\mathbf{\Sigma}}_1$ (because $\tilde{\mathbf{V}}$ consists of the largest eigenvectors of $\tilde{\mathbf{\Sigma}}_1$). Cauchy's interlacing theorem establishes the following relation between the ν_i and u_i, namely $\nu_i \leqslant u_i$. Note that equality only holds if $\mathbf{V}$ and $\tilde{\mathbf{V}}$ are the same (irrespectively of permutation). Together with the fact that $f(\nu) = \frac{1-\nu}{\nu} + \frac{\nu}{1-\nu}$ is maximized at the borders (i.e. for largest ν in this case) this implies

$$\frac{1-\nu_i}{\nu_i} + \frac{\nu_i}{1-\nu_i} \leqslant \frac{1-u_i}{u_i} + \frac{u_i}{1-u_i},$$

Thus $\mathcal{L}_{kl}(\mathbf{V}) \leqslant \mathcal{L}_{kl}(\tilde{\mathbf{V}})$ and consequently

$$\mathcal{L}_{kl}(\tilde{\mathbf{R}}) = \mathcal{L}_{kl}(\tilde{\mathbf{U}}) + \mathcal{L}_{kl}(\tilde{\mathbf{V}}) < \mathcal{L}_{kl}(\tilde{\mathbf{U}}) + \mathcal{L}_{kl}(\tilde{\mathbf{V}}) = \mathcal{L}_{kl}(\tilde{\mathbf{B}}).$$

This contradicts the assumption that $\tilde{\mathbf{R}}$ maximizes $\mathcal{L}_{kl}(\tilde{\mathbf{R}})$.

BIBLIOGRAPHY

Ahn, M., Cho, H., Ahn, S., and Jun, S. C. High theta and low alpha powers may be indicative of BCI-illiteracy in motor imagery. *PLoS ONE*, 8(11):e80886, 2013.

Alamgir, M., Grosse-Wentrup, M., and Altun, Y. Multitask learning for brain-computer interfaces. In *JMLR Workshop and Conference Proceedings Volume 9: AISTATS 2010*, pages 17–24, 2010.

Amari, S. Information geometry in optimization, machine learning and statistical inference. *Frontiers of Electrical and Electronic Engineering in China*, 5(3):241–260, 2010.

Amari, S. and Cichocki, A. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

Amari, S. and Nagaoka, H. Methods of information geometry. volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, 2000.

Ang, K. K. and Guan, C. Brain-computer interface in stroke rehabilitation. *Journal of Computing Science and Engineering*, 7(2):139–146, 2013.

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 2390–2397, 2008.

Ang, K. K., Guan, C., Chua, K. S. G., Ang, B. T., Kuah, C. W. K., Wang, C., Phua, K. S., Chin, Z. Y., and Zhang, H. A large clinical study on the ability of stroke patients to use an EEG-based motor imagery brain-computer interface. *Clinical EEG and Neuroscience*, 42(4):253–258, 2011.

Arvaneh, M., Guan, C., Ang, K. K., and Quek, H.-C. Spatially sparsed common spatial pattern to improve BCI performance. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2412–2415, 2011.

Arvaneh, M., Guan, C., Ang, K. K., and Quek, C. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(4): 610–619, 2013a.

Arvaneh, M., Guan, C., Ang, K. K., and Quek, C. EEG data space adaptation to reduce intersession nonstationarity in brain-computer interface. *Neural Computation*, 25(8): 2146–2171, 2013b.

Baik, J. and Silverstein, J. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

Baldeweg, T., Richardson, A., Watkins, S., Foale, C., and Gruzelier, J. Impaired auditory frequency discrimination in dyslexia detected with mismatch evoked potentials. *Annals of Neurology*, 45(4):495–503, 1999.

Bamdadian, A., Guan, C., Ang, K. K., and Xu, J. Online semi-supervised learning with kl distance weighting for motor imagery-based BCI. In *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 2732–2735, 2012.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. Common spatial pattern revisited by riemannian geometry. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 472–476, 2010.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. Multiclass brain-computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

Barry, R. J., Johnstone, S. J., and Clarke, A. R. A review of electrophysiology in attention-deficit/hyperactivity disorder: II. Event-related potentials. *Clinical Neurophysiology*, 114(2):184 – 198, 2003.

Bartz, D., Hatrick, K., Hesse, C. W., Müller, K.-R., and Lemm, S. Directional variance adjustment: Bias reduction in covariance matrices based on factor analysis with an application to portfolio optimization. *PLoS ONE*, 8:67503, 2013.

Başar, E., Başar-Eroglu, C., Karakaş, S., and Schürmann, M. Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International Journal of Psychophysiology*, 39(2):241–248, 2001.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

Berger, H. Über das elektrenkephalogramm des menschen II. *Journal für Psychologie und Neurologie*, 40:160–179, 1930.

Bhatia, R. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, 1997.

Bießmann, F., Plis, S. M., Meinecke, F. C., Eichele, T., and Müller, K.-R. Analysis of multimodal neuroimaging data. *IEEE Rev. Biomed. Eng.*, 4:26 – 58, 2011.

Biessmann, F., Dähne, S., Meinecke, F. C., Blankertz, B., Görgen, K., Müller, K.-R., and Haufe, S. On the interpretability of linear multivariate neuroimaging analyses: Filters, patterns and their relationship. In *2nd NIPS Workshop on Machine Learning and Inference in Neuroimaging*, 2012.

Binnie, C. EEG in clinical practice. *Journal of the Royal Society of Medicine*, 88(9):518, 1995.

Birbaumer, N. and Cohen, L. G. Brain–computer interfaces: communication and restoration of movement in paralysis. *The Journal of Physiology*, 579(3):621–636, 2007.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., and Flor, H. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Blankertz, B., Dornhege, G., Lemm, S., Krauledat, M., Curio, G., and Müller, K.-R. The berlin brain-computer interface: Machine learning based detection of user specific brain states. *Journal of Universal Computer Science*, 12(6):581–607, 2006a.

Blankertz, B., Müller, K.-R., Krusienski, D., Schalk, G., Wolpaw, J., Schlögl, A., Pfurtscheller, G., del R. Millán, J., Schröder, M., and Birbaumer, N. The BCI competition III: validating alternative approaches to actual BCI problems. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):153–159, 2006b.

Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F. U., Nikulin, V., and Müller, K.-R. Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 113–120, 2008a.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K.-R. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):41–56, 2008b.

Blankertz, B., Sannelli, C., Halder, S., Hammer, E. M., Kübler, A., Müller, K.-R., Curio, G., and Dickhaus, T. Neurophysiological predictor of SMR-based BCI performance. *NeuroImage*, 51(4):1303–1309, 2010a.

Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L., Sturm, I., Curio, G., et al. The berlin brain–computer interface: non-medical uses of BCI technology. *Frontiers in Neuroscience*, 4, 2010b.

Blankertz, B., Lemm, S., Treder, M. S., Haufe, S., and Müller, K.-R. Single-trial analysis and classification of ERP components – a tutorial. *NeuroImage*, 56(2):814–825, 2011.

Blum, A. S. and Rutkove, S. B., editors. *The Clinical Neurophysiology Primer*. Humana Press, Totowa, NJ, 2007.

Bonnet, L., Lotte, F., and Lécuyer, A. Two brains, one game: Design and evaluation of a multi-user BCI video game based on motor imagery. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(2):185–198, 2013.

Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *{USSR} Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967.

Carlson, T. and del R. Millán, J. Brain-controlled wheelchairs: A robotic architecture. *IEEE Robotics Automation Magazine*, 20(1):65–73, 2013.

Cichocki, A. and Amari, S. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Collura, T. F. History and evolution of electroencephalographic instruments and techniques. *Journal of Clinical Neurophysiology*, 10(4):476–504, 1993.

Courtine, G., Micera, S., DiGiovanna, J., and del R. Millán, J. Brain-machine interface: closer to therapeutic reality? *The Lancet*, 381(9866):515–517, 2013.

Daly, J. J. and Wolpaw, J. R. Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, 7(11):1032–1043, 2008.

del R. Millán, J., Ferrez, P. W., and Seidl, T. Chapter 14 validation of brain-machine interfaces during parabolic flight. volume 86 of *International Review of Neurobiology*, pages 189–197. Academic Press, 2009.

del R. Millán, J., Rupp, R., Müller-Putz, G. R., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kübler, A., Leeb, R., et al. Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in Neuroscience*, 41:161, 2010.

Devlaminck, D., Wyns, B., Grosse-Wentrup, M., Otte, G., and Santens, P. Multi-subject learning for common spatial patterns in motor-imagery BCI. *Computational Intelligence and Neuroscience*, 2011(217987):1–9, 2011.

Dietrich, D., Lang, R., Bruckner, D., Fodor, G., and Muller, B. Limitations, possibilities and implications of brain-computer interfaces. In *3rd International Conference on Human System Interactions (HSI)*, pages 722–726, 2010.

Dornhege, G., Blankertz, B., Curio, G., and Müller, K.-R. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Transactions on Biomedical Engineering*, 51(6):993 –1002, 2004.

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

Dunne, S., Leeb, R., Nijholt, A., and del R. Millán, J. *Towards Practical Brain-Computer Interfaces*. ser. Biological and Medical Physics, Biomedical Engineering. Springer, 2013.

Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, pages 793–803, 1983.

Eguchi, S. and Kano, Y. Robustifying maximum likelihood estimation. *Tokyo Institute of Statistical Mathematics, Tokyo, Japan, Tech. Rep*, 2001.

Falzon, O., Camilleri, K. P., and Muscat, J. The analytic common spatial patterns method for EEG-based BCI data. *Journal of Neural Engineering*, 9(4):045009, 2012.

Farwell, L. A. and Donchin, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

Fattahi, D., Nasihatkon, B., and Boostani, R. A general framework to estimate spatial and spatio-spectral filters for EEG signal classification. *Neurocomputing*, 119(0):165 – 174, 2013.

Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.-R., and Grozea, C. Subject-independent mental state classification in single trials. *Neural Networks*, 22(9):1305–1312, 2009.

Fazli, S., Danóczy, M., Schelldorfer, J., and Müller, K.-R. L1-penalized Linear Mixed-Effects Models for high dimensional data with application to BCI. *NeuroImage*, 56 (4):2100–2108, 2011.

Févotte, C. and Idier, J. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

Fischer, C., Morlet, D., Bouchet, P., Luaute, J., Jourdan, C., and Salord, F. Mismatch negativity and late auditory evoked potentials in comatose patients. *Clinical Neurophysiology*, 110(9):1601–1610, 1999.

Ford, J. M. Schizophrenia: The broken p300 and beyond. *Psychophysiology*, 36(6):667–682, 1999.

Galán, F., Nuttin, M., Lew, E., Ferrez, P. W., Vanacker, G., Philips, J., and del R. Millán, J. A brain-actuated wheelchair: asynchronous and non-invasive brain–computer interfaces for continuous control of robots. *Clinical Neurophysiology*, 119(9):2159–2169, 2008.

Goksu, F., Ince, N., and Tewfik, A. Sparse common spatial patterns in brain computer interface applications. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 533–536, 2011.

Gouy-Pailler, C., Congedo, M., Brunner, C., Jutten, C., and Pfurtscheller, G. Nonstationary brain source separation for multiclass motor imagery. *IEEE Transactions on Biomedical Engineering*, 57(2):469–478, 2010.

Grosse-Wentrup, M. and Buss, M. Multiclass common spatial patterns and information theoretic feature extraction. *IEEE Transactions on Biomedical Engineering*, 55(8):1991–2000, 2008.

Grosse-Wentrup, M., Gramann, K., and Buss, M. Adaptive spatial filters with predefined region of interest for EEG based brain-computer-interface. In *Advances in Neural Information Processing Systems 19 (NIPS)*, pages 537–544, 2007.

Grosse-Wentrup, M., Liefhold, C., Gramann, K., and Buss, M. Beamforming in noninvasive brain computer interfaces. *IEEE Transactions on Biomedical Engineering*, 56(4): 1209 –1219, 2009.

Grosse-Wentrup, M., Schölkopf, B., and Hill, J. Causal influence of gamma oscillations on the sensorimotor rhythm. *NeuroImage*, 56(2):837–842, 2011.

Guger, C., Harkam, W., Hertnaes, C., and Pfurtscheller, G. Prosthetic control by an EEG-based brain-computer interface (BCI). In *Proc. of 5th European Conference for the Advancement of Assistive Technology*, pages 3–6, 1999.

Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Hahne, J. M., Graimann, B., and Müller, K.-R. Spatial filtering for robust myoelectric control. *IEEE Transactions on Biomedical Engineering*, 59(5):1436–1443, 2012.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Haufe, S., Treder, M. S., Gugler, M. F., Sagebaum, M., Curio, G., and Blankertz, B. EEG potentials predict upcoming emergency brakings during simulated driving. *Journal of Neural Engineering*, 8(5):056001, 2011.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Biessmann, F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87(0):96 – 110, 2014.

Herculano-Houzel, S. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3, 2009.

Huber, P. J. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1981.

Hyvärinen, A. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999a.

Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999b.

Jackson, A., Moritz, C. T., Mavoori, J., Lucas, T. H., and Fetz, E. E. The neurochip BCI: towards a neural prosthesis for upper limb function. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):187–190, 2006.

Jasper, H. and Penfield, W. Electrocorticograms in man: effect of voluntary movement upon the electrical activity of the precentral gyrus. *Archiv für Psychiatrie und Nervenkrankheiten*, 183(1-2):163–174, 1949.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., and Sejnowski, T. J. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.

Kaiser, V., Kreilinger, A., Müller-Putz, G. R., and Neuper, C. First steps toward a motor imagery based stroke BCI: new strategy to set up a classifier. *Frontiers in Neuroscience*, 5, 2011.

Kamiya, J., Callaway, E., and Yeager, C. Visual evoked responses in subjects trained to control alpha rhythms. *Psychophysiology*, 5(6):683–695, 1969.

Kang, H. and Choi, S. Bayesian multi-task learning for common spatial patterns. In *International Workshop on Pattern Recognition in NeuroImaging (PRNI)*, pages 61–64, 2011.

Kang, H., Nam, Y., and Choi, S. Composite common spatial pattern for subject-to-subject transfer. *Signal Processing Letters*, 16(8):683 –686, 2009.

Kaplan, A. Y., Fingelkurts, A. A., Fingelkurts, A. A., Borisov, S. V., and Darkhovsky, B. S. Nonstationary nature of the brain activity as revealed by EEG/MEG: Methodological, practical and conceptual challenges. *Signal Processing*, 85(11):2190 – 2212, 2005.

Kawanabe, M. and Vidaurre, C. Improving BCI performance by modified common spatial patterns with robustly averaged covariance matrices. In *Proc. of IWANN 09, Part I*, LNCS, pages 279–282. Springer, 2009.

Kawanabe, M., Vidaurre, C., Blankertz, B., and Müller, K.-R. A maxmin approach to optimize spatial filters for EEG single-trial classification. In *Bio-Inspired Systems: Computational and Ambient Intelligence*, volume 5517 of *LNCS*, pages 674–682. Springer, 2009.

Kawanabe, M., Samek, W., von Bünau, P., and Meinecke, F. An information geometrical view of stationary subspace analysis. In *Artificial Neural Networks and Machine Learning - ICANN 2011*, volume 6792 of *LNCS*, pages 397–404. Springer, 2011.

Kawanabe, M., Samek, W., Müller, K.-R., and Vidaurre, C. Robust common spatial filters with a maxmin approach. *Neural Computation*, 26(2):1–28, 2014.

Király, F. J., Bünau, P. V., Meinecke, F. C., Blythe, D. A. J., and Müller, K.-R. Algebraic geometric comparison of probability distributions. *Journal of Machine Learning Research*, 13(1):855–903, 2012.

Klem, G. H., Lüders, H., Jasper, H., and Elger, C. The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology. *Electroencephalography and Clinical Neurophysiology. Supplement*, 52:3, 1999.

Koles, Z. J., Lazar, M. S., and Zhou, S. Z. Spatial patterns underlying population differences in the background EEG. *Brain Topography*, 2:275–284, 1990.

Krauledat, M. *Analysis of Nonstationarities in EEG signals for improving Brain-Computer Interface performance*. PhD thesis, Technische Universität Berlin, 2008.

Krauledat, M., Tangermann, M., Blankertz, B., and Müller, K.-R. Towards zero training for brain-computer interfacing. *PLoS ONE*, 3(8):e2967, 2008.

Krusienski, D. J., Grosse-Wentrup, M., Galán, F., Coyle, D., Miller, K. J., Forney, E., and Anderson, C. W. Critical issues in state-of-the-art brain–computer interface signal processing. *Journal of Neural Engineering*, 8(2):025002, 2011.

Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J., and Birbaumer, N. Brain-computer communication: unlocking the locked in. *Psychological Bulletin*, 127(3):358–375, 2001.

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

Lalor, E. C., Kelly, S. P., Finucane, C., Burke, R., Smith, R., Reilly, R. B., and Mcdarby, G. Steady-state vep-based brain-computer interface control in an immersive 3d gaming environment. *EURASIP Journal on Applied Signal Processing*, 2005:3156–3164, 2005.

Lance, B. J., Kerick, S. E., Ries, A. J., Oie, K. S., and McDowell, K. Brain–computer interface technologies in the coming decades. *Proceedings of the IEEE*, 100(13):1585–1599, 2012.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Leeb, R., Friedman, D., Müller-Putz, G. R., Scherer, R., Slater, M., and Pfurtscheller, G. Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: a case study with a tetraplegic. *Computational Intelligence and Neuroscience*, vol. 2007, 2007.

Lemm, S., Blankertz, B., Curio, G., and Müller, K.-R. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Transactions on Biomedical Engineering*, 52:1541–1548, 2005.

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387–399, 2011.

Li, X., Zhang, H., Guan, C., Ong, S. H., Ang, K. K., and Pan, Y. Discriminative learning of propagation and spatial pattern for motor imagery EEG analysis. *Neural Computation*, 25(10):2709–2733, 2013.

Li, Y. and Guan, C. An extended em algorithm for joint feature extraction and classification in brain-computer interfaces. *Neural Computation*, 18(11):2730–2761, 2006.

Li, Y.-h. and Savvides, M. Kernel fukunaga-koontz transform subspaces for enhanced face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

Lim, C. G., Lee, T. S., Guan, C., Fung, D. S. S., Zhao, Y., Teng, S. S. W., Zhang, H., and Krishnan, K. R. R. A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder. *PLoS ONE*, 7(10):e46692, 2012.

Liyanage, S. R., Guan, C., Zhang, H., Ang, K. K., Xu, J., and Lee, T. H. Dynamically weighted ensemble classification for non-stationary EEG processing. *Journal of Neural Engineering*, 10(3):036007, 2013.

Lotte, F. Brain-computer interfaces for 3d games: hype or hope? In *Proceedings of the 6th International Conference on Foundations of Digital Games*, pages 325–327, 2011.

Lotte, F. and Guan, C. Spatially regularized common spatial patterns for EEG classification. In *International Conference on Pattern Recognition (ICPR)*, pages 3712–3715, 2010a.

Lotte, F. and Guan, C. Learning from other subjects helps reducing brain-computer interface calibration time. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 614–617, 2010b.

Lotte, F. and Guan, C. Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*, 58 (2):355 –362, 2011.

Lu, H., Plataniotis, K., and Venetsanopoulos, A. Regularized common spatial patterns with generic learning for EEG signal classification. In *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 6599–6602, 2009.

Lu, H., Eng, H.-L., Guan, C., Plataniotis, K., and Venetsanopoulos, A. Regularized common spatial pattern with aggregation for EEG classification in small-sample setting. *IEEE Transactions on Biomedical Engineering*, 57(12):2936–2946, 2010.

Lu, S., Guan, C., and Zhang, H. Unsupervised brain computer interface based on inter-subject information. In *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 638–641, 2008.

MacKay, D. J. C. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

McFarland, D. J. and Wolpaw, J. R. Brain-computer interface operation of robotic and prosthetic devices. *Computer*, 41(10):52–56, 2008.

Mihoko, M. and Eguchi, S. Robust blind source separation by beta divergence. *Neural Computation*, 14(8):1859–1886, 2002.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A., and Müller, K.-R. Invariant feature extraction and classification in kernel spaces. In *Advances in Neural Information Processing System 12 (NIPS)*, pages 526–532, 2000.

Mollah, M. N. H., Sultana, N., Minami, M., and Eguchi, S. Robust extraction of local structures by the minimum beta-divergence method. *Neural Networks*, 23(2):226–238, 2010.

Montavon, G., Braun, M., Krüger, T., and Müller, K.-R. Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment. *Signal Processing Magazine, IEEE*, 30(4):62–74, 2013.

Mousavi, E. A., Maller, J. J., Fitzgerald, P. B., and Lithgow, B. J. Wavelet common spatial pattern in asynchronous offline brain computer interfaces. *Biomedical Signal Processing and Control*, 6(2):121 – 128, 2011.

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

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. Machine learning for real-time single-trial EEG-analysis: from brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, 167(1):82–90, 2008.

Müller-Putz, G. R., Scherer, R., Pfurtscheller, G., and Rupp, R. EEG-based neuroprosthesis control: a step towards clinical practice. *Neuroscience Letters*, 382(1):169–174, 2005.

Murata, N., Takenouchi, T., and Kanamori, T. Information geometry of u-boost and bregman divergence. *Neural Computation*, 16:1437–1481, 2004.

Nguyen, T.-H., Park, S.-M., Ko, K.-E., and Sim, K.-B. Multi-class stationary CSP for optimal feature separation of brain source in BCI system. In *International Conference on Control, Automation and Systems (ICCAS)*, pages 1035–1039, 2012.

Nijboer, F., Sellers, E., Mellinger, J., Jordan, M., Matuz, T., Furdea, A., Halder, S., Mochty, U., Krusienski, D., Vaughan, T., Wolpaw, J., Birbaumer, N., and Kübler, A. A p300-based brain-computer interface for people with amyotrophic lateral sclerosis. *Clinical Neurophysiology*, 119(8):1909 – 1916, 2008.

Nijholt, A. and Tan, D. Playing with your brain: Brain-computer interfaces and games. In *Proc. of the International Conference on Advances in Computer Entertainment Technology*, pages 305–306, 2007.

Nijholt, A., Plass-Oude Bos, D., and Reuderink, B. Turning shortcomings into challenges: Brain–computer interfaces for games. *Entertainment Computing*, 1(2):85–94, 2009.

Nunez, P. L. *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, 2006.

Palva, S. and Palva, J. M. New vistas for α-frequency band oscillations. *Trends in Neurosciences*, 30(4):150–158, 2007.

Park, J. and Chung, W. Common spatial patterns based on generalized norms. In *IEEE International Winter Workshop on Brain-Computer Interface*, pages 39–42, 2013.

Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. Recipes for the linear analysis of EEG. *NeuroImage*, 28:326–341, 2005.

Patterson, J., Michalewski, H., and Starr, A. Latency variability of the components of auditory event-related potentials to infrequent stimuli in aging, alzheimer-type dementia, and depression. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 71(6):450 – 460, 1988.

Petersen, K. B. and Pedersen, M. S. The matrix cookbook, 2012. Version 20121115, http://www2.imm.dtu.dk/pubdb/p.php?3274.

Pfurtscheller, G. and Aranibar, A. Evaluation of event-related desynchronization (ERD) preceding and following voluntary self-paced movement. *Electroencephalography and Clinical Neurophysiology*, 46(2):138–146, 1979.

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

Pfurtscheller, G. and Neuper, C. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*, 239(2):65–68, 1997.

Pfurtscheller, G., Stancak Jr, A., and Neuper, C. Post-movement beta synchronization. a correlate of an idling motor area? *Electroencephalography and Clinical Neurophysiology*, 98(4):281–293, 1996.

Pfurtscheller, G., Neuper, C., Brunner, C., and da Silva, F. L. Beta rebound after different types of motor imagery in man. *Neuroscience Letters*, 378(3):156–159, 2005.

Pineda, J. A. The functional significance of mu rhythms: translating seeing and hearing into doing. *Brain Research Reviews*, 50(1):57–68, 2005.

Plumbley, M. D. Geometrical methods for non-negative ICA: Manifolds, lie groups and toral subalgebras. *Neurocomputing*, 67(161-197), 2005.

Popescu, F., Fazli, S., Badower, Y., Blankertz, B., and Müller, K.-R. Single trial classification of motor imagination using 6 dry EEG electrodes. *PLoS ONE*, 2(7):e637, 2007.

Porbadnigk, A. K., Treder, M. S., Blankertz, B., Antons, J.-N., Schleicher, R., Möller, S., Curio, G., and Müller, K.-R. Single-trial analysis of the neural correlates of speech quality perception. *Journal of Neural Engineering*, 10(5):056003, 2013.

Priestley, M. B. *Spectral analysis and time series*. Academic press, 1981.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, USA, 2009.

Ramoser, H., Müller-Gerking, J., and Pfurtscheller, G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 1998.

Rebsamen, B., Guan, C., Zhang, H., Wang, C., Teo, C., Ang, V., and Burdet, E. A brain controlled wheelchair to navigate in familiar environments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(6):590–598, 2010.

Reuderink, B. *Robust Brain-Computer Interfaces*. PhD thesis, University of Twente, 2011.

Robinson, N., Ang, K., Tee, K., and Guan, C. EEG-based classification of fast and slow hand movements using wavelet-csp algorithm. *IEEE Transactions on Biomedical Engineering*, 60(8):2123–2132, 2013.

Rousseeuw, P. J. and Driessen, K. V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Salmelin, R. and Hari, R. Spatiotemporal characteristics of sensorimotor neuromagnetic rhythms related to thumb movement. *Neuroscience*, 60(2):537–550, 1994.

Samek, W. and Kawanabe, M. Robust common spatial patterns by minimum divergence covariance estimator. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2059–2062, 2014.

Samek, W. and Müller, K.-R. Information geometry meets BCI – spatial filtering using divergences. In *IEEE International Winter Workshop on Brain-Computer Interface*, pages 1–4, 2014.

Samek, W., Kawanabe, M., and Vidaurre, C. Group-wise stationary subspace analysis - a novel method for studying non-stationarities. In *Proc. of International Brain-Computer Interface Conference*, pages 16–20. Verlag der TU Graz, 2011.

Samek, W., Müller, K.-R., Kawanabe, M., and Vidaurre, C. Brain-computer interfacing in discriminative and stationary subspaces. In *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 2873–2876, 2012a.

Samek, W., Vidaurre, C., Müller, K.-R., and Kawanabe, M. Stationary common spatial patterns for brain-computer interfacing. *Journal of Neural Engineering*, 9(2):026013, 2012b.

Samek, W., Binder, A., and Müller, K.-R. Multiple kernel learning for brain-computer interfacing. In *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 7048–7051, 2013a.

Samek, W., Blythe, D., Müller, K.-R., and Kawanabe, M. Robust spatial filtering with beta divergence. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1007–1015, 2013b.

Samek, W., Meinecke, F. C., and Müller, K.-R. Transferring subspaces between subjects in brain-computer interfacing. *IEEE Transactions on Biomedical Engineering*, 60(8): 2289–2298, 2013c.

Samek, W., Kawanabe, M., and Müller, K.-R. Divergence-based framework for common spatial patterns algorithms. *IEEE Reviews in Biomedical Engineering*, 7:50–72, 2014.

Sannelli, C., Braun, M., and Müller, K.-R. Improving BCI performance by task-related trial pruning. *Neural Networks*, 22(9):1295–1304, 2009.

Sannelli, C., Vidaurre, C., Müller, K.-R., and Blankertz, B. CSP patches: an ensemble of optimized spatial filters. an evaluation study. *Journal of Neural Engineering*, 8(2): 025012, 2011.

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

Schürmann, M. and Başar, E. Functional aspects of alpha oscillations in the EEG. *International Journal of Psychophysiology*, 39(2):151–158, 2001.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P., and Müller, K.-R. Towards adaptive classification for BCI. *Journal of Neural Engineering*, 3(1):R13–R23, 2006.

Sterman, M. B. Physiological origins and functional correlates of EEG rhythmic activities: implications for self-regulation. *Biofeedback and Self-regulation*, 21:3–33, 1996.

Suffczynski, P., Pijn, J. P. M., Pfurtscheller, G., and Da Silva, F. L. Event-related dynamics of alpha band rhythms: a neuronal network model of focal ERD-surround ERS. *Event-related desynchronization: Handbook of Electroencephalography and Clinical Neurophysiology*, pages 67–87, 1999.

Sugiyama, M. and Kawanabe, M. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT Press, Cambridge, MA, USA, 2011.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Tangermann, M., Krauledat, M., Grzeska, K., Sagebaum, M., Blankertz, B., Vidaurre, C., and Müller, K.-R. Playing pinball with non-invasive BCI. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 1641–1648, 2008.

Thiébaux, H. J. and Zwiers, F. W. The interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology*, 23(5):800–811, 1984.

Thomas, K. P., Guan, C., Lau, C. T., Vinod, A. P., and Ang, K. K. A new discriminative common spatial pattern method for motor imagery brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 56(11):2730–2733, 2009.

Tomioka, R. and Müller, K.-R. A regularized discriminative framework for EEG analysis with application to brain-computer interface. *NeuroImage*, 49(1):415–432, 2010.

Tomioka, R., Hill, J., Blankertz, B., and Aihara, K. Adapting spatial filter methods for nonstationary BCIs. In *Proc. of Workshop on Information-Based Induction Sciences (IBIS)*, pages 65 – 70, 2006.

Treder, M. S. and Blankertz, B. (C)overt attention and visual speller design in an ERP-based brain-computer interface. *Behavioral and Brain Functions*, 6:28, 2010.

van Erp, J., Lotte, F., and Tangermann, M. Brain-computer interfaces: Beyond medical applications. *Computer*, 45(4):26–34, 2012.

Van Leeuwen, W. S., Wieneke, G., Spoelstra, P., and Versteeg, H. Lack of bilateral coherence of mu rhythm. *Electroencephalography and Clinical Neurophysiology*, 44(2): 140–146, 1978.

Vapnik, V. N. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Vidal, J. J. Toward direct brain-computer communication. *Annual Review of Biophysics and Bioengineering*, 2(1):157–180, 1973.

Vidaurre, C., Kawanabe, M., von Bünau, P., Blankertz, B., and Müller, K.-R. Toward unsupervised adaptation of LDA for brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 58(3):587 –597, 2011a.

Vidaurre, C., Sannelli, C., Müller, K.-R., and Blankertz, B. Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural Computation*, 23(3):791–816, 2011b.

Vidaurre, C., Sannelli, C., Müller, K.-R., and Blankertz, B. Co-adaptive calibration to improve BCI efficiency. *Journal of Neural Engineering*, 8(2):025009 (8pp), 2011c.

von Bünau, P. *Stationary Subspace Analysis - Towards understanding non-stationary data*. PhD thesis, Technische Universität Berlin, 2012.

von Bünau, P., Meinecke, F. C., Király, F. C., and Müller, K.-R. Finding stationary subspaces in multivariate time series. *Physical Review Letters*, 103(21):214101+, 2009.

von Bünau, P., Meinecke, F., Scholler, S., and Müller, K.-R. Finding stationary brain sources in EEG data. In *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 2810–2813, 2010.

Vuckovic, A. and Sepulveda, F. A four-class BCI based on motor imagination of the right and the left hand wrist. In *1st International Symposium on Applied Sciences on Biomedical and Communication Technologies*, pages 1–4, 2008a.

Vuckovic, A. and Sepulveda, F. Delta band contribution in cue based single trial classification of real and imaginary wrist movements. *Medical & Biological Engineering & Computing*, 46(6):529–539, 2008b.

Wang, D., Miao, D., and Blohm, G. Multi-class motor imagery EEG decoding for brain-computer interfaces. *Frontiers in Neuroscience*, 6, 2012a.

Wang, H. Harmonic mean of kullback-leibler divergences for optimizing multi-class EEG spatio-temporal filters. *Neural Processing Letters*, 36(2):161–171, 2012.

Wang, H. Discriminant and adaptive extensions to local temporal common spatial patterns. *Pattern Recognition Letters*, 34(10):1125–1129, 2013.

Wang, H. and Xu, D. Comprehensive common spatial patterns with temporal structure information of EEG data: Minimizing nontask related EEG component. *IEEE Transactions on Biomedical Engineering*, 59(9):2496–2505, 2012.

Wang, H. and Zheng, W. Local temporal common spatial patterns for robust single-trial EEG classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(2):131–139, 2008.

Wang, H., Tang, Q., and Zheng, W. L1-norm-based common spatial patterns. *IEEE Transactions on Biomedical Engineering*, 59(3):653–662, 2012b.

Winkler, I., Haufe, S., and Tangermann, M. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, 7(1): 30, 2011.

Wojcikiewicz, W., Vidaurre, C., and Kawanabe, M. Stationary common spatial patterns: Towards robust classification of non-stationary EEG signals. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 577–580, 2011.

Wolpaw, J. and Wolpaw, E. W., editors. *Brain-Computer Interfaces: Principles and Practice*. Oxford Univ. Press, 2012.

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

Wu, W., Chen, Z., Gao, S., and Brown, E. N. A probabilistic framework for learning robust common spatial patterns. *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2009:4658–61, 2009.

Yong, X., Ward, R., and Birch, G. Robust common spatial patterns for EEG signal preprocessing. In *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 2087–2090, 2008.

Zhang, H., Chin, Z. Y., Ang, K. K., Guan, C., and Wang, C. Optimum spatio-spectral filtering network for brain-computer interface. *IEEE Transactions on Neural Networks*, 22(1):52–63, 2011.

Zhang, H., Yang, H., and Guan, C. Bayesian learning for spatial filtering in an EEG-based brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 24(7):1049–1060, 2013.