Towards EEG *source* connectivity analysis

Stefan Haufe

Towards EEG *source* connectivity analysis

Dipl.-Inform. Stefan Haufe

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

Prüfungsausschuss

Vorsitzender:	Prof. Dr. Klaus Obermayer
Gutachter:	Prof. Dr. Klaus-Robert Müller
	Prof. Dr. Lars Kai Hansen
	Dr. Guido Nolte

Tag der wissenschaftlichen Aussprache: 22.11.2011

Berlin, 2012

D83

Abstract

Due to its temporal resolution in the millisecond range and other compelling properties such as noninvasiveness, portability and relatively low costs, electroencephalography (EEG) is the tool of choice for studying brain dynamics in humans. However, any neurophysiological interpretation of EEG data is hindered by the fact that the signals related to electrical activity in source brain regions are spread across the EEG sensors due to a process called volume conduction, the inversion of which is an ill-posed inverse problem. In the EEG-based analysis of directed information flow between brain regions, volume conduction poses a serious challenge, since multiple active sources have to be assumed, the contributions of which, however, mix into all EEG sensors.

With this thesis we contribute to the field of EEG-based brain connectivity analysis. We start by conducting an extensive survey of relevant approaches and go on to assess their performance on simulated EEG data, which are realistically generated as mixtures of interacting sources. We observe that most of the tested methods are not able to determine the connectivity structure of the underlying sources because either i) the effect of volume conduction is neglected or ii) the assumptions made to estimate mixed sources do not comply with the requirements of source connectivity analysis or even with the physics of the EEG.

The central contribution of this thesis is the development of source demixings that meet theoretical requirements for source connectivity analysis and are applicable in practice. We achieve this using two different strategies. In inverse source reconstruction, a given physical model of EEG generation is inverted under physiologically motivated constraints. We derive an inverse methodology, S-FLEX, which is able to recover multiple source of arbitrary shape and depth using a sparsity penalty that assures rotational invariance of the solution. Our method is applicable to entire time series as a preprocessing step to source connectivity analysis. Its favorable localization properties are empirically evaluated on real data as well as in simulations and are demonstrated to be the key to correct source connectivity determination. In the course of the thesis, we also present a blind source separation (BSS) technique, SCSA, which estimates the underlying brain sources and their mixing patterns jointly under assumptions on connectivity structure of the sources. We demonstrate that SCSA performs well on simulated data, if mild assumptions regarding the non-Gaussianity of the source variables hold. By way of an outlook, we present a hybrid approach that combines S-FLEX inverse source reconstruction with SCSA blind source separation and thereby fuses physical and dynamical assumptions on the sources.

In the final part of the thesis, we analyze information transfer between sources of the human alpha rhythm during rest. Our analysis of S-FLEX source estimates yields a number of insights which could not have been obtained from mere sensor-space analysis. We observe that generators of alpha band activity are mostly symmetrically distributed on the cortex. However, the strength of the sources nor the corresponding interaction patterns are exclusively symmetric. These findings support the hypothesis of a consistent dominant laterality in the population.

Zusammenfassung

Aufgrund seiner hohen zeitlichen Auflösung, Nichtinvasivität und Portabilität, sowie seiner relativ geringen Kosten ist Elektroenzephalografie (EEG) die momentan gebräuchlichste Methode zur Messung dynamischer Informationsverarbeitung im Gehirn. Die Interpretation von EEG-Daten unter neurophysiologischen Gesichtspunkten wird jedoch dadurch erschwert, dass die im Gehirn entstehenden Signale durch Volumenleitung in alle EEG-Sensoren ausstrahlen. Bei der Schätzung von Interaktionen zwischen Gehirnregionen ist dies besonders hinderlich, da hier stets mehrere aktive Quellregionen angenommen werden müssen, deren Beiträge sich jedoch infolge von Volumenleitung in den EEG-Kanälen überlagern.

Mit dieser Dissertation leisten wir Beiträge zur EEG-basierten Analyse von Gehirnkonnektivität. Dazu unterziehen wir zunächst gängige Ansätze einer umfangreichen Evaluation anhand simulierter EEG-Daten, die in realistischer Weise als Mischung interagierender Quellen erzeugt werden. Wir beobachten, dass die Mehrzahl der getesteten Methoden nicht in der Lage ist, die Konnektivitätsstruktur der zugrundeliegenden Quellen zu ermitteln. Dies kann in den meisten Fällen darauf zurückgeführt werden, dass entweder i) Volumenleitung bei der Modellierung nicht berücksichtigt wird, oder dass ii) die Annahmen zur Quellschätzung der Existenz von Interaktionen oder gar physikalischen Gegebenheiten widersprechen.

Das Hauptaugenmerk der Arbeit liegt auf der Entwicklung von Methoden, die sowohl theoretisch zur Untersuchung von Quellenkonnektivität geeignet, als auch praktisch einsetzbar sind. Hierbei verfolgen wir zwei verschiedene Ansätze. Bei der inversen Quellenrekonstruktion wird ein gegebenes physikalisches Modell der Volumenleitung unter physiologisch motivierten Bedingungen invertiert. Wir präsentieren eine Inversionsmethode namens S-FLEX, die durch geeignete Annahmen in der Lage ist, aktive Gehirnregionen beliebiger Form und Tiefe anhand von EEG-Daten zu rekonstruieren. Unsere Methode eignet sich auch zur Rekonstruktion kompletter Quellzeitreihen als Ausgangspunkt für die Konnektivitätsanalyse. Unsere empirischen Studien belegen die höhere Lokalisierungsgenauigkeit von S-FLEX im Vergleich zu Standardmethoden, sowie die damit einhergehende Verbesserung der Schätzung von Quellenkonnektivität. Im weiteren Verlauf der Arbeit präsentieren wir eine Methode zur sogenannten blinden Quellenrekonstruktion, die in der Lage ist, die Zeitreihen der aktiven Gehirnquellen sowie deren Mischungsverhältnisse im EEG allein anhand der Annahme, dass einige der Quellen untereinander interagieren, zu schätzen. Wir zeigen empirisch, dass diese, SCSA genannte, Methode korrekte Ergebnisse liefert, sofern bestimmte Voraussetzungen die Verteilung der Quellvariablen betreffend erfüllt sind.

Im letzten Teil dieser Arbeit analysieren wir den Informationsaustausch zwischen Gehirnregionen mit hoher alpha Aktivität im Ruhezustand. Unsere Untersuchung ergibt, dass die entsprechenden Regionen symmetrisch auf der Großhirnrinde angeordnet sind. Jedoch ist sowohl die Stärke der alpha Generatoren als auch deren geschätzte Vernetzung asymmetrisch, was auf eine über die Stichprobe konsistente dominante Körperhälfte hindeutet.

Acknowledgements

I am grateful for the support I received from Prof. Klaus-Robert Müller and Dr. Guido Nolte during the entire duration of my Ph.D. studies. I conducted this work as a member of Klaus' "Intelligent Data Analysis" group at Fraunhofer FIRST, Berlin, which later became the "Machine Learning" group at TU Berlin. The great working conditions in this lab allowed me throughout to focus on research without having to worry about some of the typical woes of a Ph.D. candidate. My honest thanks go to Klaus for letting me choose my own research theme, for encouraging me to proceed in this direction and for providing personal and scientific advice in many situations. I am equally happy to have worked with Guido Nolte, who introduced me to the field of neuroimaging. His broad scientific interest and knowledgeability made him my most important scientific advisor and the discussions with him often pointed me towards interesting new research directions. I would furthermore like to express my gratitude to Prof. Lars Kai Hansen, who kindly agreed to be the external referee of this thesis, and to Katja Biermann and Duncan Blythe, who proof-read the manuscript. During my time in Berlin I benefited from being surrounded by many excellent scientists. I feel particularly grateful to Benjamin Blankertz, Ryota Tomioka, Motoaki Kawanabe, Vadim Nikulin, Joaquin Quiñonero Candela and Gabriel Curio who invested their time to teach me about their specialties. I am also thankful to Andrea Gerdes, Dominik Kühne and Imke Weitkamp for the organisational and technical support, as well as to Eike Schmidt for providing me data. My thanks go to all those colleagues with whom I had a great time in- and outside the office without necessarily being engaged in scientific exchange. While this list is certainly incomplete, I would like to mention Andreas Ziehe, Basti Venthur, Cecilia Maeder, Claudia Sannelli, Carmen Vidaurre, Felix Biessmann, Frank Meinecke, Katja Hansen, Leo Jugel, Konrad Rieck, Marius Kloft, Martijn Schreuder, Marton Danóczy, Mikio Braun, Nicole Krämer, Nikolay Jetchev, Paul von Bünau, Siamac Fazli, Steven Lemm, Tobias Lang, Ulf Brefeld and Yakob Badower in this respect. Finally, I would like to thank my family for the moral support and my girlfriend Katja for the good times during this period.

This work was financially supported by the Technische Universität Berlin, the Fraunhofer-Institute for Computer Architecture and Software Technology, Berlin, the german Federal Ministry of Education and Research (grant Nos. 16SV2234 and 01GQ0850), the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886, and the German Academic Exchange Service.

Contents

1 Introduction

1.1 EEG-based brain connectivity analysis

In the endeavor towards a better understanding of the human brain's functioning, the analysis of task-dependent information transfer between brain regions plays a crucial role.

Modern neuroscience began with the first recording of human electroencephalography (EEG) by Pravdich-Neminsky (1913). About ten years later, Berger started to systematically study the EEG signal and discovered the human alpha rhythm (Berger, 1938). Since then, a huge number of neurophysiological phenomena have been discovered and related to experimental variables in uncountably many psychophysiological studies (Davis, 1939; Sutton et al., 1965; Kornhuber and Deecke, 1965; Spehlmann, 1965; Sutton et al., 1967; Jeffreys and Axford, 1972; Farwell and Donchin, 1988; Pfurtscheller and Lopez da Silva, 1999, to name some). However, these phenomena could only be linked to brain anatomical structures with the advent of EEG inverse source reconstruction (Jeffs et al., 1987; Ioannides et al., 1990; Scherg and Ebersole, 1993; Hämäläinen and Ilmoniemi, 1994; Pascual-Marqui et al., 1994; Matsuura and Okabe, 1995; Mosher and Leahy, 1999) and alternative functional neuroimaging modalities such as positron emission tomography (PET, Ter-Pogossian et al., 1975), functional magnetic resonance imaging (fMRI, Roy and Sherrington, 1890; Ogawa et al., 1990) and invasive electrophysiological recordings such as electrocorticography (ECoG, Penfield et al., 1954). These novel techniques facilitated the releases of comprehensive atlases of brain functions (Talairach and Tournoux, 1988; Rohlfing et al., 2010), as well as atlases depicting the structural connections between brain regions (Sporns et al., 2005; Murayama et al., 2006; Hagmann et al., 2007). Concomitant with and even prior to these developments, tremendous progress had been made in understanding the structural organization of the brain and the physicochemical mechanisms underlying cerebral information storage and transfer on the cellular level (Deiters, 1865; Golgi, 1885; Gotch, 1902; Ramón y Cajal, 1904; Dale, 1914; Loewi, 1921; Erlanger and Gasser, 1924; von Euler, 1946; Hodgkin and Huxley, 1952; Carlsson et al., 1957; Eccles, 1964; Kebabian and Greengard, 1971; Kandel et al., 2000). However, there is still a considerable gap to bridge in linking today's knowledge about the dynamics of single neurons to the macroscopic effects observable with present neuroimaging technologies, although large-scale integrative models are under development (Markram, 2006).

An intermediate step towards unified brain modeling are models that describe the dynamics and interaction patterns of brain regions on a macroscopic scale. This is the strategy pursued by the majority of studies in *brain connectivity analysis*. There exist various competing definitions of "connectivity" and there is an even greater disagreement on how to properly measure connectivity according to the various definitions. Regarding the first point, a distinction between *structural*, *functional* and *effective connectivity* has been widely agreed on (Friston, 1994; Horwitz, 2003; Jirsa and McIntosh, 2007). Structural connectivity refers to the static anatomical structure of the brain, which can be acquired, for example, by a single scan using a high-resolution anatomical MRI

(Lauterbur, 1973) or diffusion tensor imaging (DTI, Moseley et al., 1990) device. Functional and effective connectivity are defined with respect to a mental task and refer to "coupled" activity of two neuroanatomical entities during task execution. The common distinction between the two is that effective connectivity is directed, i. e., describes a driver-receiver relationship, while functional connectivity is not. Originally, functional connectivity had been equated with instantaneous correlation (Friston, 1994) but it is useful to extend this definition to arbitrary measures of undirected functional dependencies which are symmetric in their arguments. Analogously, effective connectivity might be quantified by any asymmetric function. There is an infinite number of functions complying with these definitions and indeed, the proposed connectivity measures are numerous and originate from diverse fields such as graph theory, signal processing and Bayesian statistics. Here, we are primarily concerned with measures of *effective* connectivity. These can be roughly divided into dynamic causal modeling and Granger-causal modeling approaches, although there exist also differing concepts (Sun et al., 2008; Janzing and Schölkopf, 2010).

Dynamic causal modeling (DCM, Friston et al., 2003; Kiebel et al., 2008; Friston, 2009) assumes that there exist a number of different models of how the observation sequences are generated. The specifications of these models include hypothetical source regions, the signal transformation from the source regions to the measurement sensors and a directed graph modeling the causal links between the souce regions, which are to be tested. There is no standard choice for the models to be compared; they have to be specified manually using domain knowledge about the mental task under study. In the DCM methodology, the parameters of the various models are optimized to match the data as well as possible, while fulfilling certain prior expectations, which must also be pre-specified. The most likely model is selected using Bayesian decision theory and its network topology (including the effective connections) as well as the fitted model parameters are subjected to neurophysiological interpretation.

A huge class of effective connectivity measures is commonly subsumed under the term *Granger-causal modeling* (GCM), which is used to express the presence of the following two properties. First, in contrast to DCM, the estimation of effective connections between the variables of interest (i. e., sources) is usually not restricted by a predetermined network topology. Rather, the presence of connections is estimated exhaustively for all pairs of variables in a completely data-driven manner. Second, driver-receiver relationships are defined using the argument that *the cause* (measured through the driving variable) *temporally precedes the effect* (measured through the receiving variable). In terms of time series, this implies that the sending variable's time series contains information about future values of the receiving variable's time series. Granger-causality is one way of quantifying this effect. It is based on Granger's consideration that knowledge of the driver's time series at a time should improve the *prediction* of the receiver's time series at a later time. This practical definition has been implemented by a number of estimators (Granger, 1969; Kamiński and Blinowska, 1991; Baccalá and Sameshima, 2001; Valdés-Sosa et al., 2005) using autoregressive models and has found widespread applications (e. g., in neuroscience Kamiński et al., 1997; Hesse et al., 2003; Brovelli et al., 2004; Babiloni et al., 2004; Astolfi et al., 2004; Roebroeck et al., 2005; Babiloni et al., 2005; Eichler, 2005; Supp et al., 2007; Blinowska et al., 2010). However, not every approach that adopts the temporal definition of effective connectivity falls into the category of Granger causality. Counterexamples are Nolte et al. (2004) and Nolte et al. (2008), which are completely model free approaches to connectivity analysis. Both are based on detecting nonzero phase lags by analyzing the imaginary part of the time series' cross-spectrum.

There is an ongoing debate on whether DCM, GCM or the analysis of phase lags should be prefered in brain effective connectivity analysis (Valdes-Sosa et al., 2011). A crucial point to consider is the properties of the measurement modality used to study connectivity. Measures of metabolic functions such as fMRI and PET allow one to directly study internal brain dynamics with high spatial resolution, but their temporal resolution lies in the range of a second, which is disadvantageous for methods which define drivers and receivers in terms of temporal precedence. Indeed, Granger-causal analyses of fMRI data have resulted in only very few significant connections (Valdés-Sosa et al., 2005; Eichler, 2005), indicating that the temporal resolution of fMRI is too low to capture time-delayed neuronal information transfer at the relevant scales. Electrophysiological measurements reflect neuronal activity more directly than fMRI and PET, and with sampling rates of up to several kHz. However, unlike for fMRI and PET, it is not possible to exactly reconstruct the brain's internal activity from electrophysiology measured outside the head. Therefore, electrophysiology-driven brain effective connectivity analysis requires either invasive measures, or the (explicit or implicit) solution to an ill-posed *inverse problem*. Invasive recordings in humans are rarely indicated, which leaves one with electroencephalography (EEG) and the related magnetoencephalography (MEG). These modalities utilize extracranial sensors to pick up signals related to the electric activity of gross populations of synchronized neurons. In the of case EEG, these are scalp electric potentials, while MEG measures the corresponding magnetic fields. In both cases the signal is spatially diffused while traversing from the source regions to the sensors by a process called *volume conduction*. Regarding the interpretability of EEG/MEG data in terms of the underlying active source regions, this poses a serious challenge. The situation is most severe in brain connectivity studies, where multiple sources must be assumed, the contributions of which, however, mix into all sensors. The fact that volume conduction has to be accounted for in EEG- and MEG-based brain connectivity studies is receiving more and more attention nowadays (Nolte et al., 2004; Schlögl and Supp, 2006; Nolte et al., 2006, 2008; Gómez-Herrero et al., 2008; Nolte and Müller, 2010) and has led to the request for genuine *EEG/MEG source connectivity analysis*, i. e., the analysis of brain connectivity on source estimates derived from EEG/MEG data (Schoffelen and Gross, 2009), which is the subject of this thesis.

1.2 Scientific proposal

The purpose of this work is to contribute to the field of EEG/MEG-based brain connectivity analysis. We do this in three steps. First, we conduct an extensive survey of the field and evaluate relevant approaches on simulated data. These simulations signify the two core theses of this work:

i. *Volume conduction needs to be accounted for explicitly in EEG/MEG analyses, the results of which should be subjected to neurophysiological interpretation.*

ii. *The assumptions used to recover sources in EEG/MEG connectivity studies need to comply with the theoretical requirements of source connectivity analysis.*

In the second step, we are concerned with the development of novel methods for EEG/MEG source reconstruction and connectivity analysis, in which these insights fully enter. The efficacy of the novel approaches is validated on the same data. In the third step, the applicability of our concepts to real data is demonstrated.

We here focus on EEG, which is far more widely used than MEG due to dramatically lower costs and better portability. Virtually any analysis discussed in this thesis can, however, be transferred to the MEG domain without amendments to the mathematics involved (an exception is the calculation of lead fields, which differs for EEG and MEG). Likewise, we restrict ourselves here to the analysis of *effective* connectivity following the *temporal definition*. Nevertheless, some of the results presented in this thesis are also relevant for other branches of connectivity analysis.

1.3 Outline of the thesis

Following this introduction, the thesis begins with a chapter on background (Chapter 2), in which we give an overview on EEG signal generation and the most important neurophysiological EEG phenomena, before coming to mathematical concepts from machine learning, statistics and signal processing that are of general use. The last part of the chapter introduces popular measures of effective connectivity, as well as the two general strategies for obtaining source estimates from EEG data, inverse source reconstruction and blind source separation.

The following chapter (Chapter 3) introduces a simple simulated EEG dataset, which we use to evaluate representative measures of effective connectivity under standardized conditions. We also evaluate a number of preprocessing steps, including prominent inverse source reconstruction and blind source separation techniques.

Chapters 4 and 5 contain the core methodological contributions of this thesis. Chapter 4 deals with inverse source reconstruction. Here, we successively develop a novel methodology for estimating possibly interacting sources by inverting a physical model describing the effect of volume conduction. The core features of the final approach are i) invariance with respect to rotations of the coordinate system, ii) the ability to reconstruct source regions of arbitrary shape and depth, iii) the ability to spatially distinguish multiple neighboring source regions and iv) the applicability to time series data under the assumption of time-invariant spatial signatures of the sources. While i) and ii) are purely physiologically-motivated, iii) and iv) are crucial requirements for source connectivity analysis. We propose a measure for comparing the localization accuracy of arbitrary reconstructed sources, which we use to demonstrate that our approach outperforms the state-of-the-art on simulated data. This result is confirmed on real data, for which the "ground truth" is known. Finally, our inverse source reconstruction is successfully applied to the problem of EEG source connectivity analysis using the simulated data introduced in Chapter 3.

Chapter 5 describes the development of a novel blind source separation (BSS) approach to source connectivity estimation. We start by proposing two Granger-causal measures of effective connectivity, by which it is possible to estimate sparse connectivity graphs (i. e., graphs with few connections). We extend one of these approaches to a BSS setting, in which the mixing of the sources caused by volume conduction is estimated jointly with the connectivity graph in a completely data-driven fashion. We further outline an extension, in which the mixing patterns are constrained to reflect only brain sources with physiologically meaningful spatial signatures. That is, we integrate inverse source reconstruction techniques developed in Chapter 4 into our blind source separation framework. Note that our approaches stand in contrast to the majority of BSS techniques, the assumptions of which either prohibit the analysis of interactions between the underlying brain sources or disregard physical constraints imposed by volume conduction.

Before summarizing the results of this thesis in Chapter 7, we present an EEG study of human resting state brain connectivity (Chapter 6). Here, we are able to reproduce results from the literature, which were obtained by means of sensor-space connectivity analysis. We, however, show that results obtained in sensor space are generally sensitive to changes of certain technical parameters, which severely limits their interpretability under neurophysiological aspects. Notably, this limitation does not apply to source-space approaches such as those presented here. Source connectivity analysis using the inverse source reconstruction methodology derived in Chapter 4 reveals a number of symmetric source regions, which are active during rest. Interestingly, these regions exhibit complex and partly asymmetric interaction patterns.

1.4 List of included published work

The following publications are included in large parts into this thesis.

[1] Haufe, S., Nikulin, V. V., Ziehe, A., Müller, K.-R., Nolte, G., 2008. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. NeuroImage 42, 726–738.

[2] Haufe, S., Nikulin, V. V., Ziehe, A., Müller, K.-R., Nolte, G., 2009. Estimating vector fields using sparse basis field expansions. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21. MIT Press, pp. 617–624.

[3] Haufe, S., Nolte, G., Müller, K.-R., Krämer, N., 2010. Sparse causal discovery in multivariate time series. In: Guyon, I., Janzing, D., Schölkopf, B. (Eds.), Causality: Objectives and Assessment. Vol. 6 of JMLR W&CP. pp. 97–106.

[4] Haufe, S., Tomioka, R., Dickhaus, T., Sannelli, C., Blankertz, B., Nolte, G., Müller, K.-R., 2010. Localization of class-related mu-rhythm desynchronization in motor imagery based brain-computer interface sessions. In: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE. pp. 5137–5140.

[5] Haufe, S., Tomioka, R., Nolte, G., Müller, K.-R., Kawanabe, M., 2010. Modeling sparse connectivity between underlying brain sources for EEG/MEG. IEEE Trans Biomed Eng 57, 1954–1963.

[6] Haufe, S., Tomioka, R., Dickhaus, T., Sannelli, C., Blankertz, B., Nolte, G., Müller, K.-R., 2011. Large-scale EEG/MEG source localization with spatial flexibility. NeuroImage 54, 851–859.

2 Fundamentals

In this chapter, we introduce concepts that are fundamental to understanding the thesis. We start with a section on EEG neurophysiology (Section 2.1), where we explain the generation of the EEG signal and describe the various features seen in the EEG. The remaining part of the chapter introduces mathematical tools that are useful for the development of our methods or for general data analysis, as well as methods from the literature with which we shall compare our methods. We devote a section to notation and basic definitions (Section 2.2), which is followed by a description of basic machine learning concepts in the context of EEG modeling (Section 2.3). Section 2.4 contains a brief description of the statistical testing approach used throughout the thesis. Signal processing topics cover the last four sections of the chapter. In Section 2.5, we discuss spectral filtering, the Fourier transform and autoregressive models. The latter two techniques are fundamental to the definition of the most prominent measures of time-lagged effective connectivity, which are introduced in Section 2.6. The last two sections describe techniques for the decomposition of EEG data. Since the EEG signal is known to be a superposition of source signals due to volume conduction, most decompositions techniques can be regarded as factorizations of the EEG signal into the volume conduction (mixing) part and the source signals. If the mixing matrix is obtained from a physical model we speak of inverse source reconstruction, which is explained in detail in Section 2.7 along with prominent examples. An alternative to using a constant mixing matrix are data-driven blind-source separation approaches that identify spatial mixing/demixing coefficients based on statistical assumptions on the sources. We discuss relevant approaches in Section 2.8.

2.1 Neurophysiology

The electroencephalographic (EEG) signal is an electric potential that is measured on the scalp. It comprises brain activity caused by electric activity of (mainly) cortical neurons as well as several types of physiological and non-physiological artifacts. In this section, we briefly discuss the mechanisms underlying the transformation of cerebral electrical activity into EEG potentials and describe the two most important neurophysiological phenomena observed in EEG signals, event-related potentials and oscillations.

2.1.1 EEG signal generation

Information processing in the brain takes place in approximately one-hundred billion interconnected *neurons*, which are specialized cells that consist of a cell body (the *soma*), *dendrites*, an *axon* and an enclosing *membrane*. The electroencephalographic signal arises as a result of synchronous activity of large populations of neurons with similar spatial orientation. Following Baillet et al. (2001) and Wolters and de Munck (2007), this process can be summarized as follows (see Figure 2.1 for a depiction). Neurons are electrically charged through transport proteins that

pump ions across their membranes. An axonal potential leads to the generation of excitatory postsynaptic potentials (ESPs) at the apical dendritic tree, which causes the dendrite to release ions through its membrane. The resulting *depolarization* of the membrane establishes an electrical potential difference between the apical dendrite and the non-excited cell soma and basal dendrites. This causes two types of ionic current flows. Currents that directly travel within the neuronal dentritic trunk are called (intracellular) *primary* currents. The rule of conservation of electric charges implies that there is also current flow in the opposite direction. The respective curents are called (extracellular) *secondary* currents, because they travel through the exterior of the neuron. In certain cerebral structures, there exist large populations of equally-aligned neurons. If these neurons are synchronously activated, their primary currents add. The corresponding secondary currents, which spread over the whole volume conductor, are strong enough to be measurable as scalp potentials. The propagation of secondary currents from the *sources* (the generators of the primary currents) through the biological tissue towards the measurement sensors is called *volume conduction*. Importantly, this process is governed by the geometric and conductive properties of the traversed media, which are the brain, scull and scalp tissues and the cerebrospinal fluid (CSF). It is possible to mathematically model the propagation of secondary currents for a given (primary) current source and volume conductor model using the fact that all currents are *passive* in the frequency ranges of interest (see Section 2.7). In general, the electric potential observed at the scalp surface is more widespread the deeper the generating source is, while it is stronger, the more neurons are acting synchronously, the more similar their spatial alignment is and the more superficially they are located. Pyramidal cortical neurons are the likely main contributors to EEG potentials, because they are superficially located and spatially similarly aligned (perpendicular to the cortical surface). The dynamics observed in the EEG signals is assumed to be caused by interacting networks of such active cortical patches (Baillet et al., 2001).

2.1.2 Event-related potentials

Event-related potentials (ERPs) are characteristic reproducable EEG potential changes due to internal or external stimulation. Internal stimulation may refer, e. g., to the semantic processing as occuring during the perception of infrequent stimuli (Sutton et al., 1965, 1967), or to the preparation of movements (Kornhuber and Deecke, 1965). In contrast, external stimulation is associated with sensory input. Respective ERPs are reported for tactile, electric (Penfield and Boldrey, 1937), visual (Spehlmann, 1965; Jeffreys and Axford, 1972) and auditory (Davis, 1939) stimulation. The study of event-related potentials has widespread applications in clinical diagnosis and psychophysiology (Fabiani et al., 2000), as well as brain-computer interfacing (Farwell and Donchin, 1988; Schreuder et al., 2010; Blankertz et al., 2011) and mental state monitoring (Haufe et al., 2011b). A sequence of event-related potentials occuring during induced emergency braking in a simulated driving task is depicted in Figure 2.2 (a).

Event-related potentials exhibit a spatio-temporal signature that depends on the location of the active cerebral current sources involved in the mental processing and the spatio-temporal dynamics of source activition. The amplitude of ER potentials typically lies in the range of 1–20 μV, which is approximately 1–2 orders of magnitude below the noise (cerebral background activity and artifacts) level. To increase the signal-to-noise ratio, ERPs are usually estimated from multiple repetitions of the same mental task by averaging the respective stimulus-locked EEG segments.

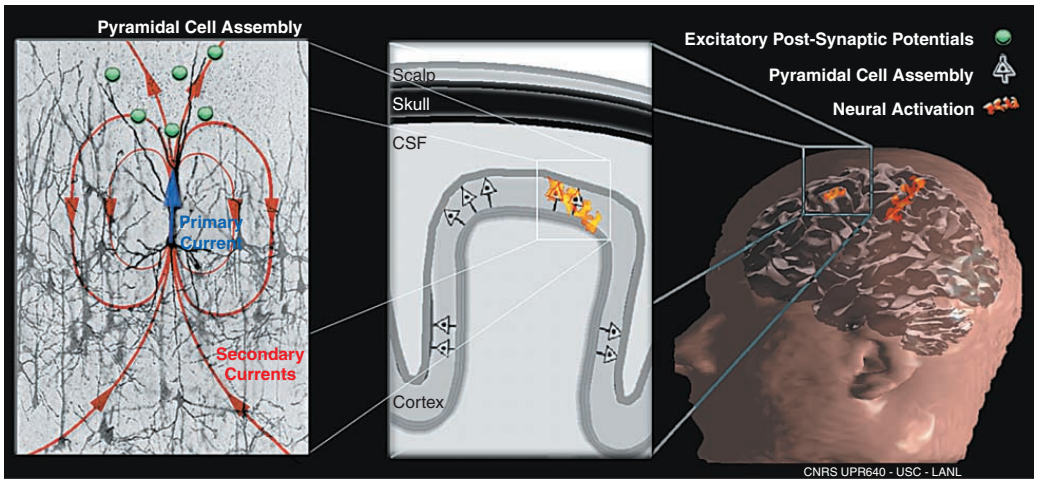


Figure 2.1: Generation of the EEG signal (figure taken from Baillet et al. (2001)). Left: Excitatory postsynaptic potentials at the apical dendritic tree cause its membrane to depolarize, by which means an electrical potential difference between apical dendrite and cell soma on one hand and the basal dendrites on the other hand is established. This causes two types of ionic current flows. Primary currents (blue) travel directly within the neuronal dentritic trunk. Secondary currents (red) travel through the entire volume conductor and are measurable by EEG, if the corresponding primary currents are strong enough. Center: Synchronously active populations of pyramidal cortical neurons are the likely main contributors of EEG potentials due to their superficial location and their homogeneous spatial orientation (perpendicular to the cortical surface), which causes their primary currents to add. Right: Interacting networks of several active cortical patches are the assumed main causes for the dynamics observed in the EEG signal.

2.1.3 Rhythmic activity

The EEG power spectrum exhibits a characteristic $1/f$ (*pink noise*) shape, which is in many cases superimposed by one ore more spectral peaks representing strong oscillatory activity in narrow frequency ranges. Most notably, a peak within the *alpha band* (8 to 14 Hz), which sometimes co-occurs with a peak at the doubled frequency range (the *beta band*, extending from 15 to 30 Hz), is observed in most subject's EEG. Alpha activity was recorded in one of the earliest EEG experiments conducted by Berger (1938). Other physiologically relevant spectral bands include the *delta band* (up to 4 Hz), the *theta band* (5 to 8 Hz) and the *gamma band* (30 to more than 100 Hz). The various spectral peaks highly differ in their spatial distribution on the scalp, indicating that they are generated by differing brain networks. Already the alpha band is known to comprise at least two functionally distinct rhythms, which differ in their EEG topographies as well as the shapes of their waveforms. While the so-called *mu rhythm* has a more central alignment, the (stronger) alpha activity is observed in more parieto-occipital scalp sites.

The posterior alpha rhythm has been related to a number of behavioral markers including vigilance (Schmidt et al., 2009; Schubert et al., 2009), fatigue (Simon et al., 2011) and the inhibitions of actions (Klimesch et al., 2007). Most notably, alpha power is modulated by the amount of relaxation of the visual system, and is strongest when the eyes are closed. The strength of the mu rhythm, on the other hand, is related to the level of relaxation of the motor system, and

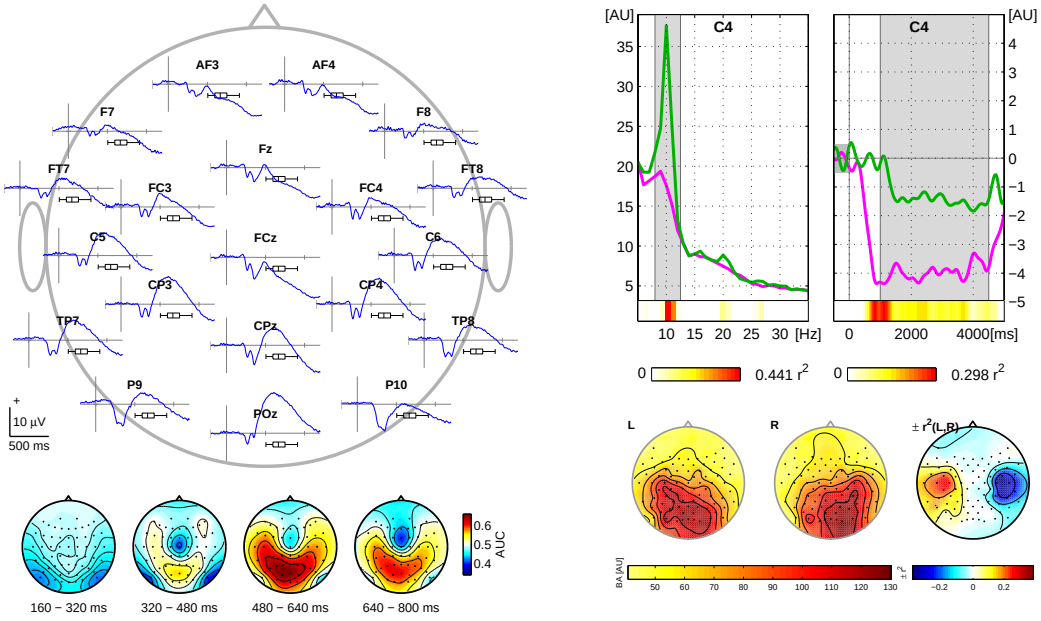


Figure 2.2: (a) Grand-average (N = 18) event-related potentials (ERP) occuring during induced emergency braking in a simulated driving task (figure taken from Haufe et al. (2011b)). Upper panel: ERP curves locked to the brakelight onset of a preceding car. Lower panel: Topographical maps depicting how well emergency braking and normal driving situations can be distinguished based on ERPs. (b) Average (N = 1) event-related desynchronization (ERD) in the mu band during motor imagery of the left hand (magenta color) and right hand (left color); figure taken from Haufe et al. (2010b). Upper left panel: Spectrum at electrode C4 (right hemisphere) depicting desynchronization in the mu band during left hand motor imagery. Upper right panel: time course of log-power in the mu band depicting desynchronization during 3 s of left hand motor imagery. Lower panel: topographical maps of power in the mu band (left two plots) and differences in mu power between conditions (red and blue colors depicting the areas in which the power is stronger during left and right hand imagery, respectively)

decreases when movements are observed, planned, executed or even only imagined (Pfurtscheller and Lopez da Silva, 1999). Moreover, a dependence on the frontal gamma rhythm has been noted (Grosse-Wentrup et al., 2011). In line with these findings, the "idling" hypothesis (see, e. g., Palva and Palva, 2007; Sabate et al., 2011) states that EEG rhythms represent the default mode of the brain in which large neuronal populations are synchronized in a "feedback loop", while any recruitment of neurons for task engagement decreases the number of "idling" neurons and hence decreases the strength of the oscillation. This process is called (*event-related*) *desynchronization* (ERD, Pfurtscheller and Lopez da Silva, 1999). Figure 2.2 (b) illustrates the phenomenon of event-related desynchronisation during motor imagery of the hands.

2.2 Notation and basic definitions

In this work we denote matrices by italic upper-case letters, vectors by bold lower-case letters and scalars by italic upper- or lower-case letters. Vectors are understood to be in columnar shape unless otherwise mentioned. The i-th entry of a vector $\mathbf{x}$ is denoted by $\mathbf{x}_i$, the i-th column of a matrix A by $\mathbf{a}_i$ and the entry in the i-th row and j-th column of a matrix A by $a_{i,j}$. For time series data, indexing may also be indicated using parentheses, i. e., $\mathbf{x}(t)$ denotes the value of the multivariate times series $\mathbf{x}$ at time t, which can also be regarded as the t-th column of a matrix X. The notation $\mathbf{x}_S$ with set-valued S denotes a vector composed of the stacked entries $\mathbf{x}_i,\ i \in S$. The transpose of a matrix A is denoted by $A^\top$ and the inverse (if it exists) of a square matrix is denoted by A^{-1}. A real-valued matrix A is *symmetric* if $A = A^\top$ and *antisymmetric* if $A = -A^\top$. An M-dimensional vector of ones is denoted by $\mathbf{1}$, while a zero vector is denoted by $\mathbf{0}$. The $M \times M$ *identity matrix* is denoted by I_M. The *unit vector* $\mathbf{e}_i$ is a vector with all zeros except for $e_{ii} = 1$. The *vectorization* operator $\mathbf{vec}\,(A) = (\mathbf{a}_1^\top, \ldots, \mathbf{a}_m^\top)^\top$ stacks the columns of A vertically. The operator $\mathrm{diag}(A) = (a_{1,1}, \ldots, a_{M,M})^\top$ extracts the *diagonal* entries of the $M \times M$ matrix A. Analogously, $\mathrm{off}(A)$ is a vector containing the *off-diagonal* terms of A. The *trace* of an $M \times M$ matrix A is denoted by $\mathrm{Tr}\{A\} = \sum_{m=1}^{M} a_{m,m}$. The *determinant* of a square invertible matrix A is denoted by $|A|$. The *ℓ_p-norm* of an M-dimensional vector $\mathbf{x}$ is defined by $\|\mathbf{x}\|_p = (\sum_{m=1}^{M} |x_m|^p)^{1/p}$ for $p \geq 1$.

A *probability density function* (pdf) $f(\mathbf{x})$ describes the likelihood of an M-dimensional real-valued continuous random variable $\mathbf{x}$ to occur at any point $\mathbf{u} \in \mathbb{R}^M$. If f is a Lebesgue-integrable nonnegative function that is normalized such that $\int_{\mathbf{u}\in\mathbb{R}^M} f(\mathbf{u})\mathrm{d}\mathbf{u} = 1$, and if $\mathcal{R} \subset \mathbb{R}^M$ has nonzero measure, then the probability of $\mathbf{x}$ to fall into $\mathcal{R}$ is $Pr[\mathbf{x} \in \mathcal{R}] = \int_{\mathbf{u}\in\mathcal{R}} f(\mathbf{u})\mathrm{d}\mathbf{u}$. The *cumulative density function* (cdf) induced by the pdf f is $\mathrm{cdf}(\mathbf{u}) = \int_{\mathbf{u}^*\leq\mathbf{u}} f(\mathbf{u}^*)\mathrm{d}\mathbf{u}^*$, where $\mathbf{u}^* \leq \mathbf{u}$ is true if and only if $u_m^* \leq u_m$ holds for all $m \in \{1, \ldots, M\}$. The *expected value* of a multivariate random variable $\mathbf{x}$ is defined by $\mathrm{E}[\mathbf{x}] = \mu_\mathbf{x} = \int_{\mathbf{u}\mathbb{R}^M} \mathbf{u} f(\mathbf{u})\mathrm{d}\mathbf{u}$, while the *covariance* is defined by $\mathrm{Cov}[\mathbf{x}] = \Sigma_\mathbf{x} = \mathrm{E}[(\mathbf{x} - \mathrm{E}[\mathbf{x}])(\mathbf{x} - \mathrm{E}[\mathbf{x}])^\top]$. The empirical estimator of the expected value is the *mean* $\widehat{\mu}_\mathbf{x} = 1/T \sum_{t=1}^{T} \mathbf{x}(t)$, where $\mathbf{x}(t)$ are realizations of $\mathbf{x}$. An empirical estimator of the covariance is $\widehat{\Sigma}_\mathbf{x} = 1/(T-1) \sum_{t=1}^{T} [(\mathbf{x}(t) - \widehat{\mu}_\mathbf{x})(\mathbf{x}(t) - \widehat{\mu}_\mathbf{x})^\top]$. The vector of empirical *standard deviations* is defined by $\boldsymbol{\sigma}_\mathbf{x} = \sqrt{\mathrm{diag}(\widehat{\Sigma}_\mathbf{x})}$, where the square-root is applied entrywise. The probability density function of a *multivariate Gaussian* distribution with mean $\mu_\mathbf{x}$ and covariance $\Sigma_\mathbf{x}$ is given by

$$f(\mathbf{x}) = |2\pi\Sigma_\mathbf{x}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \mu_\mathbf{x})^\top \Sigma_\mathbf{x}^{-1}(\mathbf{x} - \mu_\mathbf{x})\right), \tag{2.1}$$

where $\exp(x) = e^x$, $e \approx 2.7183$ and $\pi \approx 3.1415$. The notation $\mathbf{x} \sim \mathcal{N}(\mu_\mathbf{x}, \Sigma_\mathbf{x})$ is used to indicate that a random variable $\mathbf{x}$ is multivariate Gaussian distributed. A random variable $\mathbf{x}$ is *standard normally* distributed if $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$.

A *complex* number z has the form $z = a + ib$, where $i = \sqrt{-1}$ is the *imaginary unit*, and $a = \Re(z)$ and $b = \Im(z)$ are the *real* and *imaginary parts* of z, respectively. The *modulus* of a complex number z is defined as $|z| = \sqrt{(\Re(z)^2 + \Im(z)^2)}$, whereas the *power* is defined as $|z|^2$ and the *phase* as $\arg(z) = \tan^{-1}(\Im(z)/\Re(z))$. The *complex conjugate* of $z = a + ib$ is denoted by $z^* = a - ib$. The matrix of complex conjugate transposed entries of a complex-valued matrix A is called the *Hermitian* and is denoted by A^H. A matrix A is *Hermitian* if $A = A^H$. A real-valued square matrix A is *orthogonal* if $A^{-1} = A^\top$. A complex-valued square matrix A is *unitary* if $A^{-1} = A^H$.

2.3 Machine learning

2.3.1 A model of EEG data

The most general generative model of EEG data is given by

$$\mathbf{x}(t) = A\mathbf{s}(t) + \boldsymbol{\eta}(t) \,, \tag{2.2}$$

where $\mathbf{x}(t) \in \mathbb{R}^M$ is the signal of M EEG electrodes recorded at time t, $\mathbf{s}(t) \in \mathbb{R}^K$ is the activity of K brain sources at time t and $A \in \mathbb{R}^{M\times K}$ is a matrix representing instantaneous source mixing due to volume conduction. The *noise term* $\boldsymbol{\eta}(t)$ comprises uncorrelated measurement (sensor) noise as well as correlated noise, which could be due to non-task-related background activity or artifacts as caused by, e. g., eye blinks or 50 Hz line noise. Notably, EEG activity of cerebral origin is always correlated due to volume conduction, which is modeled here explicitly using the matrix A. The variables A and $\mathbf{s}$ are not identifiable given the observations $\mathbf{x}$ without further assumptions. All methods introduced later in this thesis are specifications of the general model (2.2) using various assumptions on A, $\mathbf{s}$ and the model error $\boldsymbol{\eta}$.

2.3.2 Maximum-likelihood estimation

Setting $A = \text{const.}, \operatorname{rank}(A) = K$, $K \leq M$ and assuming $\boldsymbol{\eta}(t) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_M)$, the sources $\mathbf{s}$ are uniquely defined by the *maximum-likelihood* (ML) principle. Let the number of recorded samples be T. Setting $X = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$, $S = (\mathbf{s}(1), \ldots, \mathbf{s}(T))$ and $E = (\boldsymbol{\eta}(1), \ldots, \boldsymbol{\eta}(T))$, we have $E = X - AS$. The noise probability density as a function of the unknown parameters is called the *likelihood* of the observations. In our case, the likelihood reads

$$p(S) = \prod_{t=1}^{T} \prod_{M=1}^{M} \left(2\pi\sigma^2\right)^{-½} \exp\left(-\frac{\eta_m^2(t)}{2\sigma^2}\right) \,. \tag{2.3}$$

The maximum-likelihood estimate is the variable assignment $\widehat{S}$ that maximizes the likelihood. However, it is equivalent and more convenient to obtain $\widehat{S}$ as the minimizer of the negative log-likelihood $\mathcal{L}(S) = -\log p(S)$, i. e.,

$$\begin{aligned} \widehat{S} &= \arg\min_S \frac{TM \log\left(2\pi\sigma^2\right)}{4\sigma^2} + \sum_{t=1}^{T}\sum_{M=1}^{M} \eta_m^2(t) \\ &= \arg\min_S \left\|\mathbf{vec}\,(X - AS)\right\|_2^2 \\ &= \arg\min_S \operatorname{Tr}\{(X - AS)^\top (X - AS)\} \,. \end{aligned} \tag{2.4}$$

This derivation shows that, as a result of assuming a Gaussian noise model, the maximum-likelihood estimate coincides with the so-called *ordinary least squares* (OLS) estimate, which minimizes the sum of the squared error terms. Note that this estimate is independent of the noise variance σ^2. The solution is obtained by setting the first derivative to zero

$$\frac{\partial\mathcal{L}\,(S)}{S} = 2A^\top AS - 2A^\top X = 0 \,, \tag{2.5}$$

from what follows that

$$\widehat{S} = \left(A^\top A\right)^{-1} A^\top X \,. \tag{2.6}$$

2.3.3 Regularization

When the number of observations is small compared to the number of variables, the maximum-likelihood estimate might *overfit*, i. e., fit the non-systematic components (noise) too. For example, if A is square and $K = M$, the OLS estimate of S in the above example is $\widehat{S} = A^{-1}X$ and the model error $\widehat{E} = X - A\widehat{S}$ is zero regardless of the presence of noise. Overfitting can be counteracted by imposing constraints that limit the complexity of the solution. This technique is called *regularization*. In the underdetermined case $K > M$, the OLS/ML estimator is not even uniquely defined. Here, regularization does not only prevent overfitting; it is rather a necessity in order to obtain a well-defined estimator.

While there are many ways to perform regularization (Engl et al., 2000), we here focus on approaches that add a regularization term to the negative log-likelihood. The purpose of such a *regularizer* is to penalize variable assignments with high complexity. One important family of complexity measures are norms. The combination of a least squares error measure and a regularizer that measures the squared ℓ_2-norm of linearly transformed variables is called *Tikhonov-regularization* (Tikhonov and Arsenin, 1977). A Tikhonov-regularized estimate of the sources under our model is given by

$$\begin{aligned} \widehat{S} &= \arg\min_S \|\mathbf{vec}\,(X - AS)\|_2^2 + \lambda \|\mathbf{vec}\,(\Gamma S)\|_2^2 \\ &= \left(A^\top A + \lambda \Gamma^\top \Gamma\right)^{-1} A^\top X\,, \end{aligned} \tag{2.7}$$

where the parameter λ controls the relative influence of the error and regularization terms. Note that λ is proportional to the (typically unknown) noise variance σ^2, which hence does enter the estimation here. The matrix $\Gamma \in \mathbb{R}^{K\times K}$ is used to encode desired properties of the source variables. For example, if $\Gamma = I_K$, sources with minimal energy are sought. This variant is known as *ridge regression*. If spatial relations between the sources exist, Γ can enforce, for example, spatial smoothness of the sources. This is achieved using a matrix of spatial second derivatives, called the discrete *Laplace operator* (see also Section 2.7). It is also possible to enforce constraints in the spatio-temporal domain by considering the more general Tikhonov-regularized solution

$$\begin{aligned} \mathbf{vec}\left(\widehat{S}\right) &= \mathbf{vec}\left(\arg\min_S \left\|\mathbf{vec}\,(X) - \widetilde{A}\,\mathbf{vec}\,(S)\right\|_2^2 + \lambda \left\|\widetilde{\Gamma}\,\mathbf{vec}\,(S)\right\|_2^2\right) \\ &= \left(\widetilde{A}^\top \widetilde{A} + \lambda \widetilde{\Gamma}^\top \widetilde{\Gamma}\right)^{-1} \widetilde{A}^\top\, \mathbf{vec}\,(X)\,, \end{aligned} \tag{2.8}$$

where $\widetilde{\Gamma} \in \mathbb{R}^{KT\times KT}$ is a spatio-temporal filter matrix and $\widetilde{A} = I_T \otimes A$.

The Tikhonov-regularized estimate has an interpretation as the *maximum a-posteriori* (MAP) estimate in a *Bayesian* sense, since it is the most probable variable assignment assuming a zero-mean Gaussian *prior* probability distribution of the (linearly transformed) variables (Berger, 1985). Similarly, the regularizers discussed in the following can be interpreted as MAP estimates according to different prior distributions. If the ℓ_2-norm in the regularizer is replaced by the ℓ_1-norm, *sparsity* of the solution can be achieved. For example, the so-called *lasso* (Tibshirani, 1996) estimate

$$\widehat{S} = \arg\min_S \|\,\mathbf{vec}\,(X - AS)\,\|_2^2 + \lambda \|\,\mathbf{vec}\,(S)\,\|_1 \tag{2.9}$$

contains (many) zero entries. The use of the ℓ_1-norm here corresponds to imposing the *Laplace* distribution as a prior on the source variables. The Laplace distribution is a *super-Gaussian*

distribution, which means that it has "heavier tails" (more probability mass far away from the mean) than the Gaussian distribution. A generalized lasso regularizer $\lambda \| \widetilde{\Gamma} \mathbf{vec}(S) \|_1$ yields solutions which are sparse after linear transformation, which can be used, for example, to estimate piecewise homogeneous functions (Kim et al., 2009).

If the variables have matrix structure, this can be taken into account using dedicated regularizers. One example is the *spectral norm* of a matrix, which is the ℓ_1 norm of its singular values (see Section 2.8). Penalizing the spectral norm leads to sparsity of the singular values and hence general low-rank matrix estimates (Recht et al., 2010). Another example is the $\ell_{1,2}$-norm, which is defined as the sum of the ℓ_2-norms of the columns of a matrix

$$\|S\|_{1,2} = \sum_i \|\mathbf{s}_i\|_2 \,. \tag{2.10}$$

The use of the $\ell_{1,2}$-norm (or *group lasso*) regularizer leads to columnwise sparsity. That is, the columns of the estimated matrix $\widehat{S}$ are either jointly zero or jointly nonzero (Yuan and Lin, 2006). Note that $\ell_{1,2}$-norm penalties are not restricted to matrices but can be applied to any set of variables with group structure. Geometric arguments for the sparsifying properties of lasso and group lasso regularizers are given in Tibshirani (1996) and Yuan and Lin (2006).

2.3.4 Model selection

Penalized models have at least one so-called *hyperparameter* (such as λ), which adjusts the degree of regularization. To prevent overfitting, we are interested in choosing hyperparameters such that the estimated model *generalizes* well in the sense that it reasonably explains data that has not been used for the estimation (Hastie et al., 2001; Bishop, 2007). This selection process can be regarded as a maximization of the *out-of-sample likelihood*.

Cross-validation (CV) is one way to obtain an estimate of this out-of-sample likelihood and by this means to select hyperparameters. In *k-fold* cross-validation, the dataset is split into k parts. In each fold, the model is fitted on $k-1$ (*training*) parts and the likelihood is evaluated on the remaining (*test*) part. The procedure is carried out for different assignments of the hyperparameters. The assignment that maximize the average test likelihood is selected.

If the data samples are independent and identically distributed (*i.i.d.*), cross-validation provides an unbiased estimate of the out-of-sample likelihood. If, however, dependencies exist, their influence should be mitigated by a suitable splitting scheme. For example, if the data has temporal structure, it is essential to ensure that the training samples temporally precede the test samples. Such issues are discussed in Lemm et al. (2011).

Cross-validation is not the only strategy for hyperparameter selection. Another popular approach, which, however, works only for a single regularization parameter, is to find the corner of the *L-curve*, which is a log-log plot of the model error vs. the regularization term (Hansen, 1992). Other approaches involve the evaluation of the *Bayesian Information Criterion* (BIC, Schwarz, 1978) or Akaike's "*an information criterion*" (AIC, Akaike, 1974).

2.4 Statistical testing

When data are analyzed empirically, it is important to assess whether the observed effects are systematic or only due to random fluctuations. This is done by statistical hypothesis testing. The common practice is to formulate the *null hypothesis* that no systematic effect is present. Using assumptions on the distribution of a *test statistics* derived from the collected data, it is then calculated how "likely" this null hypothesis is. The *p-value* is the probability of obtaining the observed test statistics or a more extreme one given the null hypothesis. If the p-value is smaller than a certain predefined *significance level*, the null hypothesis is rejected and the observed effect is considered "significant".

If N independent samples are available, the true (population) standard deviation σ_x of which is unknown, the statistics $\mathrm{t} = \widehat{\mu}_x / \widehat{\sigma}_x \sqrt{(N)}$ follows Student's univariate *t-distribution* with $\nu = N - 1$ *degrees of freedom* (df), where $\widehat{\mu}_x$ is the sample mean and $\widehat{\sigma}_x$ the sample standard deviation (Gosset, 1908). This is used by the *one-sample t-test* to test whether the mean of the population distribution is different from zero. The p-value is derived by evaluating the cumulative distribution function of the t-distribution, which is denoted by $\mathrm{cdf_t}(\mathrm{t}, \nu)$. The p-value according to a *two-sided* test (testing for both positive and negative deviation from zero) is $\mathrm{p} = 2\mathrm{cdf_t}(-|\mathrm{t}|, \nu)$.

In some cases we are also interested in testing whether the means of two population distributions differ. If the samples are paired, that is, if each experiment yields one sample from each distribution, the one-sample t-test can be performed on the inter-population differences. If there is no pairing, the *two-sample t-test* is appropriate. Assuming equal, but unknown, population standard deviations $\sigma_{X_1} = \sigma_{X_2}$ and $N_1 + N_2$ independent samples, the test statistics

$$\mathrm{t} = \frac{(\widehat{\mu}_{x_1} - \widehat{\mu}_{x_2})}{s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \quad \text{, where} \quad s_p^2 = \frac{(N_1 - 1)\,\widehat{\sigma}_{x_1}^2 + (N_2 - 1)\,\widehat{\sigma}_{x_2}^2}{N_1 + N_2 - 2} \tag{2.11}$$

and $\widehat{\sigma}_{X_{1/2}}$ are the sample standard deviations, follows a t-distribution with $N_1 + N_2 - 2$ degrees of freedom. The p-value can be obtained from the cdf of this distribution.

When working with neural signals with high inter-subject variability, it is often desirable to obtain a *grand-average* p-value for the entire subject population. To this end, it is convenient to transform the subjects' individual t-scores (following different t-distributions depending on the numbers of samples) to standard normally distributed z-scores using the transformation $\mathrm{z} = \mathrm{cdf_z^{-1}}(\mathrm{cdf_t}(\mathrm{t}, \nu))$ where $\mathrm{cdf_z^{-1}}$ is the inverse of the cdf of the univariate standard normal distribution. That is, the obtained z-score leaves the p-value unchanged. The z-scores related to K subjects can be combined to a *grand-average* z-score $\overline{\mathrm{z}} = (\mathrm{z}_1 + \mathrm{z}_2 + \cdots \mathrm{z}_K)/\sqrt{K}$, which is also standard normally distributed. A grand-average p-value for the entire population of subjects can be derived from $\overline{\mathrm{z}}$ using $\mathrm{cdf_z}$.

In this thesis, an effect is considered significant if $\mathrm{p} \leq 0.05$, which corresponds to $|\mathrm{z}| \geq 2$. However, when many statistical tests are carried out simultaneously (e. g., when multiple hypotheses are tested in exploratory analyses), the probability of obtaining spurious significant results just by chance is high. In this case, p-values must be *corrected*. The most simple way to do this is the Bonferroni procedure (Bonferroni, 1936). The Bonferroni-corrected p-value is the original p-value multiplied by the number of tests (and thresholded to 1).

2.5 Spectral decompositions

2.5.1 The discrete Fourier transform

Electroencephalographic data typically contains several rhythmic components with narrow-band frequency signatures, which are believed to be linked to "idling" of certain brain functional networks (see Section 2.1). In order to extract these oscillations and to distinguish them from artifacts, it is helpful to work with a spectral representation of the data. This can be achieved by means of the *discrete Fourier transform* (DFT), which decomposes the data into a sum of sinusoids of differing frequencies and variable delays. The DFT *spectrum* of a finite discrete univariate time series $x(t)$, $t = 1, \ldots, T$ is given by

$$\widetilde{x}(f) = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} x(t) \exp\left(-\tfrac{2\pi i}{T} f t\right) . \tag{2.12}$$

The function value $\widetilde{x}(f)$ is a complex number, the modulus of which encodes the strength of the oscillation at frequency f, whereas its phase is the delay of that wave relative to the start of the time series. The DFT may be evaluated for any real-valued frequency f but it is sufficient to evaluate all integers between 0 and $T-1$ in order to have a complete representation of the signal. To see this consider that the DFT is a linear operation that can be written as a matrix multiplication $\widetilde{\mathbf{x}} = F\mathbf{x}$, where $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_T)^\top$ and $\mathbf{x} = (x_1, \ldots, x_T)^\top$. For frequencies $f = 0, \ldots, T-1$, F is a unitary $T \times T$ matrix. This implies that the transformation preserves energy, i. e., $\|\widetilde{\mathbf{x}}\|_2 = \|\mathbf{x}\|_2$ and that the inverse transformation exists and is defined by $F^{-1} = F^H$.

The empirical *cross-spectrum* S of two time series x_i and x_j is defined by

$$s_{i,j}(f) = \tfrac{1}{K} \sum_{k=1}^{K} \widetilde{x}_{i,k}(f)^* \widetilde{x}_{j,k}(f) , \tag{2.13}$$

which is a mean taken over K repeated simultaneous measuremens of $x_i(f)$ and $x_j(f)$. Diagonal entries $s_{i,i}(f)$ of the cross-spectrum correspond to the mean power of $x_i(f)$, while for $i \neq j$, $s_{i,j}(f)$ is a general complex number, the amplitude of which is an estimate of the combined strength of the two signals and the phase of which is an estimate of their phase difference.

2.5.2 The autoregressive model

The linear autoregressive (AR) process is the most simple model for the dynamics of a discrete time series (Brockwell and Davis, 1998). It assumes that the present state of a time series can be approximated by a linear combination of its past P samples, i. e., $x(t) = \sum_{p=1}^{P} b(p)x(t-p) + \varepsilon(t)$, where, $\varepsilon(t)$ is noise and $b(p)$ are scalar coefficients describing the influence of $x(t-p)$ on $x(t)$. Autoregressive processes are driven by nonzero noise terms, which are also called *innovations*. Typically, the innovations are assumed to be Gaussian distributed but this is no general requirement (Haufe et al., 2010c; Hyvärinen et al., 2010). The *multivariate AR* (MVAR) model

$$\mathbf{x}(t) = \sum_{p=1}^{P} B(p)\mathbf{x}(t-p) + \boldsymbol{\varepsilon}(t) \tag{2.14}$$

extends the univariate model to multiple time series. Here, $B(p)$ are matrices, the off-diagonal parts $b_{i,j}(p)$, $i \neq j$ of which describe influences between different time series.

There exist numerous algorithms for estimating the parameters of AR models (Brockwell and Davis, 1998; Schlögl, 2006). If T is large enough, a maximum-likelihood approach is appropriate. The matrix $E = (\boldsymbol{\varepsilon}(P+1), \ldots, \boldsymbol{\varepsilon}(T))$ of innovations can be expressed as $E = X - \widetilde{B}\widetilde{X}$, with $\widetilde{B} = (B(1), \ldots, B(P))$, $X = (\mathbf{x}(P+1), \ldots, \mathbf{x}(T))$, $\widetilde{X} = (X_1, \ldots, X_P)^\top$ and $X_p = (\mathbf{x}(P+1-p), \ldots, \mathbf{x}(T-p))^\top$. Assuming a Gaussian distribution for the innovations, the ML solution is obtained by ordinary least squares regression analogous to (2.6). The ML/OLS solution can be obtained iteratively for growing P, which is the idea of the ARFIT algorithm (Neumaier and Schneider, 2001). When T is small compared to M, it is appropriate to regularize, which could be done using any of the regularizers discussed in Section 2.3, and more. It is possible to estimate the frequency spectrum of the data by applying the Fourier transform to the vectors of estimated AR coefficients $(b_{i,j}(1), \ldots, b_{i,j}(P))^\top$. This is called the parametric approach to spectrum estimation as opposed to applying the Fourier transform to raw data.

A reasonable AR process must be *stable* in order to ensure stationary dynamics. To derive a condition for stability it is helpful to consider that an AR process of order P can be equivalently written as an order 1 process $\check{\mathbf{x}}(t) = \check{B}\check{\mathbf{x}}(t-1) + \check{\boldsymbol{\varepsilon}}(t)$ with

$$\check{B} = \begin{bmatrix} \widetilde{B} \\ I_{M(P-1)} \; 0 \end{bmatrix}, \tag{2.15}$$

$\check{\mathbf{x}}(t) = (\mathbf{x}(t)^\top, \ldots, \mathbf{x}(t-P+1)^\top)^\top$ and $\check{\varepsilon}_m(t) = 0$ for $m > M$. The system is stable if and only if the largest eigenvalue of $\check{B}$ is smaller than one. Otherwise, the innovation noise is amplified in each time step, leading to divergence. Stability must be ensured in AR estimation as well as for randomly drawn AR matrices that are used to generate artificial data.

2.5.3 Spectral filters

If signals of interest and noise have different frequency characteristics, it is beneficial to work only in the frequency range of the signal of interest. A *time-domain* representation of the signal in the frequency range of interest can be obtained by *spectral filtering*, or *bandpass filtering* if the frequency range is contiguous. A simple spectral filter can be implemented using the DFT by transforming the data into the frequency domain, setting the Fourier coefficients corresponding to undesired frequencies to zero and applying the inverse DFT. However, a DFT-based filter always uses all T measurements for estimating the filtered signal at time t. As one consequence, the DTF filter is not *causal*, that is, it uses information that is "in the future". For most types of EEG analysis, non-causal filters are either not suitable or not practical. An alternative are infinite impulse response (IIR) filters of the form $\widetilde{\widetilde{x}}(t) = \frac{1}{a_0}(\sum_{q=0}^{Q} b_q x(t-q) - \sum_{q=1}^{Q} a_q \widetilde{\widetilde{x}}(t-q))$ where $\widetilde{\widetilde{x}}(t)$ is the filtered signal, Q is the filter order, b_q and a_q are the filter coefficients and $\widetilde{\widetilde{x}}(t) = x(t) = 0$ for $t < 1$. There exist numerous ways of optimizing the filter coefficients for a particular frequency band of interest. The *Butterworth filter* is designed to damp all desired (*passband*) frequencies equally little while maximally suppressing the remaining (*stopband*) frequencies (Butterworth, 1930).

2.6 Measures of time-lagged effective connectivity

While there are multiple ways to define effective connectivity, the most widely accepted definition is based on a temporal argument: the cause must precede the effect. Respective approaches are often subsumed under the term *Granger-causal modeling*, although the technique that is known as *Granger causality* is only one way to measure time-lagged influence. In this section, we introduce popular measures of time-lagged effective connectivity, including various Granger causal approaches and the phase-slope index.

2.6.1 Granger causality based on model errors

Given a multivariate time series $\mathbf{x}(t)$, an influence of x_i on x_j according to the original Granger causality approach (Granger, 1969) is estimated as follows. A multivariate AR model is fitted for the full set $\mathbf{x}_{\{1,\dots,M\}} = \mathbf{x}$, as well as for the reduced set $\mathbf{x}_{\{1,\dots,M\}\setminus\{i\}}$ of available time series. Denoting the prediction errors of the full model by $\boldsymbol{\varepsilon}^{\text{full}}$ and those of the reduced model by $\boldsymbol{\varepsilon}^{\setminus i}$, the *Granger score* GC describing the influence of x_i on x_j is defined as the log-ratio of the mean-squared errors (MSE) of the two models with respect to x_j. I. e.,

$$\mathrm{gc}_{i,j} = \log\left(\frac{\sum_{t=P+1}^{T}\left[\varepsilon_j^{\text{full}}(t)\right]^2}{\sum_{t=P+1}^{T}\left[\varepsilon_j^{\setminus i}(t)\right]^2}\right). \tag{2.16}$$

Note that this definition, which is based on the ratio of prediction errors, is independent of the scale of the time series $x_m, 1 < m < M$. The Granger score defines a so-called *causality* or *effective connectivity graph*. Time series x_i is defined to *Granger-cause* time series x_j if $\mathrm{gc}_{i,j}$ is (significantly) greater than zero. The pairwise *net flow* is obtained by *antisymmetrizing* GC via $\mathrm{gc}^{\text{net}}_{i,j} = \mathrm{gc}_{i,j} - \mathrm{gc}_{j,i}$. Time series x_i is the *net driver* of time series x_j, if $\mathrm{gc}^{\text{net}}_{i,j}$ is (significantly) greater than zero, and the *net receiver*, if $\mathrm{gc}^{\text{net}}_{i,j}$ is (significantly) smaller than zero. These definitions apply analogously to other causal graphs defined in the following.

2.6.2 Granger causality based on AR-coefficients

One could argue that a Granger-causal dependence of time series x_j on time series x_i is already sufficiently evidenced if any of the P coefficients $b_{i,j}(p)$ of the MVAR model fitted on the full set of available time series is (significantly) different from zero. This is the basic consideration for a second class of Granger-causal approaches, which are defined directly based on the AR coefficients rather than prediction errors. While some of these methods operate in the time domain (Valdés-Sosa et al., 2005; Marinazzo et al., 2008), we here introduce two popular approaches assessing Granger-causal influence per frequency by evaluating the discrete Fourier transform of the MVAR coefficients across the time lag dimension (it is also possible to evaluate oscillations in between the DFT frequency bins). The *directed transfer function* (DTF, Kamiński and Blinowska, 1991) at frequency f is defined as

$$\mathrm{dtf}_{i,j}(f) = \frac{\left|\left(\widetilde{B}^{-1}\right)_{i,j}(f)\right|^2}{\sum_{m=1}^{M}\left|\left(\widetilde{B}^{-1}\right)_{j,m}(f)\right|^2}. \tag{2.17}$$

A related measure is *partial directed coherence* (PDC), which we define here as the squared absolute value of the original (complex-valued) quantity introduced in Baccalá and Sameshima (2001). That is,

$$\mathrm{pdc}_{i,j}(f) = \left| \frac{\bar{\bar{b}}_{i,j}(f)}{\sqrt{\bar{\bar{\mathbf{b}}}_j^H(f)\bar{\bar{\mathbf{b}}}_j(f)}} \right|^2 , \tag{2.18}$$

where $\bar{\bar{B}}(f) = I_M - \widetilde{B}(f)$ is an estimate of the strength of the information flow from x_i to x_j. Both DTF and PDC are normalized to take values from the interval $[0,1]$, but the normalization conditions differ. While DTF is normalized such that each time series has unit *inflow*, i. e., $\sum_i \mathrm{dtf}_{i,j}(f) = 1$, PDC is normalized such that each time series has unit *outflow*, i. e., $\sum_j \mathrm{pdc}_{i,j}(f) = 1$. Another difference between the two is, that DTF also reveals indirect influences not contained in non-vanishing $b(p)_{i,j}$ through the use of inverse MVAR matrices $\widetilde{B}^{-1}(f)$. It shares this property with the original Granger causality approach. Partial directed coherence, which operates directly on the (Fourier-transformed) AR matrices, only includes direct connections.

2.6.3 The phase-slope index

Another popular measure of interaction at a specific frequency is *coherency*, a generalization of correlation in the frequency domain (Nunez et al., 1997, 1999). Coherency (denoted by CHY) is a complex-valued measure describing the linear relationship of two time series at a specific frequency. It is defined as the normalized cross-spectrum

$$\mathrm{chy}_{i,j}(f) = \frac{s_{i,j}(f)}{\left(s_{i,i}(f)s_{j,j}(f)\right)^{1/2}} . \tag{2.19}$$

Coherence (denoted by CH) is the absolute value of coherency, i. e., $\mathrm{ch}_{i,j}(f) = |\mathrm{chy}_{i,j}(f)|$. It is often used to quantify the strength of functional connections. Since it is independent of the phase difference of the two oscillations, coherence makes no distinction between instantaneous (zero-lag) correlation and truly time-delayed (cross-) correlation. Instantaneous correlations, however, do not reflect time-lagged interactions and can occur for trivial reasons. Especially for EEG data, channels are highly correlated due to volume conduction in the head. These instantaneous correlations tend to dominate coherence. As a remedy, it has been proposed to look at the imaginary part of coherency only rather than at the absolute value (Nolte et al., 2004). This is motivated by the fact that the imaginary part of the cross-spectrum (and coherency) is zero if the phase difference (the phase of $s_{i,j}(f)$ rsp. $\mathrm{chy}_{i,j}(f)$) is zero. Thus, by looking at the imaginary part only, instantaneous effects are ignored.

In general, a positive imaginary part of $\mathrm{chy}_{i,j}(f)$ indicates that z_i is earlier than x_j and information appears to be flowing from x_i to x_j. However, "earlier" and "later" are ambiguous. For example at 10 Hz being 10 ms earlier cannot be distinguished from being 90 ms later. In order to resolve this ambiguity, the information at different frequencies can be aggregated within a frequency band of interest. The idea behind the phase-slope index (PSI, Nolte et al., 2008) is that the phase difference between sender and recipient increases linearly with frequency, i. e., the slope of the

phase spectrum is positive. Consequently, denoting by $\mathcal{F}$ a contiguous set of frequencies and by δf the frequency resolution, PSI is defined as

$$\psi_{i,j} = \Im\left(\sum_{f\in\mathcal{F}} \mathrm{chy}^*_{i,j}\,\mathrm{chy}_{i,j}(f+\delta f)\right). \tag{2.20}$$

Note that, in the original formulation, PSI is divided by its estimated standard deviation in order to assess statistical significance. The issue of statistical testing is, however, considered separately here. From the Hermitian property of the cross-spectral matrices it follows that PSI is antisymmetric with $\psi_{i,j} = -\psi_{j,i}$. Hence, PSI already measures net flows. Using the same property, it can be shown that PSI exactly flips its sign when being applied to temporally reversed data. Moreover, due to the use of the normalized cross-spectrum, PSI is invariant with respect to rescaling of the data.

2.7 Inverse source reconstruction

Recall that the general model of EEG data is $\mathbf{x}(t) = A\mathbf{s}(t) + \boldsymbol{\eta}(t)$ (see Section 2.3) with generally unknown source time series $\mathbf{s}(t)$ and mixing matrix A. In reality, the matrix A describes a physical process, namely the propagation of the brain electric currents from the source regions to the EEG electrodes. If a suitable physical model of the head exists, it can be used to compute the mixing patterns of idealized brain sources. This step is called *forward modeling* and leads to an estimate of A. *Inverse source reconstruction* is concerned with the estimation of $\mathbf{s}$ given $\mathbf{x}$ and A, which amounts to solving the so-called electromagnetic *inverse problem* (Baillet et al., 2001; Nunez and Srinivasan, 2006).

2.7.1 The EEG forward model

We here describe the steps needed to obtain the matrix A via forward modeling. The first step towards this is the definition of a model of the head as a volume conductor. We consider "realistic" models, which account for different conductivities of the various tissues, as well as arbitrarily-shaped tissue compartments. Precisely, the realistic model consists of three nested *shells* representing (from inner- to outermost shell) brain, skull and skin. Within each shell, homogeneous electric conductivity is assumed. The geometry of the three shells is acquired from anatomical *magnetic resonance* (MR) images of an individual head. These are segmented into the three compartments based on the grayscale value. The boundaries of the three compartments are triangularized and stored as 3D meshes.

The EEG potential at time t (omitting the time index in the following) is a scalar function $x : \mathcal{S} \to \mathbb{R}$ describing the potential difference between a point $\mathbf{v} \in \mathcal{S} \subset \mathbb{R}^3$ on the *border of the skin shell* $\mathcal{S}$ and a *reference* point $\mathbf{v}_{\mathrm{ref}} \in \mathcal{S}$. In practice, a set of M electrodes (plus one reference) is used, the locations $\mathbf{v}_{\mathrm{ref}}, \mathbf{v}_1, \ldots, \mathbf{v}_M$ of which are known and remain constant during the recording. The observable EEG potential may therefore be summarized as $\mathbf{x} = (x_1, \ldots, x_M)^\top$ with $x_m = x(\mathbf{v}_m)$. The reference electrode is commonly placed on the nose or the linked mastoids. For EEG electrodes, the standard positioning scheme is the *10-20 system*, which places electrodes along geodesic lines with 10 (20) degrees offset relative to inion, nasion and mastoid reference points. The original 10-20 system (Klem et al., 1999) defines $M = 19$ electrode positions, but it has been extended

to up to 256 electrodes (Sharbrough et al., 1991; Oostenveld and Praamstra, 2001). Standard electrode positions can be easily registered within a head model by, for example, using chemical indicators that mark the reference points in the anatomical MR images. If electrodes are placed in nonstandard positions, it is useful to acquire their exact coordinates using 3D tracking hardware.

The primary cerebral *current density* (at time t) is a vector-valued function (*vector-field*) $\mathbf{s} : \mathcal{B} \to \mathbb{R}^3$, which describes the primary electrical current at each location $\mathbf{u} \in \mathcal{B} \subset \mathbb{R}^3$ inside the brain shell. The *forward mapping* describes the functional dependence of the EEG electric potentials (due to secondary currents) on the primary current density. For frequencies below 1 kHz, all involved current flows are Ohmic, and the *quasi-static approximation of Maxwell's equations* holds (Sarvas, 1987; Baillet et al., 2001). This leads to

$$\mathbf{x} = \int_{\mathbf{u}\in\mathcal{B}} \mathrm{a}\left(\mathbf{u}\right) \mathbf{s}\left(\mathbf{u}\right) \mathrm{d}\mathbf{u} \;, \tag{2.21}$$

where the *electric lead field* $\mathrm{a} : \mathcal{B} \to \mathbb{R}^M \times \mathbb{R}^3$ is a general nonlinear function that is specific to the volume conductor and the location of reference and EEG electrodes. The tuple $(\mathbf{u}, \mathbf{s}(\mathbf{u}))$ defines a so-called *dipole*, which is an idealized electrical source of infinitesimal spatial extent at position $\mathbf{u}$ with current moment vector $\mathbf{s}(\mathbf{u})$. The EEG electric potential caused by the single dipole $(\mathbf{u}, \mathbf{s}(\mathbf{u}))$ is given by $\mathrm{a}(\mathbf{u})\mathbf{s}(\mathbf{u})$. While $\mathrm{a}(\mathbf{u})$ is generally hard to compute, it can be evaluated for certain types of volume conductors. In spherical head models, the solution is analytic (Baillet et al., 2001). For the realistic model discussed here the method of Nolte and Dassios (2005) provides a numerical solution based on semi-analytic expansions of the lead fields. An interesting approach is provided by Stahlhut et al. (2010). By adapting the head geometry to empirically recorded data using Bayesian reasoning, their approach potentially obviates the need for individual head models.

If the integral in (2.21) is replaced by a finite sum, a discretization of the current density is obtained, which leads to a computable global approximation of the EEG forward mapping. There are two major strategies for identifying the parameters of the involved dipolar sources. *Dipole fits* jointly estimate location $\mathbf{u}$ and current moment parameters $\mathbf{s}(\mathbf{u})$ of a small number of dipoles, while *distributed inverses* estimate only the moments of a large number of dipoles with fixed locations. While there also exist methods that estimate the location of the sources based on subspace criteria (Schmidt, 1986; Mosher and Leahy, 1999) or adaptive spatial filters (Van Veen and Buckley, 1988; Van Veen et al., 1997; Sekihara et al., 2005), we focus on dipole fits and distributed source imaging in the following.

2.7.2 Dipole fits

The N dipole model assumes that the EEG electric potential is generated by N pointlike activities, while the largest part of the brain is electrically silent (Scherg and Ebersole, 1993; Baillet et al., 2001). Denoting by $\mathbf{u}_n$ and $\mathbf{s}_n \equiv \mathbf{s}(\mathbf{u}_n)$ the parameters of the n-th dipole, the approximate forward mapping reads $\mathbf{x} = \sum_{n=1}^{N} \mathrm{a}(\mathbf{u}_n)\mathbf{s}_n$. If $6N$ is small compared to M, the parameters can be found by maximum-likelihood. Assuming uncorrelated Gaussian sensor noise, the associated cost function reads

$$\min_{\mathbf{u}_n,\mathbf{s}_n} \left\| \mathbf{x} - \sum_{n=1}^{N} \mathrm{a}\left(\mathbf{u}_n\right) \mathbf{s}_n \right\|_2^2 . \tag{2.22}$$

For time series, the N dipole model is easily extended to $\mathbf{x}(t) = \sum_{n=1}^{N} \mathrm{a}(\mathbf{u}_n)\mathbf{s}_n(t)$, which leads to a similar cost function. Note that in this model the dipole moments depend on t, since they describe the neuronal dynamics. The dipole locations are assumed to be constant, which encodes the assumption that the active brain areas do not move over time, which is reasonable regarding a potential subsequent effective connectivity analysis of the source brain regions. A local minimum of (2.22) can be found by means of nonlinear optimization. However, due to the nonconvexity of this function, it cannot be guaranteed that the global minimum is found. Indeed, for $N > 2$, dipole fits tend to get trapped in local minima easily.

2.7.3 Distributed inverse imaging

Assuming constant locations $\mathbf{u}_1, \ldots, \mathbf{u}_N$, the discrete forward model becomes the linear function

$$\mathbf{x} = \sum_{n=1}^{N} A_n \mathbf{s}_n = A\mathbf{s} \,, \tag{2.23}$$

where $A_n = \mathrm{a}(\mathbf{u}_n)$, $\mathbf{s} = (\mathbf{s}_1^\top, \ldots, \mathbf{s}_N^\top)^\top$ and $A = (A_1, \ldots, A_N)$. The large matrix A is called the *lead field matrix*. Notably, it is this equation that motivates the basic linear model (2.2) of the EEG introduced in the beginning.

The idea behind *distributed inverse imaging* is to model dipolar sources at many locations within the brain (or alternatively, only in the cortical areas), and to estimate the activity at those locations jointly by inverting the linear system (2.23). In most cases, the dipoles are arranged in a grid, where each dipole represents the activity in a cubic *voxel*. The inter-voxel distance is denoted by h. The number of sources do not have to be specified a-priori. Rather, the local maxima of the current distribution are interpreted as the active (source) regions. Naturally, the number of voxels N is large (in the thousands) when h is small. As a result, (2.23) is highly underdetermined, so that the maximum-likelihood estimator is not uniquely defined. To illustrate how severe the lack of information is, consider that the space of exact solutions to (2.23) is $(3N - M)$-dimensional, while vectors that are not solutions span only an M-dimensional space. In a typical scenario with $M \sim 100$ and $N \sim 2\,000$, the solution space is thus orders of magnitude more high-dimensional. In this setting, it is crucial to regularize the maximum likelihood solution by introducing additional penalties. Regularization then serves three purposes. Besides resolving the ambiguity of the ML estimator and preventing overfitting, it can also be used to constrain the solution to be consistent with prior domain knowledge. Note that this latter point is particularly important for the solution of inverse problems, where the goal is not only to achieve good generalization performance but also to interpret the model parameters.

Historically, there exist two contradicting assumptions on the spatial distribution of brain sources: *smoothness* and *sparsity*. At least one of these assumptions is (directly or indirectly) encoded in virtually any distributed source imaging method. Smoothness of the current density is motivated by the argument that neighboring voxels are likely to be functionally related and hence to be co-activated. The sparsity assumption, on the other hand, is based on the argument that cognitive processing related to a specific task should only activate a small part of the brain. This does not imply an overall sparse current density in general, since there might be non-task-related activity (brain noise) in other parts of the brain. However, the argument holds for averaged data, in which brain noise cancels.

Low resolution tomography (LORETA)

Smoothness is explicitly enforced in the *low resolution tomography* (LORETA) estimate (Pascual-Marqui et al., 1994)

$$\widehat{\mathbf{s}}^{\mathrm{LOR}} = \arg\min_{s} \|\mathbf{x} - A\mathbf{s}\|_2^2 + \lambda \|(I_3 \otimes D^{\mathrm{LOR}})W\mathbf{s}\|_2^2 \tag{2.24}$$

by means of the $N \times N$ Laplace operator D^{LOR}, which measures the sum of the discrete second derivatives of the current density in all three spatial directions. The entries of D^{LOR} are given by

$$d_{i,j}^{\mathrm{LOR}} = \frac{1}{h^2} \begin{cases} -6 & \|\mathbf{u}_i - \mathbf{u}_j\|_2 = 0 \\ 1 & \|\mathbf{u}_i - \mathbf{u}_j\|_2 = h \\ 0 & \text{else .} \end{cases} \tag{2.25}$$

The matrix W is a full-rank depth compensation matrix, the purpose of which is explained below. The LORETA problem is an instance of Tikhonov-regularization and can be solved analytically (see Section 2.3). The solution is the smoothest current density that explains the data to a certain extent (adjusted by λ). The Laplacian utilized by LORETA employs vanishing boundary conditions for the sake of invertibility, which leads to uniqueness of the LORETA solution. As a result, the activity towards the brain boundary as estimated by LORETA always approaches zero, which is a disadvantage given that superficial (i. e., cortical) sources are assumed to be the main generators of the EEG signal.

Pascual-Marqui et al. (1994) propose transforming the lead field matrix and the data into a *common average reference* before performing the inverse calculation, which is reasonable especially if the true position of the reference electrode and the position assumed in the head model do not match. Common-average-reference-transformed data and lead field matrices are obtained by $A \leftarrow HA$ and $\mathbf{x} \leftarrow H\mathbf{x}$, where the transformation matrix H is defined by $H = I_M - \mathbf{1}\mathbf{1}^\top / \mathbf{1}^\top\mathbf{1}$.

The weighted minimum-norm estimate

Interestingly, the *weighted minimum-norm* (WMN) estimate (Jeffs et al., 1987; Ioannides et al., 1990; Hämäläinen and Ilmoniemi, 1994)

$$\widehat{\mathbf{s}}^{\mathrm{WMN}} = \arg\min_{s} \|\mathbf{x} - A\mathbf{s}\|_2^2 + \lambda \|W\mathbf{s}\|_2^2 \tag{2.26}$$

tends to be very blurred (smooth), although no spatial filtering operator is involved. The reason for this "implicit" smoothness is that WMN (just as LORETA) is an instance of Tikhonov-regularization, which implies that its solution can be expressed as a linear combination of the EEG data (see Section 2.3). As a result, linear methods are characterized by a low spatial resolution of the source estimates, which means that the estimated current density can (at best) be a spatially lowpass-filtered version of the true source distribution (Grave de Peralta-Menendez and Gonzalez-Andino, 1998). The occurrence of side-lobes (sometimes called "ghost sources") is another phenomenon occuring with linear methods (Matsuura and Okabe, 1995).

The minimum-current estimate

Sparsity-inducing inverse methods promise spatial resolution in the range of the grid-size h, since their estimated active regions consist of single isolated voxels. The original sparse inverse solution is the weighted *minimum-current* (MC) estimate (Matsuura and Okabe, 1995)

$$\widehat{\mathbf{s}}^{\mathrm{MC}} = \arg\min_{s} \|\mathbf{x} - A\mathbf{s}\|_2^2 + \lambda \|W\mathbf{s}\|_1 \,. \tag{2.27}$$

Another well-known sparse inverse method is the FOCUSS (Gorodnitsky et al., 1995) approach. The FOCUSS solution is defined algorithmically as the iteratively reweighted ℓ_2-norm-regularized solution. It has been shown by Wipf and Nagarajan (2009) that this corresponds to (approximately) minimizing the ℓ_0-*quasinorm* of the sources.

Depth compensation

The matrix W that occurs in the cost functions of LORETA, WMN and MC is called a weighting or *depth-compensation* matrix, because its purpose is to counteract a location bias in the estimation. It is known that for $W = I_{3N}$ all these methods tend to estimate superficial sources as a result of the fact that the measurable electric potential falls off quadratically with the distance between source and sensor, which implies that deep sources must be stronger than superficial sources in order to reach similar explanatory power. Since, however, the regularization terms in (2.24), (2.26) and (2.27) effectively penalize the source strengths (as a side-effect of measuring smoothness or sparsity), solutions with all-superfical sources are often selected instead of deep (but smooth/sparse) sources. As a remedy, W is used to increase the cost of superficial sources. Often, W is chosen to be diagonal with entries $w_{i,i} = \|\mathbf{a}_i\|_2$ (Jeffs et al., 1987; Ioannides et al., 1990; Pascual-Marqui et al., 1994; Matsuura and Okabe, 1995).

Standardized LORETA

A depth-compensation is implicitly performed in the *standardized low resolution tomography* (sLORETA, Pascual-Marqui, 2002) approach, which employs post-hoc standardization of the conventional (unweighted) minimum-norm solution (i. e., the solution of (2.26) with $W = I_{3N}$). Using the common-average-reference-transformed lead field matrix A, Pascual-Marqui (2002) derive the covariance estimate

$$\widehat{V} = A^\top (AA^\top + \lambda I_M)^{-1} A \tag{2.28}$$

of the minimum-norm solution, which is used to compute the voxel-wise standardized source power estimate

$$\widehat{p}_n = (\widehat{\mathbf{s}}_n^{\mathrm{WMN}})^\top V_n^{-1} (\widehat{\mathbf{s}}_n^{\mathrm{WMN}}) \,, \tag{2.29}$$

where V_n is the n-th 3×3 block on the diagonal of $\widehat{V}$ (the intra-voxel covariance at the n-th voxel). It can be shown, that sLORETA has "zero location bias", in the sense that the estimated source power $\widehat{p}$ has a global maximum at $\widehat{p}_n$, if there is no noise and only one underlying dipolar source at voxel n (Pascual-Marqui, 2002).

2.8 Blind source separation

Inverse source reconstruction is a powerful approach, by which it is possible to reconstruct not only source activity but also respective source locations. However, it relies on the availability of an accurate forward model, which usually requires an *individual* head model in combination with exact electrode positions (although Stahlhut et al. (2010) devise an approach for adjusting the head model to the individual subject based on recorded data). Instead of using a possibly inaccurate forward model and carrying out its ill-posed inversion, the mixing matrix can also be estimated jointly with the source time series in a completely data-driven way. This approach is called *blind source separation* (BSS). Basically any matrix factorization can be regarded as a BSS technique. However, not all of them deliver meaningful estimates of the source and mixing matrices given the physical restrictions of the EEG signal generation; and only an even smaller fraction of methods is suitable for source connectivity analysis.

2.8.1 Principal component analysis

Principal component analysis (PCA) is an *eigendecomposition*

$$\Sigma = ADA^\top \tag{2.30}$$

of the empirical covariance matrix Σ of the data $\mathbf{x}(t)$, where $A \in \mathbb{R}^{M\times M}$ is orthogonal and $D \in \mathbb{R}^{M\times M}$ is diagonal with $d_{1,1} \geq d_{2,2}, \ldots, \geq d_{M,M}$. The columns of A are called *eigenvectors* in PCA terminology. In principle, they could be regarded as source mixing patterns, and the transformed time series $\mathbf{s}(t) = A^\top \mathbf{x}(t)$ could be interpreted as source time series, the variance of which is encoded in the diagonal entries (*eigenvalues*) of D. Notably, the time series $\mathbf{s}(t)$ are uncorrelated, which can be seen from the fact that their covariance matrix $A^\top \Sigma A = D$ is diagonal.

While uncorrelatedness of the source time series is reasonable to assume for EEG sources originating in different parts of the brain, orthogonality of the mixing patterns cannot generally be assumed for different brain sources. For that reason, PCA-filtered time series are not commonly interpreted as source estimates. However, PCA is a useful dimensionality-reducing preprocessing due to the fact that it provides uncorrelated components which are ordered by variance. Denoting by $A_\mathcal{K}$, $\mathcal{K} = \{1, \ldots, K\}$ the first K columns of A, (corresponding to the K largest eigenvalues), the projected data $A_\mathcal{K}^\top \mathbf{x}$ represents the full-rank data up to a reconstruction error of

$$\sum_{t=1}^{T} \|\mathbf{x}(t) - A_\mathcal{K} A_\mathcal{K}^\top \mathbf{x}\|^2 = \sum_{k=K+1}^{M} d_{k,k} \,, \tag{2.31}$$

and there is no K-dimensional subspace in which the data can be represented with lower error (Pearson, 1901). The question of optimally selecting the number of principal components is discussed in Hansen et al. (1999).

Note that a PCA decomposition of Σ is strongly related to a *singular value decomposition* (SVD) of the data matrix X. In particular, the eigenvectors of Σ are identical to the *singular vectors* of X, and the eigenvalues of Σ are the squared *singular values* of X. The concept of PCA/SVD has been generalized to *multi-way data* (tensors), which is helpful for decomposing time-frequency EEG data, or if there are multiple trials of EEG activity recorded under the same experimental condition (Mørup et al., 2006, 2008).

2.8.2 Independent component analysis

Diagonalization of covariance matrices as performed by PCA corresponds to decorrelation of the data. For multivariate Gaussian distributed data without time structure this is equivalent to removing any dependence between the source time series: the distribution of the sources factorizes into univariate Gaussians. For non-Gaussian source signals, however, decorrelation is not sufficient for achieving independence. Since, from a macroscopic perspective, many cognitive processes appear to occur independently of each other, it is therefore natural to search for "maximally independent" sources, which is the idea of *independent component analysis* (ICA, Molgedey and Schuster, 1994; Cardoso and Souloumiac, 1996; Belouchrani et al., 1997; Ziehe and Müller, 1998; Hyvärinen and Oja, 2000; Højen-Sørensen et al., 2002; Koldovský et al., 2006).

Statistical dependence cannot be measured as such, for which reason practical ICA implementations focus on minimizing certain aspects of dependence. Joint diagonalization of appropriate matrices is one way to achieve this. In the joint approximate diagonalization (JADE) approach (Cardoso and Souloumiac, 1996), slices of the fourth-order cumulant tensor are simultaneously diagonalized in order to extract source time series with distinct non-Gaussian distributions. Molgedey and Schuster (1994) employ second-order statistics (correlation) only but take time-lagged cross-correlation into account as an additional measure of temporal dependence. Their approach employs a generalized eigenvalue decomposition in order to simultaneously diagonalize the covariance matrix and a time-lagged cross-covariance matrix

$$C(p) = \frac{1}{2(T-1)} \sum_{t=p+1}^{T} \left(\mathbf{x}(t) - \widehat{\mu}_{\mathbf{x}}\right)\left(\mathbf{x}(t-p) - \widehat{\mu}_{\mathbf{x}}\right)^{\top} + \left(\mathbf{x}(t-p) - \widehat{\mu}_{\mathbf{x}}\right)\left(\mathbf{x}(t) - \widehat{\mu}_{\mathbf{x}}\right)^{\top} , \tag{2.32}$$

related to a single nonnegative integer time lag p. The second order blind identification (SOBI) approach (Belouchrani et al., 1997) and the related temporal decorrelation source separation (TDSEP) approach (Ziehe and Müller, 1998) extend this idea to the joint diagonalization of more than two symmetrized cross-correlation matrices related to multiple time lags p (including $p = 0$). In this case, exact joint diagonalization is not possible in general, for which reason a cost function is defined, which is minimized using nonlinear optimization (e. g., Ziehe et al., 2004). This cost is the sum of squares of the off-diagonal elements of the transformed matrices. In the case of TDSEP the optimization criterion reads

$$\begin{aligned} \widehat{A} \quad &= \quad \arg\min_{A} \sum_{p=0}^{P} \left\| \mathrm{off}\left(A^{-1} C(p) A^{-T}\right) \right\|_2^2 \\ &\quad \text{s. t.} \quad |A| = 1 \, . \end{aligned} \tag{2.33}$$

Independent component analysis has been successfully applied to the demixing of EEG signals (Makeig et al., 1997; Calhoun et al., 2001; Eichele et al., 2008) and artifact reduction (Jung et al., 2000a,b; Ziehe et al., 2000; McKeown et al., 2003; Tangermann et al., 2009; Winkler et al., 2011). Some authors have also used ICA in source synchronization and connectivity analyses, where sources are actually assumed to be dependent (Beckmann et al., 2005; Meinecke et al., 2005; Astolfi et al., 2007).

2.8.3 Pairwise interacting sources analysis

Nolte et al. (2006) remark that antisymmetric parts of cross-correlation matrices cannot be explained by independent sources and therefore relate to genuine interaction. Consequently, they devise a method called pairwise interacting sources analysis (PISA), which decomposes the sensor data into pairs of interacting sources by jointly diagonalizing *antisymmetrized* cross-correlation matrices

$$D(p) = \frac{1}{2(T-1)} \sum_{t=p+1}^{T} \left(\mathbf{x}(t) - \widehat{\mu}_{\mathbf{x}}\right) \left(\mathbf{x}(t-p) - \widehat{\mu}_{\mathbf{x}}\right)^{\top} - \left(\mathbf{x}(t-p) - \widehat{\mu}_{\mathbf{x}}\right) \left(\mathbf{x}(t) - \widehat{\mu}_{\mathbf{x}}\right)^{\top} . \tag{2.34}$$

It turns out that this is only possible using a complex-valued matrix A^{-1}, the columns of which contain the interacting sources' field patterns in their real and imaginary parts. However, since the joint diagonalization operation is invariant with respect to phase transformations, the sender's and receiver's patterns are arbitrarily mixed within these two-dimensional subspaces and have to be disentangled using additional assumptions, for example, on spatial properties of their inverse source reconstructions (Marzetti et al., 2008; Nolte et al., 2009).

Instead of diagonalizing antisymmetrized cross-correlation matrices, it is also possible to diagonalize their respective Fourier-transforms, which are identical to the imaginary parts of the cross-spectrum (Nolte et al., 2006). Thus, PISA can be regarded as a BSS technique that identifies source pairs exhibiting genuine time-lagged interaction. Since the temporal delay between the estimated sources might, however, be different in each frequency bin, PISA is more suitable for studying functional (undirected) than effective (directed) connectivity.

2.8.4 Convolutive ICA

The concept of BSS has been extended to the so-called *blind deconvolution* setting

$$\mathbf{x}(t) = \sum_{p=0}^{P} A(p)\mathbf{s}(t-p) + \boldsymbol{\varepsilon}(t) , \tag{2.35}$$

where the observation sequence is a spatio-temporally-filtered version of source signals (Attias and Schreiner, 1998; Parra and Spence, 2000; Anemüller et al., 2003; Dyrholm et al., 2007). Methods that estimate $A(p)$ and $\mathbf{s}$ under the assumption of mutual independence of the convolutional sources $\mathbf{s}$ are called *convolutive ICA* (CICA) approaches. The convolutive ICA with an auto-regressive inverse model approach (CICAAR, Dyrholm et al., 2007) estimates the parameters $A(p)$ under the assumption of non-Gaussian distributed source signals $\mathbf{s}(t)$ and no noise. For $K = M$ (the case $K < M$ can be modeled, too) this leads to the maximum likelihood solution

$$\mathcal{L}^{\text{CICAAR}}(\{A(p)\}) = \frac{T}{2} \log |A(0)^{\top} A(0)| - \sum_{t=1}^{T} \log p(\widehat{\mathbf{s}}(t)) \tag{2.36}$$

$$\widehat{\mathbf{s}}(t) = A^{-1}(0) \left(\mathbf{x}(t) - \sum_{p=1}^{P} A(p) \widehat{\mathbf{s}}(t-p) \right) , \tag{2.37}$$

where in practice the convolutive sources are modeled using the super-Gaussian hyperbolic secant (sech) distribution as $p(\mathbf{s}(t)) = \prod_{m=1}^{M} \operatorname{sech}(s_m(t))$.

The relevance of the convolutive ICA model for EEG data is not directly apparent when one considers that the propagation of secondary brain electrical currents is purely instantaneous. However, we show in Chapter 5 that the variables $\mathbf{s}$ of the CICAAR model have an interpretation as innovations of a source MVAR process (where "sources" refers to the instantaneously demixed time series as in (2.2)), for which an independence assumption is reasonable. Hence, CICAAR can be used to assess source effective connectivity.

2.8.5 MVARICA

The multivariate autoregression + independent component analysis (MVARICA) approach by Gómez-Herrero et al. (2008) is a blind source separation technique that is especially tailored to the analysis of source effective connectivity. It assumes the model $\mathbf{x}(t) = A\mathbf{s}(t)$ and a multivariate AR model $\mathbf{s}(t) = \sum_{p=1}^{P} B(p)\mathbf{s}(t-p) + \boldsymbol{\varepsilon}(t)$ for the sources. This implies that the EEG time series follows a multivariate AR model

$$\mathbf{x}(t) = \sum_{p=1}^{P} AB(p)A^{-1}\mathbf{x}(t-p) + A\boldsymbol{\varepsilon}(t) \tag{2.38}$$

with transformed coefficients $\widetilde{B}(p) = AB(p)A^{-1}$ and correlated noise $\widetilde{\boldsymbol{\varepsilon}}(t) = A\boldsymbol{\varepsilon}(t)$. The MVAR model is fitted using the ARFIT algorithm (Neumaier and Schneider, 2001) under the assumption that the innovation terms are Gaussian distributed. The remaining mutual dependence of the innovations is removed in a subsequent ICA step on the innovation time series, which is performed using efficient fast independent component analysis (EFICA, Koldovský et al., 2006). This leads to an estimate of the mixing matrix A, which is used to obtain the underlying sources' time series $\mathbf{s}(t) = A^{-1}\mathbf{x}(t)$ and the source MVAR coefficients $B(p) = A^{-1}\widetilde{B}(p)A$. The latter can be used to perform source connectivity analysis using, e. g., the directed transfer function or partial directed coherence.

3 Reconstruction of simulated source connectivity using existing approaches

Electroencephalographic recordings have been widely used in neuroscience to estimate brain (functional or effective) connectivity (e. g., Kamiński et al., 1997; Babiloni et al., 2004; Astolfi et al., 2004; Babiloni et al., 2005; Silberstein, 2006; Srinivasan et al., 2007; Supp et al., 2007; Blinowska et al., 2010). However, to our knowledge, there has been no simulation study so far that could approve the general eligibility of EEG-based connectivity analysis. Neither has the performance of different popular connectivity measures and preprocessing schemes been evaluated in a unified way. Efforts in that direction have been carried out in Astolfi et al. (2006b), where partial directed coherence is applied to simulated data. Similarly, the phase-slope index has been compared to Granger causality on simulated data under (mixed) noise influence (Nolte et al., 2008, 2010). However, the results of these studies do not transfer to the EEG case, since none of the respective simulation settings includes *source mixing* as caused by volume conduction.

In this chapter, we assess existing measures of effective connectivity in terms of their ability to infer source interactions from realistic EEG recordings under various conditions and after various common preprocessing steps. The connectivity measures we consider are Granger-causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI). A fourth measure, the directed transfer function (DTF), has been found to perform in a similar way as PDC in all the simulations carried out here (in fact, it can be shown that both are equivalent for $M = 2$), for which reason we omit reporting on DTF separately. We focus on the most simple model that includes source interaction, namely a two-dipole model with linear dynamics and a time-delayed linear influence of one source on the other. To ensure comparability, we use the same sources throughout the various experiments conducted in this study.

The results obtained in each of the experiments are underpinned by statistical analysis. These statistics are merely related to the findings as such and do not indicate the correctness of the findings, since it is an open question to quantify correctness of a connectivity estimate under source mixing. For certain source-space approaches such as blind source separation techniques, it is possible to quantify the success of both the demixing and the source connectivity analysis steps separately, if the numbers of true and simulated sources coincide. This is the approach pursued in Section 3.10 and in an exhaustive numerical evaluation of BSS approaches presented in Section 5.2. Regarding sensor-space analyses, we are not aware of any objective measure that could be used to judge the quality of a connectivity estimate. The same holds for inverse source reconstruction approaches. For this reason, we feel that it is important to provide a qualitative assessment of such analyses in terms of the plots that would normally be interpreted by neurophysiologists. To increase comprehensibility, we do not restrict ourselves here to the analysis of a subset of electrodes but utilize so-called *head-in-head plots* (see Section 3.2), which are capable of visualizing full pairwise connectivity graphs in a manner that allows one to link the observed results directly to the known topographical field patterns of the underlying simulated sources.

The study consists of ten experiments. In Section 3.1, we apply measures of effective connectivity to simulated unmixed source time series. Section 3.2 deals with realistic EEG comprising volume conduction and noise. Section 3.3 demonstrates the effect of data normalization on PDC. In Section 3.4, we devise a test that allows one to judge whether estimated connections are related to time-lagged interaction. Section 3.5 investigates whether the problem of biased PDC estimates can be overcome by a permutation test using reshuffled data. Sections 3.6 and 3.7 demonstrate the influence of the choice of the reference electrode and the signal-to-noise ratio (SNR) on connectivity estimates, respectively. The last three experimental sections evaluate the performance of existing EEG demixings as preprocessing steps for connectivity estimation. In particular, we test the effectiveness of the scalp Laplacian (Section 3.8), linear inverse source reconstruction (Section 3.9) and various blind source separation techniques (Section 3.10). As part of our discussion in Section 3.11, we motivate the development of new algorithms for demixing connected sources. The results presented in this chapter are part of an upcoming publication.

3.1 Experiment 1: two interacting sources

Setting

We start with assessing effective connectivity analysis methodologies on the source level, where no instantaneous mixing of signals due to volume conduction takes place. We simulate a system with only two sources $s_{1/2}(t)$ following a bivariate AR process of order $P = 5$, from which we generate $T = 10\,000$ source samples. The MVAR coefficients are sampled independently from $\mathcal{N} \sim (0, 0.01)$. By setting the off-diagonal coefficients $b_{1,2}(p), 1 \leq p \leq P$ to zero, while all other coefficients remain nonzero, unidirectional flow from s_1 to s_2 is modeled. The sampling process is repeated until a stable system is obtained. The innovations $\boldsymbol{\varepsilon}(t)$ of the source AR process are drawn from the univariate standard normal distribution. We perform 100 repetitions of the experiment. For each repetition, we generate a dataset comprising distinct innovation terms and source AR coefficients. Note that, since in practice source time series cannot be observed, the present experiment serves mainly as a proof-of-concept for effective connectivity measures and a baseline for source demixing algorithms. We consider the noiseless case here, while the influence of noise has been investigated in the literature (see below).

We apply Granger causality, partial directed coherence and the phase-slope index to the source time series. The phase-slope index is computed using an implementation provided by Nolte et al. (2008)[1], while the "Granger Causal Connectivity Analysis" toolbox (Seth, 2010)[2] is used to compute GC, and the MVARICA toolbox (Gómez-Herrero et al., 2008, code not anymore publicly available) is used to compute PDC. The AR model underlying the computation of PDC is estimated using the ARFIT package (Neumaier and Schneider, 2001)[3]. Since the source time series are generated using time-domain MVAR models, interactions affect all frequency bins, for which reason information flow is estimated here using the complete frequency spectrum of the data. For PDC, that means that we average the respective frequency-wise scores to obtain a global measure of interaction. A two-sided one-sample t-test is performed to assess whether the estimated net information flow

[1] http://ml.cs.tu-berlin.de/causality
[2] http://www.informatics.sussex.ac.uk/users/anils/aks_code.htm
[3] http://www.gps.caltech.edu/~tapio/arfit

according to each of the connectivity measures significantly differs from zero in one or the other direction. For reporting the results, t-scores are converted into z-scores as outlined in Section 2.4.

Results

All connectivity measures correctly indicate significant net information flow from s_1 to s_2 with z-scores exceeding the significance level by far ($z > 11$). This is not surprising given that the generating MVAR model introduces a time-lagged influence of s_1 on s_2, which is exactly the assumption of all three measures. Moreover, the dynamical model underlying GC and PDC is exactly the linear multivariate AR model, so that correct connectivity estimation here boils down to the problem of correct AR estimation. Note that, when the source activity is superimposed by noise, the results can be quite different from those obtained here. Nolte et al. (2008) consider unmixed sources in the presence of correlated mixed noise and discover that Granger causality tends to yield spurious interaction if the SNR is low. Moreover, Nolte et al. (2010) show that this behaviour occurs even with temporally white noise and that the effect depends on the autocorrelation of the sender's time series.

3.2 Experiment 2: realistic EEG

Setting

In this experiment, we consider realistic simulated EEG comprising volume conduction and noise. We assess connectivity directly between sensor-space signals, which is common practice in the literature (e. g., Kamiński et al., 1997; Hesse et al., 2003; Blinowska et al., 2010). The pseudo-EEG signal is generated according to

$$\mathbf{x}(t) = \frac{(1-\gamma)\left(\frac{1}{2}\sum_{i=1}^{2}\frac{\mathbf{a}_i s_i(t)}{\|(S^T)_i\|_2}\right)}{\|\,\mathbf{vec}\,(S)\,\|_2} + \gamma\frac{\boldsymbol{\eta}(t)}{\|\,\mathbf{vec}\,(E)\,\|_2}\,, \tag{3.1}$$

where $\mathbf{x}$ is the EEG signal, $s_{1/2}$ are the source time series, the generation of which is described in the previous section, $\mathbf{a}_{1/2}$ are the spread patterns of the dipolar sources evaluated at 59 EEG electrode positions, $\boldsymbol{\eta}$ is noise and γ, $0 \le \gamma \le 1$ is a parameter that adjusts the SNR. Moreover, $E = (\boldsymbol{\eta}(1), \ldots, \boldsymbol{\eta}(T))$ and $S = (\mathbf{s}(1), \ldots, \mathbf{s}(T))$. The normalizing terms $\|(S^\top)_i\|_2$ are used to equalize the power of driver and receiver time series, while the normalization by $\|\,\mathbf{vec}\,(S)\,\|_2$ and $\|\,\mathbf{vec}\,(E)\,\|_2$, respectively, allows precise adjustment of the SNR by means of γ. Here, we set $\gamma = 0.5$, corresponding to a balanced SNR.

We use a head model with realistically-shaped brain, skull and skin shells (Holmes et al., 1998). Moreover, we assume standard electrode positions as defined in the extended 10-20 system, and a nose reference. The source dipoles are placed in the left and right hemispheres of the brain, 3 cm below C3 (s_1) and C4 (s_2), respectively. The moment vectors of both dipoles are chosen to be tangentially oriented, leading to bipolar field patterns, which do not overlap much. This scenario can be regarded as a simulation of active sources in the left and right sensorimotor cortices, where information flows from the left to the right cortex.

The field patterns describing the spread of the source dipoles to the EEG sensors are computed according to Nolte and Dassios (2005). Both the dipolar sources and their corresponding EEG field patterns are depicted in Figure 3.2. These field patterns are indeed similar to those obtained from electrical stimulation of the Medianus nerve at the left and right hand, respectively (see Figure 4.4 (a)). We consider a set of ten designated electrodes, for which we report results explicitly. These comprises three frontal electrodes (F3, Fz and F4), three central electrodes (C3, Cz and C4), three parietal electrodes (P3, Pz and P4) and one occipital electrode (Oz). Results on Oz are only reported in Section 3.9. Three of the electrodes are located in the left hemisphere (F3, C3 and P3), three in the right hemisphere (F4, C4 and P4) and four over the central sulcus (Fz, Cz, Pz and Oz). The positions of these electrodes are marked by circles in Figure 3.2 and their relative arrangement is shown in Figure 3.1. Notably, the scalp electrical activity at the electrodes under which the sources are located (C3 and C4) is close to zero, owing to the bipolar structure of the field patterns. The maximal signal deflection of the patterns is instead achieved at the F3 and F4 electrodes, respectively, while the minimal (but maximal in absolute terms) deflection is measured at P3 and P4.

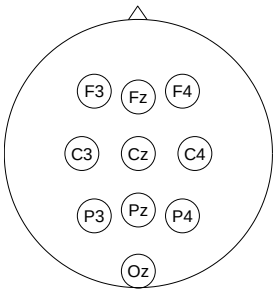


Figure 3.1: Spatial arrangement of ten representative electrodes.

The noise terms $\boldsymbol{\eta}(t) = \boldsymbol{\eta}^{\text{sensor}} + A^{\text{biol.}}\boldsymbol{\eta}^{\text{biol.}}$ are composed of *sensor noise* $\boldsymbol{\eta}^{\text{sensor}}$, which is drawn independently for each sensor and time point from a standard normal distribution. Furthermore, we include *biological noise*, representing cerebral background activity. We compute time courses of ten noise sources $\eta_1^{\text{biol.}}(t), \ldots, \eta_{10}^{\text{biol.}}(t)$ using random stable univariate AR models of order ten. These are mixed into the pseudo-EEG using a spread matrix $A^{\text{biol.}}$ related to ten randomly placed dipoles with random current moment vectors. Thus, the biological noise is correlated in sensor space but causes no time-lagged dependencies (interactions). Sensor and biological noises are scaled to contribute equally to the overall noise $\boldsymbol{\eta}$.

We generate 100 pseudo-EEG datasets, in which the underlying source dipoles and their spread patterns are kept constant, whereas distinct innovations of source and noise AR processes, as well as distinct source and noise AR coefficients and distinct noise dipole locations are drawn. A z-score is computed for each entry of the sensor-space connectivity matrices that are estimated by GC, PDC and PSI, respectively.

Results

The connectivity graphs estimated by GC, PDC and PSI are visualized in Figure 3.3 (a) using *head-in-head plots* (Nolte et al., 2004, 2008). Each head-in-head plot is composed of 19 small circles representing the human scalp. These are arranged within one large scalp plot according to the positions of the 19 electrodes of the original 10-20 electrode placement system. Each of the small scalp plots thereby shows the estimated interaction of the respective electrode to all 58 other electrodes, where red and yellow colors ($z > 0$) stand for information outflow and blue and cyan colors ($z < 0$) stand for information inflow. Bonferroni correction is used to account for multiple testing. The Bonferroni factor is determined to be 551, which is the number of electrode pairs $(19 \cdot (59 - 1)/2)$ visualized. The corresponding Bonferroni-corrected significance threshold

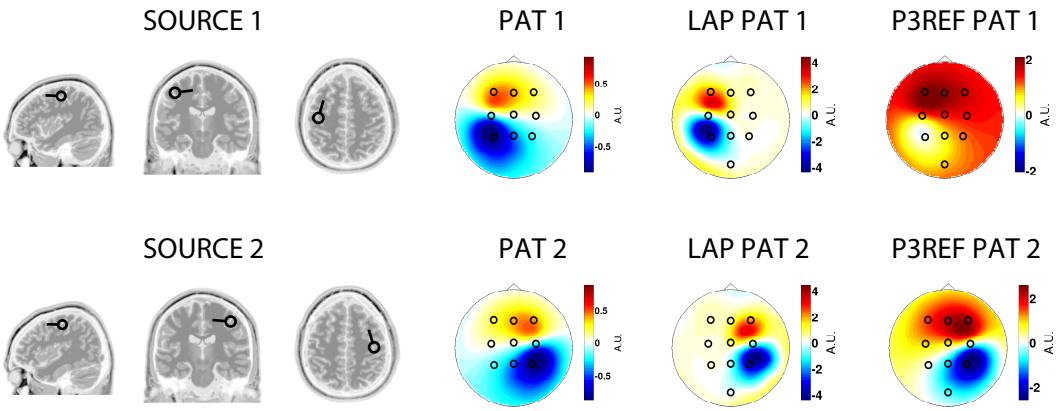


Figure 3.2: Two simulated dipolar sources (SOURCE 1/2) and their corresponding EEG field patterns (PAT 1/2). Sources are placed 3 cm below the C_3 (left) and C_4 (right) electrodes and are oriented tangentially to the scalp. Also shown are field patterns after Laplace transformation (LAP PAT 1/2) and after changing the reference to the P_3 electrode (P3REF PAT 1/2).

is $z = 4.0$ and is indicated by a thin black line in the colorbar, while the z-score corresponding to an uncorrected p-value of $p = 0.05$ ($z = 2.0$) is indicated by a thick black line.

Granger causality yields rather noisy connectivity patterns (see Figure 3.3) with no significant ($z > 4.0$) interactions between the nine frontal, central and parietal electrodes. We here report also connections that do not surpass Bonferroni correction but are otherwise significant (i. e., $z > 2.0$). The respective flows are measured from F_3, P_3 and P_4 to Fz, P_3 and P_4 to Pz, C_3 to C_4 and Cz to P_4. Hence, we observe mainly symmetric flow from lateralized electrodes, where the underlying sources are strongest, to central electrodes, where they are weakest. This is in line with an observation made by Nolte et al. (2008), where it is stated that Granger causality is affected by *asymmetries* in the noise distributions as caused, for example, by unequal signal-to-noise ratios. Notably, GC hardly makes any distinction between the driving and the receiving hemisphere. While it does indicate flow from C_3 to C_4, it is questionable whether this reflects the true flow between the underlying sources when one considers that their field patterns have almost zero activity at C_3 and C_4, respectively.

Partial directed coherence yields significant ($z > 4.0$) information flow from F_3, C_3, Fz, Cz, F_4 and C_4 to P_3, from F_3, C_3, Fz, Cz, F_4, C_4 to P_4, from C_3, Cz and C_4 to Fz, from F_3, C_3, Fz, Cz, F_4 and C_4 to Pz, from C_3 and Cz to F_3 and from Cz to F_4. That is, PDC indicates gross symmetric bilateral information exchange from electrodes with low SNR to electrodes with high SNR, regardless of whether the electrode belongs to the "driving" or "receiving" hemisphere. We identify this behaviour of PDC to be related to asymmetries in the scaling of the data rather than the SNR. In fact, we observe that PDC estimates significant interaction even on temporally and spatially white noise time series, which are differently scaled. This bias is caused by the fact that (the maximum-likelihood estimate of) the MVAR coefficients modeling the flow from low- to high-amplitude electrodes need to have larger amplitudes than those modeling the opposite direction, even though the data are completely random.

If there is no noise at all, the maximum-likelihood estimators of the MVAR models underlying PDC are not uniquely defined. The result of this method then additionally depends on regularizing assumptions made in the MVAR estimation. We observe that in this scenario the ARFIT estimate yields results vastly differing from, for example, the MVAR coefficients with minimum ℓ_2-norm.

PSI correctly reveals information flow from the left to the right hemisphere, where the respective connectivity pattern resembles the true field patterns of the underlying sources. That is, almost all electrodes in the right hemisphere receive information from almost all electrodes in the left hemisphere, while the most significant flow passes from those regions in which the driving source is most strongly expressed to those in which the receiving source is most strongly expressed. In particular, PSI estimates significant ($z > 4.0$) flow from F_3, C_3, P_3, Fz and Pz to F_4, from F_3, P_3 and Fz to C_4, from F_3, C_3, P_3, Fz and Pz to P_4, from F_3 and P_3 to Fz and from F_3 and P_3 to Pz. Generally, we can derive the empirical rule that (if driving and receiving sources are similarly strong and noise sources contribute equally much to all sensors) PSI roughly varies with $\log(|a_{1,i}||a_{2,j}|/|a_{2,i}||a_{1,j}|)$, which we call the *driver-receiver ratio* related to a pair of electrodes (i, j) and a pair of driving and receiving sources. Assuming that s_1 is the driver and s_2 is the receiver, PSI estimates flow from channel x_i to channel x_j if the driver-receiver ratio is positive and flow from x_j to x_i if the driver-receiver ratio is negative.

3.3 Experiment 3: normalization

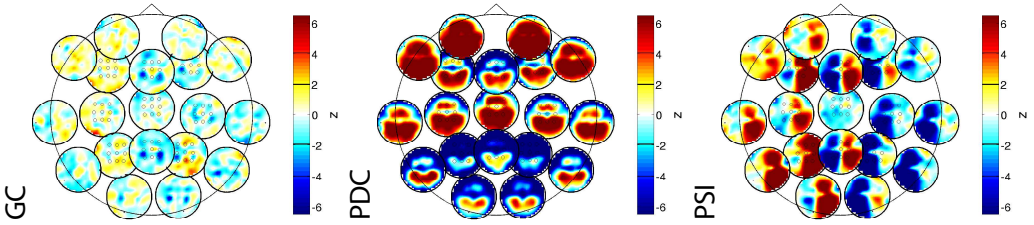
Setting

Having demonstrated PDC's dependence on the scale of the signals, we now study *normalized* EEG data $x_m(t) \leftarrow (x_m(t) - \mu_{x_m})/\sigma_{x_m},\ m = 1, \ldots, M$. We apply the normalization to the same data that is used in Section 3.2.

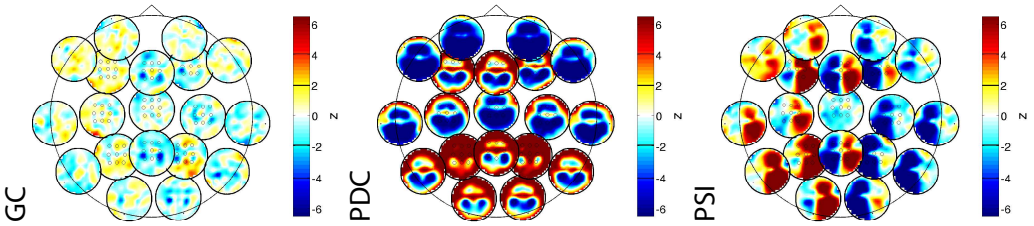
Results

Granger causality and the phase-slope index are invariant with respect to the scaling of the data (see Section 2.6), for which reason the connectivity maps related to both measures are identical to those related to the original data (c. f. Figure 3.3 (b) and Figure 3.3 (a)). In contrast, the flow according to PDC roughly *reverses* using normalized time series, indicating a further asymmetry in the MVAR

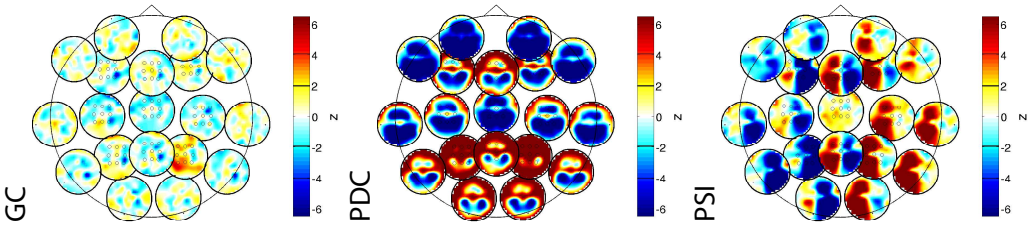
Figure 3.3 *(following page)*: Comparison of effective connectivity of simulated EEG as estimated by Granger causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI). Two source dipoles with tangential orientations are modeled 3 cm below the $C_3/4$ electrodes. Information flow from the left (C_3) to the right (C_4) source is modeled by means of a bivariate AR model. The simulated EEG is superimposed by non-interacting biological and sensor noise (SNR = 1). The significance of estimated interactions is measured in terms of z-scores and visualized as head-in-head plots, where red and yellow colors ($z > 0$) stand for information outflow and blue and cyan colors ($z < 0$) stand for information inflow. The Bonferroni-corrected significance level is indicated by a thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line. (a) Analysis of "raw" EEG data. (b) Analysis of *normalized* data. (c) Analysis of *normalized time-reversed* data. (d) Analysis of *normalized time-reshuffled* data. (e) Analysis using a two-sample statistical test for *differences between normalized original and time-reshuffled* data.



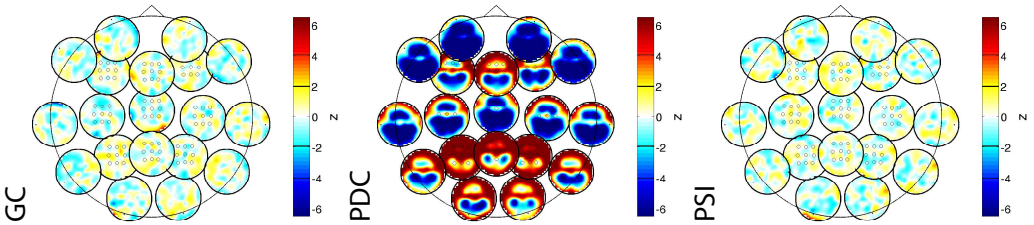
(a) Original



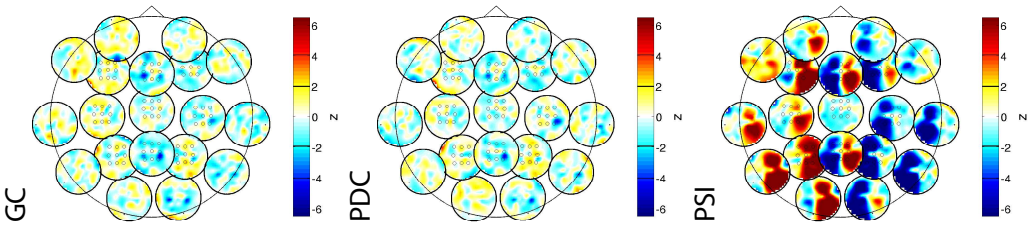
(b) Normalized



(c) Normalized time-reversed



(d) Normalized time-reshuffled



(e) Normalized original vs. time-reshuffled (two-sample test)

coefficients that was previously masked by the asymmetry introduced by the signal scale. To be precise, the direction of all significant connections reverses, except for new connections from F3 and F4 to C4, and from P3 and P4 to Pz, as well as the lack of connections from F3 to C3, and Pz to F3, Fz and F4. While we could not reproduce these effects on purely (temporally and spatially) white data, it also occurs on time-reversed as well as temporally-reshuffled data (see Sections 3.4 and 3.5). Hence, we conclude that the spurious effects we observe here are related to differences in the channels' signal-to-noise ratios and the concomitant differences in inter-channel correlations. As we note later, this should be distinguished from the influence the SNR has on Granger causality, which appears only if the data is in correct temporal order.

3.4 Experiment 4: time reversal

Setting

While we have identified a number of artifactual confounders affecting Granger-type connectivity measures in the previous sections, we are now concerned with the problem of assessing whether such confounders are the predominant reason for a given detected connection. Precisely, we introduce a simple test to assess whether an estimated interaction is truly time-lagged. According to this idea, assuming that the driver precedes the receiver, the driver-receiver relationship reverses if the temporal order of the time series is reversed. Hence, if significant flow is estimated on the original data, its direction (sign) should ideally flip under time inversion, while it is also acceptable if the estimated flow in the "wrong" direction completely vanishes. If, however, original and time-reversed signals yield the same flow direction, the effect cannot be related to a measurable time-lag of the interaction and artifactual confounders are the likely reason. We apply our test on time-reversed versions of the normalized data used in the previous section.

Results

The connectivity patterns obtained on time-reversed data are shown in Figure 3.3 (c). Due to the Hermitian property of the cross-spectrum, PSI exactly changes its sign under time inversion (see Section 2.6), which is the desired behaviour. For GC, no significant connections are found. Connections associated with z-scores greater than two are found to pass from P4 to F3, C3, P3 and Pz, from F4 to C3 and C4, and from C4 to P4. That is, the original normalized and the time-reversed data share a presumed connection from P4 to Pz, while no connection reverses after time-reversal. The remaining connections estimated by GC on time-reversed data do not occur (in neither direction) on the original normalized data. Hence, the connection from P4 to Pz is unlikely to relate to time-lagged interaction, while the outcome of the test is undefined for all other connections. Partial directed coherence applied to time-reversed data shares all connections with the result obtained from original data, except for a missing connection from F3 to C4 and three additional connections from F3 and F4 to C3 and From Pz to F4. These results indicate that most effects found by PDC are artifacts not related to time-lagged interaction.

3.5 Experiment 5: permutation testing

Setting

In Kamiński et al. (2001) it is suggested to use permutation tests for assessing the significance of connections estimated by partial directed coherence. This is reasonable, since it may be the case that a two-sample test for differences between results obtained on original and randomly permuted time series cancels out artifactual effects while revealing smaller effects that are possibly related to genuine information flow. In this experiment, we adopt the idea of a permutation test. Using the generation scheme of the normalized dataset described in Section 3.3, we create an additional version in which the data are temporally permuted. Electrode-wise differences between the two datasets are assessed by means of a two-sample t-test and transformed into z-scores. For completeness, a one-sample statistical test is also performed on the reshuffled dataset.

Results

The one-sample test does not yield any significant connection with associated z-scores greater than four for GC and PSI, while z scores exceeding two are found for very few connections (from P_3 to C_3 and Cz to P_4 for GC and from P_3 to C_3 and C_4 to Fz for PSI, see Figure 3.3 (d)). The lack of estimated interaction is the desired behaviour given that the temporal structure of the data is destroyed by the reshuffling. The analysis of PDC on this data, however, yields the same picture as on non-shuffled data except for missing connections from F_4 to Cz, P_3 to Pz and F_4 to C_4, plus additional connections from F_3 to C_3 and from Pz to F_3, Fz and F_4. That is, PDC estimates gross significant information flow from high-SNR to low-SNR sensors even in the absence of any temporal structure in the data.

When comparing original and reshuffled data using the two-sample test (see Figure 3.3 (e)), GC yields connections ($z > 2.0$) from F_3 and P_3 to Fz, and from C_4 and P_4 to Pz, none of which is significant after Bonferroni correction. This means that there is almost no difference between one-sample and two-sample testing for this method (c.f. the results in Section 3.3). The same holds for PSI, which yields the same significant connections ($z > 4.0$) as in one-sample testing, except for two missing connections from C_3 to F_4 and from Fz to C_4. In contrast, PDC yields no significant connections anymore. This is explained by the dominance of artifactual confounders both in the original and the reshuffled data, which cancel out when comparing the two. Probable interactions ($z > 2.0$) are detected from F_3 to Fz and from P_4 to Pz and C_4. That is, as a result of using two-sample testing the PDC estimate is similar to the GC estimate, with which it shares two out of four connections. In sum, appears that PDC suffers in two ways from unequal signal-to-noise ratios in the data. While the larger effect, an asymmetry in the MVAR coefficients caused by unequal degrees of correlation in EEG channel pairs, can be eliminated by permutation testing, the smaller effect related to the fact that high-SNR channels are better suited for predicting the future of low-SNR channels than vice-versa, remains.

3.6 Experiment 6: influence of the reference electrode

Setting

The results obtained so far in sensor space indicate that effective connectivity measures based on the idea of Granger causality suffer from a number of estimation biases, which prevent them from recovering the true inter-hemispheric information flow in our simulated example. PSI, in contrast, reveals that flow. However, the correct interpretion of head-in-head plots is difficult even when PSI is used. In Section 3.2, we noted that PSI designates two electrodes as driver and receiver, respectively, if the corresponding driver-receiver ratio is large enough. Unfortunately, this ratio depends on how much signal the reference electrode picks up from the underlying driving and receiving sources. In this experiment, we construct an example in which the change of reference drastically affects the driver-receiver ratio. While all previous simulations were carried out using a nose reference, we now consider data that is re-referenced by subtracting the activity of the P3 electrode. Note that this choice is particularly harmful, since the driving source is most strongly expressed in P3 (see Figure 3.2). We consider the original EEG data, which are described Section 3.2. After re-referenciation, these data are normalized and subjected to effective connectivity analysis. As in all experiments except for the previous one, we use a one-sample test for statistical evaluation.

Results

Figure 3.4 (a) shows sensor-space connectivity maps obtained on re-referenced data. Granger causality does not yield significant ($z > 4.0$) information flow, whereas there are three probable connections (from F3 to Fz and F4, and from P4 to F4) with z scores greater than two. The maps related to PDC and PSI both differ substantially from those obtained from original normalized data (see Figure 3.3 (b)). For PDC, left parietal regions are designated to receive information from all other regions ($z > 4.0$), which is incorrect but understandable since PDC on normalized data is biased towards estimating flow from high-SNR to low-SNR channels. Precisely, PDC estimates connections from F3, Fz, Cz, F4, C4 and P4 to C3, from F3, C3, Fz, Cz, Pz, F4, C4 and P4 to P3, from F3, Fz and F4 to Cz, from F3, C3, Fz, Cz, F4, C4 and P4 to Pz, from F3 to F4, from F3, Fz and F4 to C4, and from F3, Fz and F4 to P4. Regarding PSI, the left frontal region is estimated as the only driving region, while no significant interaction is estimated for the left posterior region. Significant connections ($z > 4.0$) are estimated from F3 to Fz, from F3, Cz and C4 to Pz, from F3, Cz and C4 to F4, and from F3, C3, Fz, Cz, Pz and C4 to P4. For understanding this behaviour of PSI, it is helpful to inspect the re-referenced field patterns of the interacting sources, which are depicted in Figure 3.2. According to these, the sending source is much weaker than before in the region around P3, which causes changes in the global driver-receiver ratio.

In line with our findings, a problematic influence of the choice of the reference electrode on sensor-space coherency estimates has been noted in Nunez et al. (1997) and Marzetti et al. (2007). Marzetti et al. propose the *standardized infinity reference* as a potential mitigation, which is, however, only available in a spherical head model. Moreover, there exist a number of ways to transform the data into a *reference free* representation (see Sections 3.8–3.10).

3.7 Experiment 7: influence of the SNR

Setting

Whether an estimated effective connection is significant or not naturally depends on the overall signal-to-noise ratio, which is adjusted by γ. While this is somehow trivial, we, however, also observe that results obtained at different SNR's may lead to different conclusions regarding the size and location of the underlying interacting brain regions, which we consider an issue worth mentioning. In this experiment, we study the influence of low ($\gamma = 0.25$) and high ($\gamma = 0.75$) SNR. We consider normalized data generated as in Section 3.2.

Results

Connectivity estimates for $\gamma = 0.25$ and $\gamma = 0.75$ are depicted in Figure 3.4 (b) and (c), while the case $\gamma = 0.5$ is depicted in Figure 3.3 (b). Granger causality is relatively unaffected by the SNR in that, regardless of the choice of γ, the estimated connections are not significant after Bonferroni correction. Connections corresponding to z-scores greater than two are noted from C_3 to F_3, Cz to P_3, F_3 to Fz, C_3 to Cz, P_3 to Pz, F_3 to F_4, and from F_3, P_3 and Cz to P_4 for $\gamma = 0.25$, which is similar to the results obtained for $\gamma = 0.5$. Interestingly, there are fewer connections with z-scores greater than two for $\gamma = 0.75$ (that is, at high SNR). The respective connections pass from C_3 and P_4 to F_3, from C_3 and P_4 to Fz, from C_3 and P_4 to Pz, and from P_4 to C_4. Regarding PDC and PSI, we note that, the higher γ is, the broader the regions are in which the estimated information flow is significant with $z > 4.0$. In contrast to the case $\gamma = 0.5$, PDC misses all connections except those from P_3 and P_4 to Fz for $\gamma = 0.25$. For $\gamma = 0.75$, additional connections are estimated from F_3 and F_4 to C_3, from C_3 and C_4 to Cz, and from Pz to F_4, while the connection from P_3 to Pz is missing. Similarly, for PSI, additional connections are estimated from P_3 to F_3, from C_3 to Fz, from F_3, C_3 and P_3 to Cz, from C_3 to Pz and C_4, from Pz to C_4, and from F_4 to P_4 for $\gamma = 0.75$, while the connection from P_3 to P_4 is the only one retained for $\gamma = 0.25$. Importantly, the analysis of data generated with $\gamma = 0.25$ suggests smaller interacting brain areas than it is the case for $\gamma = 0.5$, since the upper lobes of the two bipolar field patterns do not reach significance, which might lead to a totally different assessment of the number of underlying sources and their location for $\gamma = 0.25$. Analogously, for $\gamma = 0.75$ one might come to the conclusion that larger areas within both hemispheres are interacting, although, in fact, the underlying simulated sources are point sources without spatial extent.

3.8 Experiment 8: Laplace filtering

Setting

Having assessed the possibility of performing connectivity analysis in sensor space, we now turn to investigating preprocessing steps that promise to provide estimates of the actual underlying sources. We start with examining the discrete Laplace operator, which is defined as the sum of the second spatial derivatives of the scalp electrical potential. The Laplacian is known to locally enhance the signal-to-noise ratio, which is beneficial, e. g., for brain-computer interfacing classification (Sannelli et al., 2011). Several studies investigate the use of scalp Laplacians in the context of

coherence analysis and come to the conclusion that Laplace spatial filtering reasonably supresses volume conduction (Srinivasan et al., 1998, 2007; Babiloni et al., 2004). Since the Laplacian is a spatial highpass filter, Laplace-transformed EEG data is sometimes regarded as an estimate of the source activity directly below the electrodes. Notably, the Laplacian is independent of the choice of the reference electrode and thereby provides a standardized way of analyzing EEG datasets.

In this experiment, we apply Laplace spatial filtering as a preprocessing step to EEG-based effective connectivity analysis. We use an implementation called *current source density* (CSD, Kayser and Tenke, 2006), which employs spherical spline interpolation of the scalp surface. We apply the Laplace filter to the original EEG time series described in Section 3.2. To restore the full rank of the data (Laplace filtering reduces the rank by one), we add a small amount of white Gaussian noise, which is several orders of magnitude smaller than the signal level. The resulting time series are then normalized and analyzed using GC, PDC and PSI.

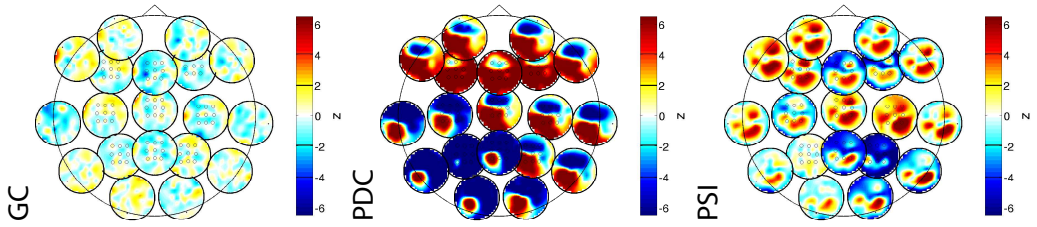
Results

The Laplace transformation does not affect Granger causality in the sense that no statistically significant (after Bonferroni correction) connections arise as a result of this preprocessing (see Figure 3.4 (d)). Connections with z-scores greater than two are noted to pass from C_3 to F_3, Fz and Cz to P_3, C_4 to Pz, Fz to F_4, and P_4 to C_4. For PDC and PSI, Laplace filtering leads to higher frequency structures in the sensor-space connectivity maps, compared to the maps obtained from the original normalized EEG data. However, the spurious effects noted previously are still prevalent in the PDC estimate (while having higher spatial complexity). In particular, we note significant ($z > 4.0$) connections from F_3, P_3, Fz, Pz, F_4 and P_4 to C_3, from F_3, P_3, F_4 and P_4 to Fz, from F_3, P_3, Fz, Pz, F_4 and P_4 to Cz, from F_3, P_3, F_4 and P_4 to Pz, and from F_3, P_3, Fz, Pz, F_4 and P_4 to C_4. For PSI, we note significant ($z > 4.0$) connections from P_3 to Fz, Cz, F_4, C_4 and P_4, from F_3 to F_4 and P_4, and from C_3, Fz and Cz to P_4. All results obtained on Laplace transformed normalized data are hence qualitatively similar to those obtained on the original normalized data. This is also apparent from the Laplace filtered field patterns of the interacting sources (see Figure 3.2), which exhibit slightly more focal active regions than the original patterns, but have the same bipolar structure. We thus conclude that Laplace filtering is not effective in demixing sources with tangential orientation. Note that, due to operating only in a neighborhood of electrodes, it is also ineffective in the case of low-frequent field patterns as induced by deep sources.

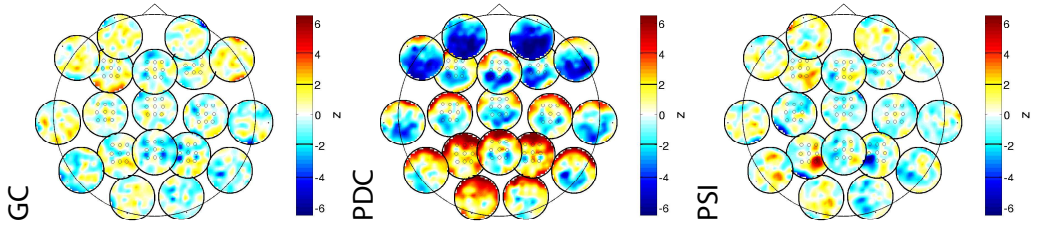
3.9 Experiment 9: linear inverse preprocessing

Setting

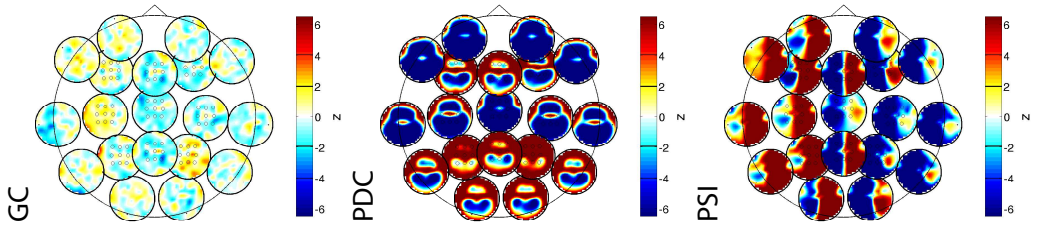
While Laplace filtering is not helpful if the underlying sources are deep and/or tangentially-oriented, this restriction does not apply to genuine solutions of the EEG inverse problem, which are also reference-free. Working on inverse solutions moreover allows one to study brain interaction directly in terms of the estimated signal-generating brain structures, which are otherwise hard to infer from sensor-space connectivity maps. And finally, inverse solutions are ideally unmixed, which may even justify Granger-causal analysis of the sources. Motivated by these considerations, a number of studies have investigated effective connectivity by means of Granger-causal analysis



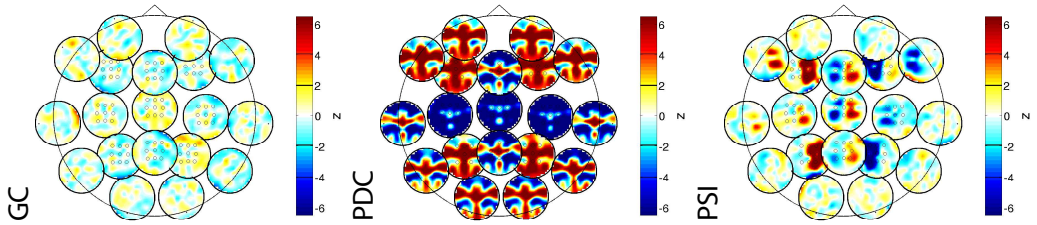
(a) P3 reference



(b) $\gamma = 0.25$



(c) $\gamma = 0.75$



(d) Laplace transformed

Figure 3.4: Comparison of effective connectivity of simulated EEG as estimated by Granger causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI). Two source dipoles with tangential orientations are modeled 3 cm below the $C_{3/4}$ electrodes. Information flow from the left (C_3) to the right (C_4) source is modeled by means of a bivariate AR model. The simulated EEG is superimposed by non-interacting biological and sensor noise (SNR = 1). The significance of estimated interactions is measured in terms of z-scores and visualized as head-in-head plots, where red and yellow colors ($z > 0$) stand for information outflow and blue and cyan colors ($z < 0$) stand for information inflow. The Bonferroni-corrected significance level is indicated by a thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line. (a) Analysis of *re-referenced normalized* data. (b/c) Analysis of *normalized* data at *different signal-to-noise ratios* (SNR). (d) Analysis of *Laplace-transformed normalized* data.

of cortically-constrained minimum-norm source estimates (e. g., Babiloni et al., 2005; Astolfi et al., 2006a; Gow et al., 2008). In this experiment, we follow a similar approach by applying the weighted-minimum norm estimate (WMN) to the original EEG data used in Section 3.2. The depth-compensation matrix W of WMN is chosen as in focal vectorfield reconstruction, which is introduced later (see Section 4.1). This choice has been shown to outperform the original suggestion (Jeffs et al., 1987; Matsuura and Okabe, 1995) in terms of localization accuracy (Haufe et al., 2008). The regularization parameter λ is chosen using 5-fold cross-validation, which is implemented by splitting the set of electrodes randomly into five folds. The same regularization parameter is selected for all time indices t. The localization is conducted in the same head model in which the data was simulated, that is, the "Montreal" head with nose reference. We define the interior of the whole brain shell as the source space, which is partitioned into $N = 2\,142$ voxels of 10 mm side length. In the center of each voxel, a dipolar source is modeled, the current moment vector of which is estimated for each time point using WMN. The resulting time courses are the basis of the subsequent connectivity assessment.

Estimating effective connectivity on inverse source estimates provided by a distributed linear inverse method poses three challenges. The first one refers to the fact that each dipolar brain source has three orientation parameters describing the amplitude of the current in each of the three spatial directions. As a consequence, no less than nine distinct combinations of source time series have to be considered when assessing the information flow between two dipoles. The second challenge is posed by the large number of dipoles that are required to sample the source space with sufficient spatial resolution. Computing the full connectivity matrix for all $3N = 6\,426$ time series is computationally expensive using PSI, while it is intractable for all Granger-causal methods. Hence, a dimensionality reduction must be performed prior to source connectivity estimation. The third challenge refers to the visualization of the effective connectivity graph, since it is not straightforward to visualize all pairwise interactions in source space in a similar way as it is possible in sensor space using head-in-head plots.

In order to reduce the dimensionality in source space, we define *regions-of-interest* (ROIs) similar to Babiloni et al. (2005) and Astolfi et al. (2006a). Since we are only concerned with simulated data, we do not define these regions anatomically but partition the source space according to the nearest (in the Euclidean sense) EEG electrode. Doing so enables us to present the results using the familiar head-in-head plots. Dipoles that are further than 5 cm away from any electrode are not assigned to any region. The source activity within each ROI is averaged separately for each spatial dimension to yield a $3M$-dimensional time series. Since this time series has at most rank M, a small amount of noise is added in order to establish full rank. The resulting time series are normalized and subjected to effective connectivity analysis. The application of each of the effective connectivity measures yields a $3M \times 3M$ connectivity matrix consisting of 3×3 blocks $R(i, j)$, which describe the interactions (net flows) between the i-th and the j-th dipole in all three spatial dimensions. We define the *total net flow* from the i-th to the j-th voxel as the sum over all entries of $R(i, j)$. This operation yields an antisymmetric $M \times M$ matrix, the entries of which are tested for being significantly different from zero using one-sample testing.

Results

The result of the WMN inverse source reconstruction is depicted in the upper left panel of Figure 3.5 as a heat map showing estimated dipole moment vector amplitudes (averaged over all time instants and repetitions). The plot is overlaid with arrows depicting the true interacting dipolar sources. Evidently, the source activity estimated by WMN is spread over the entire brain. The true sources are not well separated, as the source amplitude map exhibits only one local maximum, which is in between the two true dipoles. As a consequence, it is not possible to estimate source connectivity between "active" regions based on the WMN estimate. Rather, a ROI-based approach has to be taken, as pursued here. The mean amplitude of source activity in each ROI is depicted in the upper right panel of Figure 3.5 as a scalp map. While large portions of activity are concentrated under central electrodes around C_3 and C_4, there is also a lot of deep activity which is closer to occipital electrodes.

The estimated effective connectivity between ROIs is depicted in the lower panel of Figure 3.5 using head-in-head plots. Note that the interpretation of these plots here is not the same as in previous sections, since the depicted connectivity maps do not reflect interactions between electrodes but interactions between ROIs below these electrodes. Since information flow is not so meaningful if it is estimated between (almost) inactive regions, the information about the strength of the source activity is encoded in the visualization by means of the transparency value. Here, the ROI with maximal strength is drawn with full opacity, while 10 % opacity is used for the ROI with minimal strength. Transparency values for ROIs in between are assigned using a monotonous sigmoidal nonlinearity.

To account for the more diverse activity patterns of the WMN source estimates compared to the sensor-space field patterns, we extend the set of electrodes (and corresponding ROIs) for which we explicitly report z statistics by the occipital channel Oz, which is located posterior to all other channels (see Figures 3.2 and 3.1). We here report connections only for pairs of *active* ROIs, the combined strength (measured as the product of the mean source amplitudes) of which is at least 0.25 times as strong as the maximal strength measured for all combinations of ROIs.

The connectivity map according to PSI exhibits information flow from gross regions in the left hemisphere to gross regions in the right hemisphere. This involves the correct lateralized central regions in which strong source activity is estimated by WMN, but also an occipital area which has similarly strong presumed source activity. This latter region is estimated to transmit information from the left central area to the right central area, which is incorrect. Precisely, there are significant interactions ($z > 4.0$) from the regions below F_3, C_3 and P_3 to the region below Oz, from the region below C_3 to the regions below C_4 and P_4, and from the region below Oz to the regions below C_4 and P_4. Hence, it appears that WMN is not only incapable of localizing the true source regions accurately. It also only insufficiently demixes the sensor signals, which causes the spurious detection of interaction between active regions.

Regarding GC, we note that WMN linear inverse preprocessing does not lead to the detection of any notable ($z > 2.0$) flow between active ROIs. Partial directed coherence analysis on ROIs does not yield connections that are significant after Bonferroni correction, while we observe z scores greater than two related to connections from regions below F_3, C_3, P_3 and C_4 to the region below Oz. Note that there are no connections with z scores greater than two for GC and PDC, if two-sample instead of one-sample testing is used.

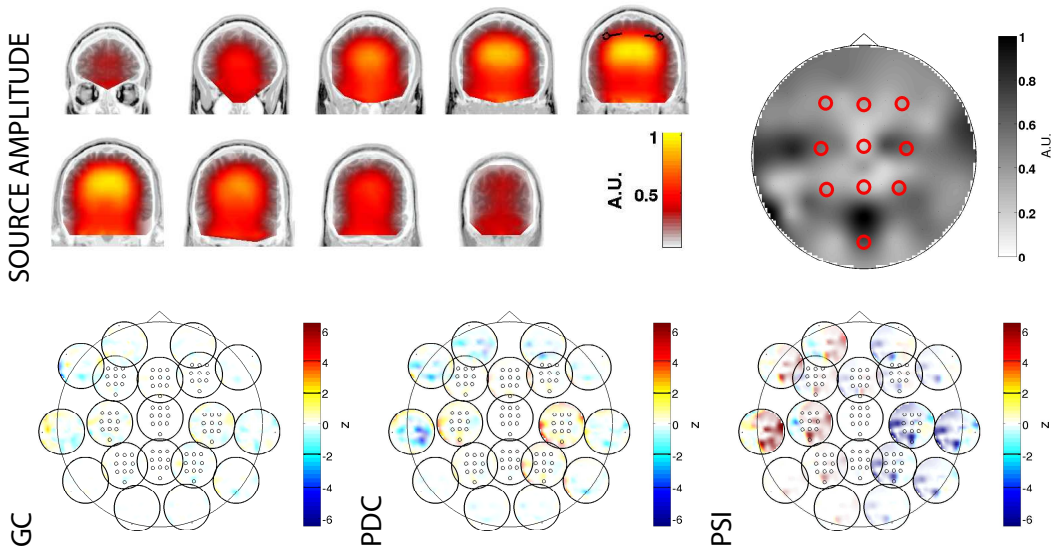


Figure 3.5: Comparison of effective connectivity of simulated EEG as estimated by Granger causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI) after application of weighted minimum norm (WMN) linear inverse source reconstruction. Two source dipoles with tangential orientations are modeled 3 cm below the $C_3/4$ electrodes. Information flow from the left (C_3) to the right (C_4) source is modeled by means of a bivariate AR process. The simulated EEG is superimposed by non-interacting biological and sensor noise (SNR = 1). Upper left panel: average source strength (dipole amplitude) per voxel. Upper right panel: average source strength in regions-of-interest associated with the closest EEG electrode. Lower panel: Connectivity between regions-of-interest in source space, which are defined based on the nearest EEG electrode. The significance of estimated interactions between regions is measured in terms of z-scores and visualized as head-in-head plots, where red and yellow colors ($z > 0$) stand for information outflow and blue and cyan colors ($z < 0$) stand for information inflow. The Bonferroni-corrected significance level is indicated by a thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line.

3.10 Experiment 10: blind source separation preprocessing

Setting

An alternative to source reconstruction based on a physical model is blind source separation (BSS), which amounts to demixing the sensor signals solely based on statistical assumptions. The aim of this experiment is to evaluate popular BSS techniques regarding their ability to demix EEG sources with mutual interaction structure, as those simulated here. We here consider four methods, namely MVARICA, which has been explicitly designed to study source effective connectivity, CICAAR, for which we note in Section 5.2 that it may be used for source connectivity analysis, and two variants of ICA, namely temporal decorrelation source separation (TDSEP) and joint approximate diagonalization (JADE). The latter are included here, because ICA has been frequently employed in EEG synchronization and connectivity studies (Beckmann et al., 2005; Meinecke et al., 2005;

Astolfi et al., 2007), although their independence assumption technically contradicts such analysis. We do not consider PISA here, since this method is tailored to detecting *functional* connectivity, while we are interested in also determining the direction of flow. Moreover, PISA itself provides only estimates of two-dimensional subspaces containing pairs of interacting sources, relying on additional post-processing steps to demix these source pairs.

The most popular BSS technique, principal component analysis (PCA), is a demixing into instantaneously decorrelated time series with orthogonal spatial patterns. Owing to the orthogonality constraint, PCA is unable to correctly estimate the mixing patterns in general, for which reason we use it here mainly to reduce the dimensionality of the data prior to an additional subsequent demixing. We apply all BSS methods to normalized source time series associated with the five strongest PCA components, where the number $K = 5$ is chosen to be lower than the known number of simulated interacting brain sources. We use implementations of MVARICA, CICAAR, TDSEP and JADE provided by the respective authors of Gómez-Herrero et al. (2008), Dyrholm et al. (2007)[4], Ziehe and Müller (1998)[5] and Cardoso and Souloumiac (1996)[6]. The joint diagonalization in TDSEP is carried out using fast Frobenius-norm joint diagonalization (FFDIAG, Ziehe et al., 2004)[7]. The number of time lags in TDSEP is chosen to be $\tau = 100$, while it is selected from $\{1, \ldots, 9\}$ for MVARICA and CICAAR using the Bayesian information criterion. We use the default parameters for JADE.

The result of applying BSS techniques is a decomposition of the data into five components consisting of a field pattern and a time series. The time series are normalized and analyzed using GC, PDC and PSI. Since only the outer product of a field pattern and a one-dimensional source time series defines a BSS component, there is a degree of freedom regarding the sign and scale of both factors (although normalization of the time series can be used to overcome the scale invariance). Moreover, the order of the BSS components is arbitrary. These indeterminacies make it difficult to match similar components obtained from differing decompositions, which is a requirement for subsequent statistical assessment. We here use the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957) to match BSS components obtained in the various repetitions of the experiment. This algorithm computes a pairing of two sets of vectors, which is optimal regarding a given measure of similarity between vectors. We use the implementation provided by Tichavský and Koldovský (2004)[8] to match the estimated field patterns. As a similarity measure for field patterns, we use the goodness-of-fit (GOF) achieved by linear OLS regression of one onto the other pattern. For two patterns $\mathbf{a}_1$ and $\mathbf{a}_2$, the optimal regression coefficient is $c(\mathbf{a}_1, \mathbf{a}_2) = \mathbf{a}_2^\top \mathbf{a}_1 / \|\mathbf{a}_2\|^2$ and the goodness-of-fit is $\mathrm{GOF}(\mathbf{a}_1, \mathbf{a}_2) = \|c(\mathbf{a}_1, \mathbf{a}_2)\mathbf{a}_2 - \mathbf{a}_1\| / \|\mathbf{a}_1\|$.

We obtain a global alignment of BSS components using the field patterns estimated in the first repetition of the experiment as a template, to which patterns obtained in later repetitions are transformed. That is, starting from the second repetition, the optimal pairing between the newly-obtained and the template field patterns is computed. Having found the pairing, the new field patterns are permuted and scaled to approximate the templates as well as possible using the optimal regression coefficients. The source connectivity graphs estimated by GC, PDC and PSI are

[4] http://www.machlea.com/mads/cicaar/index.html
[5] http://user.cs.tu-berlin.de/~ziehe/code/tdsep.zip
[6] http://perso.telecom-paristech.fr/~cardoso/Algo/Jade/jadeR.m
[7] http://user.cs.tu-berlin.de/~ziehe/code/ffdiag_pack.zip
[8] http://si.utia.cas.cz/downloadPT.htm

also permuted accordingly. We furthermore compute the goodness-of-fit related to all pairs of true and estimated field patterns. We identify the estimated components that achieve the best average GOF for each true component. These components are reported along with the respective GOF scores. Finally, one-sample testing is employed to assess whether there is significant information flow between the estimated source time series. Since there are ten possible pairwise interactions between five time series, the z-score corresponding to a Bonferroni-corrected p-value of $p = 0.05$ is given by $z = 2.8$.

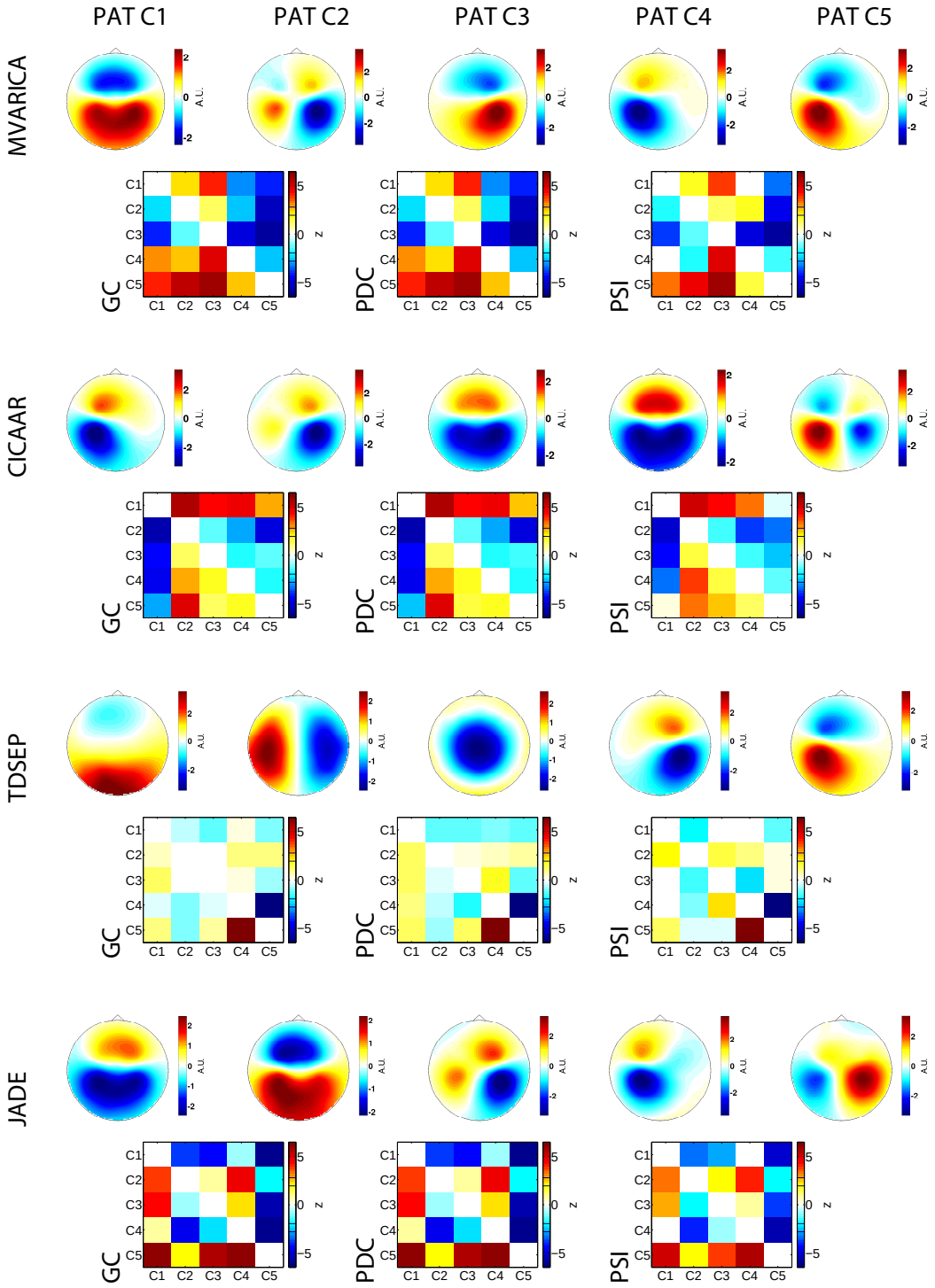
To investigate the importance of non-Gaussianity for blind source separation, we perform the analysis on the original data introduced in Section 3.2, as well as on new instances, in which the source MVAR innovations are drawn from the super-Gaussian hyperbolic secant distribution (Dyrholm et al., 2007) instead of the standard normal distribution.

Results

Figures 3.6 and 3.7 depict the source field patterns (averaged across repetitions) estimated by MVARICA, CICAAR, TDSEP and JADE, as well as the connectivity estimates calculated from the associated source time series for Gaussian (Figure 3.6) and non-Gaussian (Figure 3.7) innovations. The estimated connectivity graphs are presented as matrices, in which red and yellow colors ($z > 0$) in the intersection of the i-th row and the j-th column denote that there is net information flow from the i-th to the j-th source, while blue and cyan colors ($z < 0$) denote the opposite case.

In the case of Gaussian innovations, only TDSEP is able to reasonably recover the underlying interacting source components. Here, the pattern of component C_4 corresponds to the pattern of source in the right hemisphere, while that of component C_5 correspond to the pattern of source in the left hemisphere. The respective GOF scores are 0.71 ± 0.02 for C_4 and 0.73 ± 0.02 for C_5. The remaining components C_1, C_2 and C_3 reflect different sources, which reflect simulated biological noise. A significant ($z > 2.8$) flow from C_5 to C_4 is correctly estimated by all three measures of effective connectivity. Moreover, no additional significant connections are found by any measure. The remaining BSS methods MVARICA, CICAAR and JADE do not isolate the true two sources in single components but rather mix them to a varying degree into all sources. The best approximation of the source in the left hemisphere is achieved by component C_3 for MVARICA

Figure 3.6 *(following page)*: Comparison of effective connectivity of simulated EEG as estimated by Granger causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI) after application of blind source separation (BSS) preprocessing according to combined multivariate autoregressive estimation and independent component analysis (MVARICA), convolutive independent analysis via inverse autoregression (CICAAR), temporal decorrelation source separation (TDSEP) and joint approximate diagonalization (JADE). Two source dipoles with tangential orientations are modeled 3 cm below the $C_3/4$ electrodes. Information flow from the left (C_3) to the right (C_4) source is modeled by means of a bivariate AR process with *Gaussian* innovations. The simulated EEG is superimposed by non-interacting biological and sensor noise (SNR = 1). BSS techniques are applied to the five strongest PCA components. The significance of estimated interactions between demixed signals is measured in terms of z-scores and visualized as matrices, where entries with red and yellow colors ($z > 0$) stand for information outflow and entries with blue and cyan colors ($z < 0$) stand for information inflow of the source marked in the respective row. The Bonferroni-corrected significance level is indicated by a thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line.

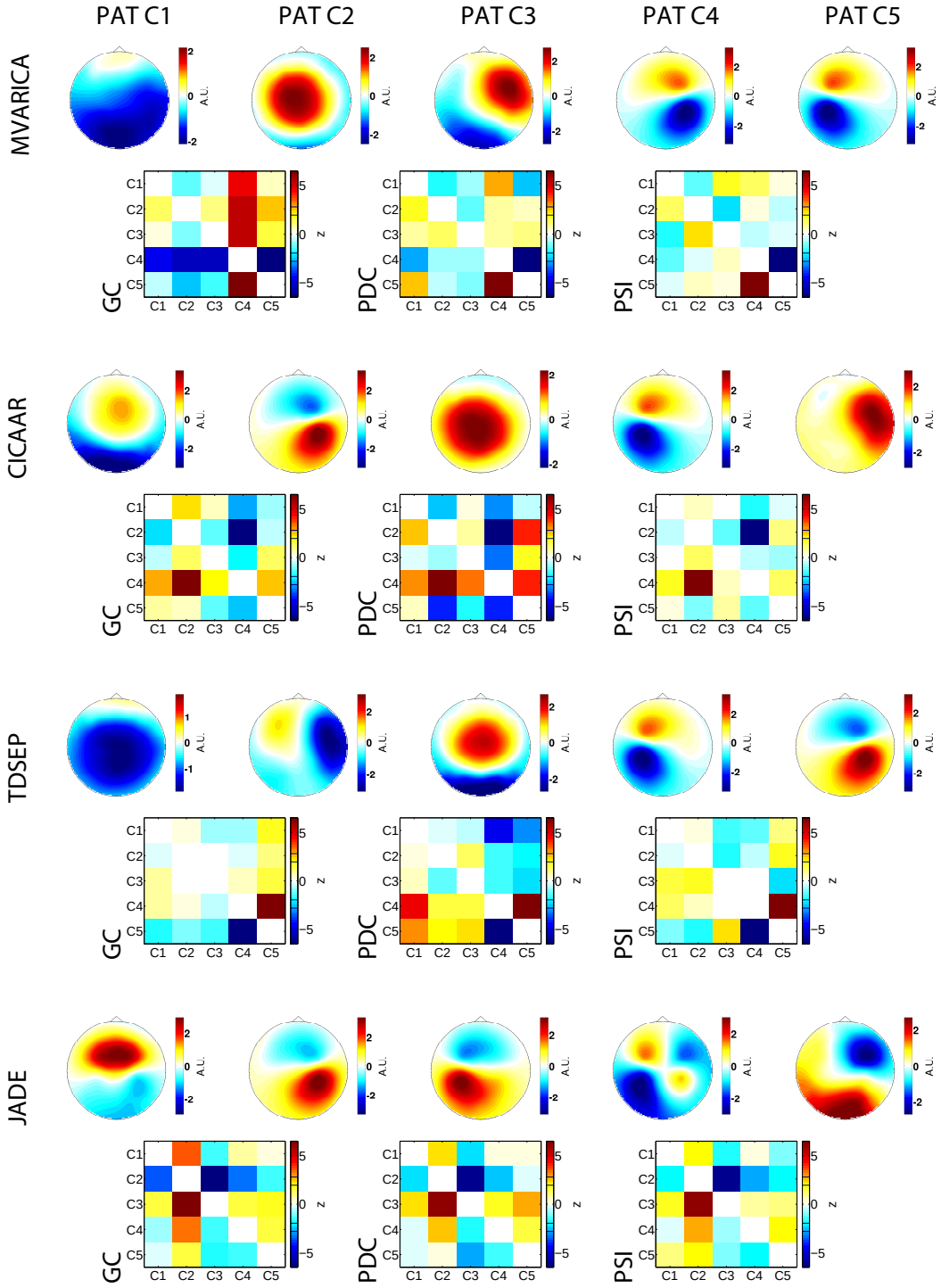


Gaussian innovations

(GOF = 0.50 ± 0.02), by component C_5 for CICAAR (GOF = 0.51 ± 0.02), and by component C_2 for JADE (GOF = 0.46 ± 0.02), while the source in the right hemisphere is best approximated by component C_1 for MVARICA (GOF = 0.48 ± 0.02), by component C_5 for CICAAR (GOF = 0.36 ± 0.02), and by component C_1 for JADE (GOF = 0.41 ± 0.02). Given the insufficient demixing of MVARICA, CICAAR and JADE, we do not report the presence of significant interactions obtained using these preprocessings.

For hyperbolic-secant-distributed source MVAR innovation terms, the picture is quite different. Here, all four BSS methods deliver source components, the patterns of which resemble the patterns of the two true interacting dipoles well. For MVARICA, the true sources are reflected in the components C_5 (left hemisphere) and C_4 (right hemisphere), while the respective components are C_4 (left hemisphere) and C_2 (right hemisphere) for CICAAR, C_4 (left hemisphere) and C_5 (right hemisphere) for TDSEP and C_3 (left hemisphere) and C_2 (right hemisphere) for JADE. The corresponding GOF scores achieved are 0.90 ± 0.01 (C_5) and 0.90 ± 0.01 (C_4) for MVARICA, 0.93 ± 0.01 (C_4) and 0.92 ± 0.01 (C_2) for CICAAR, 0.74 ± 0.01 (C_4) and 0.80 ± 0.01 (C_3) for TDSEP and 0.57 ± 0.01 (C_2) and 0.55 ± 0.01 (C_4) for JADE. The flow from the right-hemisphere source to the left-hemisphere source is significant ($z > 2.8$) for all effective connectivity measures and BSS preprocessings. For PSI, this remains the only significant connection regardless of the BSS method employed. Partial directed coherence estimates one additional significant connection when used with MVARICA, four when used with CICAAR, and two when used with TDSEP, while Granger causality estimates three additional significant connections when used in conjunction with MVARICA and two when used with JADE. This indicates that certain asymmetries influencing the results of Granger-causal measures as revealed in previous sections are also prevalent in the source estimates provided by the tested blind source separation methods. Note that the picture does not change substantially when two-sample statistical testing is emplyed instead of one-sample testing.

Figure 3.7 *(following page)*: Comparison of effective connectivity of simulated EEG as estimated by Granger causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI) after application of blind source separation (BSS) preprocessing according to combined multivariate autoregressive estimation and independent component analysis (MVARICA), convolutive independent analysis via inverse autoregression (CICAAR), temporal decorrelation source separation (TDSEP) and joint approximate diagonalization (JADE). Two source dipoles with tangential orientations are modeled 3 cm below the $C_3/4$ electrodes. Information flow from the left (C_3) to the right (C_4) source is modeled by means of a bivariate AR process with *non-Gaussian* (sech-distributed) innovations. The simulated EEG is superimposed by non-interacting biological and sensor noise (SNR = 1). BSS techniques are applied to the five strongest PCA components. The significance of estimated interactions between demixed signals is measured in terms of z-scores and visualized as matrices, where entries with red and yellow colors ($z > 0$) stand for information outflow and entries with blue and cyan colors ($z < 0$) stand for information inflow of the source marked in the respective row. The Bonferroni-corrected significance level is indicated by a thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line.



super-Gaussian innovations

3.11 Discussion (the importance of accurate source demixing for connectivity analysis)

The current simulation study demonstrates that the estimation of brain interaction from EEG measurements is challenging. We tested three widely-used measures of effective connectivity (while a fourth one, DTF, behaves similarly to PDC). While all of them estimate the correct direction of information flow when being applied directly to source time series, only one of them, PSI, is able to indicate the flow from the corresponding pseudo-EEG. Connectivity measures based on the concept of Granger causality do not succeed in recovering the simulated flow in any of our experiments in sensor space. Rather, when applied on realistic EEG time series, these measures are dominated by properties of the data that have nothing to do with source interaction. In particular, Granger causality and partial directed coherence tend to estimate information flow from electrodes with high SNR to electrodes with low SNR. Since the definition of partial directed coherence is directly based on the estimated MVAR coefficients, this measure additionally depends on the scaling of the data and the choice (of the regularizer) of the MVAR estimator. As a side effect of these dependencies, PDC is biased, and may yield spurious results even when the data have no time structure. Notably, conventional Granger causality does not yield any effect that is stable enough to survive Bonferroni correction, except in the BSS setting, where only a few simultaneous test are conducted.

Sensor-space connectivity estimates provided by PSI correctly indicate the flow from sensors in which the driving source is stronger expressed to sensors in which the receiving source is stronger expressed. However, the distribution of the source signals on the scalp depends on the location of the reference electrode. Changing the reference can hence drastically change PSI (as well as any other connectivity estimate). Moreover, when the underlying interacting sources have complex spread patterns such as bipolar maps, if their field patterns heavily overlap and/or if there are more than two interacting sources, the sensor-space connectivity maps may become almost incomprehensible. A further difficulty occuring in sensor space analyses is that the signal-to-noise ratio affects the extent of the scalp regions between which interactions are estimated. Notably, this extent is not necessarily related to the extent of the underlying interacting brain regions. In order to estimate location and extent of the underlying interacting brain sources, a genuine solution to the EEG inverse problem is required.

We devise two tests which can be used to assess whether estimated information flow is dominated by factors not related to time-lagged interaction. The first test is based on temporal reshuffling of the data. This yields a pseudo time series lacking any effective connectivity, which is to be detected by a meaningful measure of time-lagged effective connectivity. The second test is a special case of reshuffling, in which the temporal order of the time series is exactly reversed. Here, a meaningful measure must diagnose either the complete absence of effective connectivity or the presence of a flow in the opposite direction. In our simulations, the time-reversal test neither confirms nor rejects the results of GC. Partial directed coherence fails the time-reversal test, as it yields significant flow with constant direction regardless of the time-reversal. Hence, in a practical scenario, the spurious results provided by PDC could be rejected on the basis of this test. Notably, PDC also fails the reshuffling test, i. e., behaves similar on reshuffled and original data. The outcome of a statistical test for differences of PDC results obtained on reshuffled and original samples is, however, similar

to that of a one-sample test of GC, which means that PDC here lacks statistically significant results. PSI passes both tests in our empirical study. Moreover, its ability to reverse the flow direction for reversed time series can be derived analytically.

We assessed various source demixing procedures, which have been used previously in EEG-based synchronization and connectivity studies, as preprocessing steps for subsequent connectivity estimation. All of these preprocessings transform the data into a reference-free representation, which is an improvement over standard sensor-space analysis. Apart from that, source demixing using the Laplace transformation does not substantially faciliate the connectivity estimation. This may be due to the Laplacian's assumption that all sources are superficial and radially oriented, which is not the case in our simulated example.

Linear inverse source reconstruction by means of the weighted minimum norm estimate turns out to be suitable in principle for subsequent source connectivity analysis using PSI. However, the diffuse spatial distribution of the current density estimated by WMN prevents a sufficient spatial separation of the interacting sources. In our simulation setting, the sensor-space result that "there is information flow from the left to the right hemisphere" could be made only slightly more precise by taking into account the locations of the estimated underlying sources, i. e., the estimated dipole strength per voxel. Doing so, however, also leads to the discovery of a spurious occipital source, which, according to PSI, incorrectly appears to exchange information with more central sources. The application of GC and PDC in WMN source space yields no significant interaction at all. Since linear inverse solutions are generally associated with blurred source distributions and the inability to distinguish multiple sources (Grave de Peralta-Menendez and Gonzalez-Andino, 1998; Haufe et al., 2008), we conclude that they are not well suited as a preprocessing step for source connectivity analysis. Nonlinear inverses such as the minimum-current estimate, may achieve a better spatial separation of the sources but are currently computationally too expensive to be applied to longer time series. For that reason, they have not been proposed for source connectivity analysis so far.

The performance of blind source separation techniques depends crucially on the Gaussianity of the source MVAR innovations. We note that non-Gaussianity is fundamental for identifying the parameters of the correlated source model given by (5.7)/(5.8) in general. In our simulated example, TDSEP is, however, found to separate sources that are generated with Gaussian innovations fairly well. We suggest that this is the case, because any splitting of the two interacting sources (which are the strongest signal components) into multiple components introduces instantaneous cross-correlations, which increase the overall cost function TDSEP seeks to minimize. In a broader simulation presented in Section 5.2, we consider multiple connected sources which are all similarly strong. Here, the performance of TDSEP systematically breaks down.

Only two of the tested BSS methods (MVARICA and CICAAR) make an explicit assumption regarding the non-Gaussianity of the source MVAR innovations. Consequently, these are the methods that achieve the best source separation on respective data in terms of GOF scores. The performance is worst for JADE, which is understandable given that this method assumes non-Gaussian sources rather than source innovations. The performance of TDSEP falls into the intermediate range, as for Gaussian innovations, which is reasonable considering that TDSEP only uses second-order statistics. Interestingly, all BSS methods separate the sources well enough in the non-Gaussian case to yield a significant flow from the component best approximating the true driver to the component best approximating the true receiver.

To sum up, in this chapter we have studied the most simple computational model of brain interaction conceivable: a two-dipole source model in which there is a linear time-lagged influence of one source onto the other. Only the phase-slope index turned out to be capable of detecting this interaction from simulated EEG measurements. However, even the interpretability of the phase-slope index is limited in practice by overlapping field patterns of the underlying sources, and by the data-dependent bias induced by the choice of the reference electrode. Notably, Granger-causal approaches are not able to recover the true interaction in our simulated example regardless of the preprocessing step or source demixing employed, except for certain blind source separation techniques. As a result of the simulations performed here, we conclude that most of the currently available source demixing approaches are not well suited for subsequent source connectivity analysis. To improve over existing methods, we require that inverse source reconstruction methods should be able to spatially separate multiple (at least two) distinct sources, while being applicable to long EEG time series at the same time. Blind source separation techniques, on the other hand, should be able to model interactions between the demixed time series. The core of this thesis is development of methods fulfilling these criteria. These are presented in the next two chapters.

4 Source connectivity analysis via inverse source reconstruction preprocessing

Inverse source reconstruction for EEG (and MEG) is concerned with the mathematical inversion of the approximately known forward mapping from brain sources to EEG sensors. It is important to virtually any neuroscientific application, in which measurements are not only processed by black box algorithms in order to extract discriminative information (as, for example, for controlling neuroprostheses), but subjected to neurophysiological interpretation in order to understand the underlying cognitive processes. Unfortunately, the interpretability of sensor-space data is limited by the presence of volume conduction, which causes the cerebral activity to mix into sensors. A single brain source can therefore contribute substantially to quite remote locations on the scalp, depending on its depth. Moreover, these contributions can have either equal or opposite signs, depending on the orientation of the source's equivalent dipole current moment vector. Researchers not familiar with the physical process of EEG generation might easily disregard these issues and come to wrong conclusions regarding number, spatial extent and location of the underlying sources. As an example consider the sources used throughout our simulations in Chapter 3, which are both located under central electrodes but contribute mostly to frontal and parietal scalp electrodes.

Interpretability is more degraded, the more brain sources there are, which are simultaneously active and the more the field patterns of these sources spatially overlap. On the other hand, the presence of multiple sources is exactly what should be assumed in EEG connectivity analyses. Our previously conducted simulations show that sensor-space connectivity estimates are hard to interpret even in the most simple case of two interacting sources due to ambiguities caused by the choice of the reference electrode, for example. Moreover, if inappropriate measures are applied to sensor-space data, source and noise mixing due to volume conduction leads to estimation biases, which render the results completely misleading. For these reasons, reconstruction of the underlying sources in the first place is even more crucial for EEG-based connectivity analysis than it is for other neuroscientific investigations. Inverse source reconstruction, compared to blind source separation, has the advantage that the estimated sources can be mapped directly to the brain anatomy. Available approaches can be roughly divided into scanning (Schmidt, 1986; Van Veen and Buckley, 1988; Van Veen et al., 1997; Mosher and Leahy, 1999; Sekihara et al., 2005; Zwoliński et al., 2010), dipole fitting (Scherg and von Cramon, 1986; Scherg and Ebersole, 1993) and distributed source imaging (Jeffs et al., 1987; Ioannides et al., 1990; Pascual-Marqui et al., 1994; Matsuura and Okabe, 1995; Wipf et al., 2010) techniques. We here exclusively consider distributed inverse solvers, which are especially suitable for EEG source connectivity analysis since i) they provide a genuine solution to the inverse problem (that is, their source estimates may fully explain the data) and ii) they are not constrained to a predefined number of sources.

As has been pointed out previously, the EEG inverse problem is ill-defined, since any EEG measurement can be equally well explained by infinitely many different source distributions. Therefore, in order to resolve this ambiguity, it is necessary to impose constraints on the inverse

solution. In distributed source imaging, this is done by adding a penalty term to the negative log likelihood that reflects prior assumptions on the sources. The two perhaps most common assumptions are spatial smoothness and focality, respectively, both of which can be motivated by neurophysiological arguments. Representative estimators implementing either one or the other of these assumptions were introduced in Section 2.7. As pointed out there, smoothness constraints are often implemented by means of Tikhonov regularization. Such approaches lead to solutions that are linear in the observations, which means that they can be applied even to large time series via a multiplication with a data-independent inversion matrix. However, the resulting sources tend to be spread over a considerable part of the source space (cortex or brain), which is not always physiologically meaningful. Related to this, linear estimators have limitations in distinguishing multiple sources, which is especially obstructive if one is interested in source connectivity (Schoffelen and Gross, 2009).

Focal source images are usually obtained by imposing the sparsity-enforcing ℓ_1-norm penalty on the sources. However, the resulting estimates tend to be unstable and spatially scattered. Also, spatially extended sources are not modeled adequately by current sparse methods. Rather, they are approximated by a number of disconnected active dipoles, which may convey the misleading impression that multiple sources are active. A specific problem of the ℓ_1-norm concerns the fact that it leads to sparsity in the single components of the dipole moment vectors rather than sparsity in the dipoles themselves. This biases the estimated currents towards the coordinate axes, which is a highly nonphysiological behaviour given that the choice of the coordinate system is arbitrary. There are two further obstacles regarding the use of sparse source estimates for source connectivity analysis. One is their high computational cost. Unlike for the ℓ_2-norm, ℓ_1-norm optimal solutions have to be computed iteratively using nonlinear optimization for each sample, which is time-consuming for large numbers of samples. Even more severe for a potential connectivity analysis of the sources is the lack of temporal continuity, which is caused by the fact that a different spatial sparsity pattern may be chosen for each time sample.

In this chapter, we develope methods that combine the strengths of smooth and sparse distributed source imaging approaches while avoiding their weaknesses, keeping in mind the ultimate goal of improving source connectivity analyses. In Section 4.1, we introduce focal vectorfield reconstruction (FVR), which achieves a compromise between smoothness and sparsity by means of a hybrid penalty. In addtion, FVR employs a new type of depth weighting and achieves rotational invariance by inducing sparsity of entire current vectors by means of $\ell_{1,2}$-norm penalization. In Section 4.2, sparse basis field expansion inverse source reconstruction (S-FLEX) is introduced, which has similar properties as FVR. In addition, the S-FLEX solution can be calculated for a large number of samples. The assumption S-FLEX makes for joint localization is that the set of active sources is shared among the samples, which exactly matches the requirements of source connectivity analysis. In Section 4.3, we introduce the earth mover's distance (EMD) as a meaningful measure for comparing arbitrary source distributions under a common framework. We use the EMD in Section 4.4 to assess the ability of FVR and S-FLEX to separate multiple brain sources compared to purely smooth and purely sparse approaches. The applicability of our methods in a real-world setting is demonstrated in Section 4.5, where the task is to distinguish sources in the left and right sensorimotor cortices, which are activated by means of electrical stimulation of the left and right median nerves. Finally, Section 4.6 describes the application of S-FLEX to the simulated dataset introduced in Chapter 3.

4.1 Focal vectorfield reconstruction (FVR)

We are now concerned with the derivation of an inverse methodology fulfilling certain requirements on the spatial distribution of the sources, which are important for subsequent source connectivity analysis. The method, *focal vectorfield reconstruction* (FVR), was introduced in Haufe et al. (2008), from which large parts of this section are taken. Focal vectorfield reconstruction deals with the situation in which only one EEG sample is available. Hence, it is not directly applicable to the source analysis of time series but may be considered a first step towards that.

As has been mentioned in Section 2.7, the regularizing criterion of a distributed inverse method should encode prior knowledge on what a "good" solution looks like. Conventional minimum ℓ_p-norm solutions, weighted or not, are either rotationally invariant but highly non-focal ($p = 2$) or sparse but violating rotational invariance ($p = 1$). The problem with many sparse approaches is that they do not take into account the vectorial nature of currents, as a result of which the orientations of most of the estimated dipoles are axes-parallel, which is a highly non-physiological behaviour.

One possibility to alleviate this problem is to fix the orientations a-priori in a meaningful way. In Uutela et al. (1999), the dipole orientations are adopted from the minimum ℓ_2-norm estimate, while dipole amplitudes are estimated using the ℓ_1-norm penalty. A more sophisticated approach is suggested in Huang et al. (2006), where activity in a voxel is discouraged, if the orientation of the minimum-norm estimate in the respective voxel is close to one of the coordinate axes. In cortically-constrained approaches (Dale and Sereno, 1993; Kincses et al., 2003), dipoles are assumed to be perpendicular to the cortical surface based on the consideration that the apical dendrites of pyramidal neurons are oriented in this way. This approach, however, requires very precise knowledge of the individual cortical geometry, as small changes in the normal vector can quickly lead to considerably different forward equations.

We here consider the estimation of full vectorial currents, which we request to be invariant with respect to rotations of the coordinate system. Moreover, we assume that brain sources are focal. Both goals can be achieved simultaneously by penalizing a global ℓ_1-norm of local ℓ_2-norms, which leads exactly to the $\ell_{1,2}$-norm regularizer introduced in Section 2.3. The use of the $\ell_{1,2}$-norm in inverse source imaging has been proposed in three simultaneous papers by Ding and He (2008); Haufe et al. (2008) and Ou et al. (2009).

In order to obtain source estimates which are more plausible than purely sparse or purely smooth solutions, we additionally request that meaningful source estimates should be sparse with only a minimal number of continuous nonzero patches. This can be achieved by imposing sparsity of the spatial Laplacian in addition to conventional sparsity, and thereby relaxing the strict focality constraint in favor of a more robust "simplicity" requirement. Imposing sparse second derivatives leads to smoothness in the same way as minimizing their ℓ_2-norm (as in LORETA). To understand the difference it is helpful to consider that in one dimension, sparse second-order derivatives characterize exactly the piecewise linear functions.

The idea of using multiple penalties has also been proposed in the statistics literature. For example, the *fused lasso* algorithm (Tibshirani et al., 2005) considers joint regularization of variables and their derivatives, while the *elastic net* considers mixed penalties containing ℓ_1- and ℓ_2-norm terms (Zou and Hastie, 2005). The elastic net has also been applied in distributed source imaging (Vega-Hernández et al., 2008).

4.1.1 Discrete Laplace operator

Since, in contrast to LORETA, our approach does not rely on invertibility of the Laplacian, we use the following definition, which does not penalize nonzero currents at the boundaries of the source space. This is important, because cerebral activity of interest often originates from cortical structures, which are located superficially. The $N \times N$ Laplacian D^{FVR} is given by

$$d_{i,j}^{\mathrm{FVR}} = \frac{1}{h^2} \begin{cases} -\left|\{k \mid \|\mathbf{u}_i - \mathbf{u}_k\|_2 = h\}\right| & i = j \\ 1 & \|\mathbf{u}_i - \mathbf{u}_j\|_2 = h \\ 0 & \text{else} \end{cases}, \tag{4.1}$$

where $|\mathcal{S}|$ denotes the number of elements in the set $\mathcal{S}$ and h is the inter-voxel distance.

4.1.2 Depth compensation

Due to attenuation in the tissue, the signals captured by the sensors are much stronger for superficial sources compared to deep sources. This causes an estimation bias when using ℓ_p-norm penalties. We counteract this bias by means of a *depth compensation* matrix, which penalizes activity at superficial locations. While the conventional approach is based on weighting activity at each voxel with the inverse of the norm of the respective column of the lead field matrix (Jeffs et al., 1987; Ioannides et al., 1990), the weighting proposed here is inspired by the post-hoc current standardization of sLORETA (see Section 2.7). That is, we penalize activity at locations with high variance according to the variance estimate $\widehat{V}$ derived in Pascual-Marqui (2002). Defining $W_n = V_n^{\frac{1}{2}} \in \mathbb{R}^{3\times 3}$ to be the matrix square root of the n-th 3×3 blockdiagonal part of $\widehat{V}$, we define the depth-compensation matrix as

$$W = \begin{pmatrix} W_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_N \end{pmatrix} \in \mathbb{R}^{3N\times 3N}. \tag{4.2}$$

Using W instead of conventional column norm weighting has been proven to drastically increase localization accuracy of various sparse and non-sparse methods including the weighted minimum-norm estimate and the minimum-current estimate (Haufe et al., 2008).

4.1.3 Model specification and parameter estimation

Let $\mathbf{s}_i \in \mathbb{R}^3$ denote the dipole moment at the i-th voxel such that $\mathbf{s} = (\mathbf{s}_1^\top, \cdots, \mathbf{s}_N^\top)^\top$. Similarly, let $\mathbf{t} = (D^{\mathrm{FVR}} \otimes I_3)\mathbf{s}$, such that $\mathbf{t}_i \in \mathbb{R}^3$ is the moment of the Laplacian of the source field at the i-th voxel. The FVR source estimate is then given by

$$\widehat{\mathbf{s}}^{\mathrm{FVR}} = \arg\min_{\mathbf{s}} \|\mathbf{x} - A\mathbf{s}\|_2^2 + \lambda \left(\sum_{i=1}^{N} \|W_i \mathbf{s}_i\|_2 + \alpha \sum_{i=1}^{N} \|W_i \mathbf{t}_i\|_2 \right), \tag{4.3}$$

where $\mathbf{x}$ is the EEG measurement, A is the lead field matrix, λ is the regularization parameter guiding the tradeoff between model likelihood and regularization and α is an additional parameter weighting the relative importance of smoothness and sparsity in the regularizer. Note that,

compared to LORETA, the order in which the Laplacian and the depth-weighting operators are applied is interchanged here.

The formulation using a weighted sum of likelihood and regularization terms is a re-parametrization of our original formulation (Haufe et al., 2008), in which the data fidelity is encoded using the hard constraint $\|A\mathbf{s} - \mathbf{x}\|_2 \leq \epsilon$. Both formulations are equivalent in the sense that for any λ there exists an ϵ (and vice-versa), such that the solutions of (4.3) and the cost function defined in Haufe et al. (2008) coincide. Moreover, both formulations are convex, such that the existence of a unique minimum is guaranteed. The solution can be obtained by means of second-order cone programming (SOCP) optimization (see Haufe et al., 2008).

4.1.4 Rotational invariance

A central technical aspect of our method is the way sparsity of the current density is enforced. We propose penalizing the length of the current amplitudes rather than the absolute values of individual moments of the current vectors using $\ell_{1,2}$-norm penalties. With this choice not only sparsity but also rotational invariance of the FVR solution is guaranteed, as is shown in the following. If the coordinate system is rotated by an orthogonal matrix U, the lead field matrix A and the sources $\mathbf{s}$ are transformed as

$$A \longrightarrow A\widehat{U}^\top \equiv A_U \tag{4.4}$$

$$\mathbf{s} \longrightarrow \widehat{U}\mathbf{s} \equiv \mathbf{s}_U\ , \tag{4.5}$$

where $\widehat{U} \equiv I_N \otimes U$ is the rotation operator for all voxels. Then, the sLORETA variance estimate transforms as

$$\widehat{V} = A^\top \left(AA^\top\right)^{-1} A \longrightarrow A_U^\top \left(A_U A_U^\top\right)^{-1} A_U = \widehat{U}\widehat{V}\widehat{U}^\top \tag{4.6}$$

and the block diagonal entries of the square root of $\widehat{V}$ transform as

$$W_i \longrightarrow \left(UV_iU^\top\right)^{\frac{1}{2}} = U\sqrt{V}_iU^\top = UW_iU^\top. \tag{4.7}$$

Now, (4.3) is rotationally invariant since a)

$$\|W_i\mathbf{s}_i\|_2 \longrightarrow \|UW_iU^\top U\mathbf{s}_i\|_2 = \|W_i\mathbf{s}_i\|_2 \tag{4.8}$$

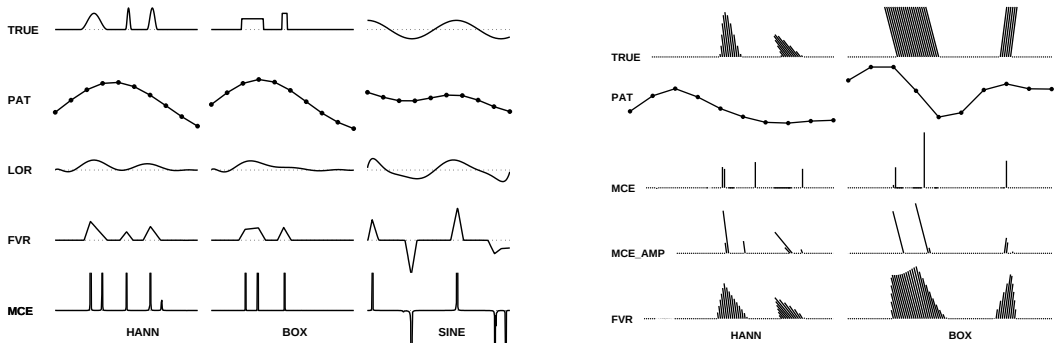
and b) the same holds for $\|W_i\mathbf{t}_i\|_2$ because the Laplacian is a scalar differential operator and the rotation is independent of space (i. e. the moment of each voxel is rotated identically). A similar invariance can be shown for transformations with general unitary matrices U, which also cover the case of complex phase shifts $U = \exp(i\phi)I_3$. That is, the phase-shifted source estimate obtained from a complex-valued field pattern is identical to the souce estimate obtained from the phase-shifted version of the field pattern. This property is important when localizing complex-valued spatial patterns as, for example, resulting from a Fourier decomposition.

Note that a rotation of the coordinate system must be distinguished from a rotation of the grid. Invariance with respect to the former is exactly fulfilled implying that the method itself does not prefer specific source orientations. Rotational invariance of the latter is an approximation limited by the discrete nature of the Laplacian, however with negligible impact for small voxel distances.

4.1.5 Illustration

Figure 4.1 (a) illustrates the main properties of LORETA and the minimum-current (MC) estimate compared to that of FVR. The current density domain is defined to be a straight line of 300 scalar sources. Three source configurations, consisting of either three Hanning windows, two boxcar windows or a single sine wave, are simulated. Source reconstruction is performed on noise-free pseudo-measurements, which are obtained by smoothing and subsampling the sources. Apparently, only FVR is able to recover the exact number of sources in all three cases. LORETA is not able to distinguish all three sources in the Hanning example. Instead, one estimated source is placed exactly in between two true sources. Minimum-current estimates consist of spikes, the number and locations of which do not match the true source configuration.

In Figure 4.1 (b), the effect of enforcing sparse current amplitudes is illustrated on the basis of two simulations. The sources are modeled as a straight line of 100 two-dimensional vectors. In one case, we simulate two sources with Hanning window envelopes. In the other example, two boxcar windows are used. All vectors related to the same source have equal orientation. Ten pseudo-measurements are constructed from the source vectors by means of lowpass-filtering. In the examples shown, MC estimates according to Matsuura and Okabe (1995) are all parallel to one of the two axes. In contrast, the modified version minimizing the ℓ_1-norm of vector amplitudes recovers the original orientations well, while being even sparser. Finally, additional sparsity in the amplitudes of the Laplacian removes the problem of source scattering.



(a) (b)

Figure 4.1: (a) One-dimensional simulation illustrating the characteristics of conventional inverse solutions, as well as focal vectorfield reconstruction (FVR). Simulated source configurations (TRUE) include three hanning windows (HANN), two boxcar windows (BOX) and a sine wave (SINE). Hypothetical measurements (PAT) are obtained by smoothing and subsampling the sources. Source reconstructions according to one-dimensional versions of low-resolution tomography (LOR), FVR and the minimum-current estimate (MCE) are shown in the three lower panels. (b) Simulation, illustrating the approaches of MCE and FVR to achieve sparsity of a vector field. A straight line of two-dimensional vectors models the current density. The vector envelopes are combinations of either two Hanning (HANN) or two boxcar (BOX) windows. Vector orientations are fixed within sources. True sources (TRUE) and pseudo EEG patterns are shown in the upper panels of the plot. The inverse solutions of MCE, a modified version of MCE working on amplitudes (MCE_AMP) and FVR are shown below.

4.2 Sparse basis field expansions (S-FLEX)

While the focal vectorfield reconstruction solution enjoys favorable properties, it can only be computed in practice for a single or few spatial patterns. This is due to the use of a mixed penalty enforcing sparsity in two different spatial bases, which prevents us from using large-scale numerical solvers for the minimization. Another downside of FVR is that the compromise between smoothness and focality of the sources has to be pre-specified via the hyper-parameter α, which practically encodes the preference for a certain spatial extent of the sources. Although α can be optimized using cross-validation, a fixed choice may still be suboptimal if multiple differently-sized brain areas are simultaneously active.

In this section, we propose EEG source reconstruction through *sparse basis field expansions* (S-FLEX) as an alternative method that achieves a compromise between smoothness and focality similar to FVR. In contrast to FVR, S-FLEX is able to model the simultaneous occurence of extended sources of different sizes and shapes. Moreover, the S-FLEX solution has a simpler mathematical form. Its cost function enables the deployment of a very efficient optimization scheme, by which it becomes possible to solve inverse source reconstruction problems involving orders of magnitude more variables than previously. These additional variables can be used to increase the spatial resolution in source space, or to localize larger datasets, as it is required, for example, in source connectivity analysis. Sparse basis field expansions were introduced in Haufe et al. (2009) for single EEG measurements. In Haufe et al. (2011a), the extension to multiple measurements was suggested, along with an efficient implementation scheme.

Model specification and parameter estimation

Instead of estimating the current moments $\mathbf{s}$ directly, S-FLEX models the current density as a linear combination of (potentially many) spatial *basis fields*, the coefficients of which are to be estimated. A basis field is defined as a vector field, in which all output vectors point in the same direction, while the magnitudes are proportional to a scalar (basis) function $g : \mathcal{B} \rightarrow \mathbb{R}$. Given a set of basis functions $g_l, l = 1, \ldots, L$ (called a *dictionary*), the basis field expansion reads

$$\widetilde{\mathbf{s}}(\mathbf{u}) = \sum_{l=1}^{L} \mathbf{c}_l \, g_l(\mathbf{u}) \; , \tag{4.9}$$

with coefficient vectors $\mathbf{c}_l \in \mathbb{R}^3, l = 1, \ldots, L$. Note that we employ the basis field expansion not on the genuine current density but on depth-weighted sources. As will become clear later, this prevents us from introducing weight matrices in the regularizer, which is the key to efficient optimization (see below). The relation between $\widetilde{\mathbf{s}}$ and the true sources $\mathbf{s}$ is given by

$$\mathbf{s}(\mathbf{u}_n) = W_n^{-1} \widetilde{\mathbf{s}}(\mathbf{u}_n), n = 1, \ldots, N \; , \tag{4.10}$$

where W_n^{-1} are the inverses of the 3×3 depth compensation matrices W_n defined in the previous section. By including one coefficient for each spatial dimension, we learn orientations as well as amplitudes of the current moment vectors in our model. Moreover, if the $\mathbf{c}_l$ are modeled to be complex-valued, phase angles are learned, too. Let $C = (\mathbf{c}_1, \ldots, \mathbf{c}_L)^\top \in \mathbb{R}^{L \times 3}$ contain the

coefficients, and let the corresponding basis function values evaluated at all locations $\mathbf{u}_n$ for which the lead field matrix has been computed be collected in the design matrix

$$G = \begin{pmatrix} g_1(\mathbf{u}_1) & \cdots & g_L(\mathbf{u}_1) \\ \vdots & \ddots & \vdots \\ g_1(\mathbf{u}_N) & \cdots & g_L(\mathbf{u}_N) \end{pmatrix} \in \mathbb{R}^{N \times L} . \tag{4.11}$$

The forward model then reads

$$\mathbf{x} = \Gamma \, \mathbf{vec}\,(C) \ , \tag{4.12}$$

where $\Gamma = AW^{-1}(G \otimes I_3) \in \mathbb{R}^{M \times 3L}$, and where W^{-1} is the inverse of the depth-compensation matrix introduced in the context of focal vectorfield reconstruction.

Solving (4.12) for C does not yield a unique solution if the number of coefficients $3L$ is larger than the number of electrodes M, which is the common situation. As usual, the ambiguity can be overcome by regularization, i. e., by imposing additional constraints on the variables. Here, we assume that, for an appropriately chosen dictionary, the current density can be well approximated by a *small number* of basis fields. This can be achieved by estimating a *sparse* coefficient matrix C. Besides the regularizing effect, sparse decompositions offer the additional advantage of providing a way of interpreting estimated current densities in terms of the selected basis functions (those having corresponding nonzero coefficients in C). The premise for such interpretability is that the basis functions themselves are simple enough, which should be ensured when designing the dictionary.

As was pointed out in the previous section, important issues arising in sparse EEG source reconstruction are rotational invariance and invariance with respect to phase shifts. As in previous approaches (Ding and He, 2008; Haufe et al., 2008; Ou et al., 2009; Bolstad et al., 2009), we utilize the $\ell_{1,2}$-norm penalty to achieve this. Our approach amounts to estimating

$$\begin{aligned} \widehat{C} &= \arg\min_C \ \mathcal{L}(C) + \lambda \mathcal{R}(C) \\ \mathcal{L}(C) &= \|\mathbf{z} - \Gamma \, \mathbf{vec}\,(C)\|_2^2 \\ \mathcal{R}(C) &= \sum_{l=1}^{L} \|\mathbf{c}_l\|_2 = \|C\|_{1,2} \ , \end{aligned} \tag{4.13}$$

where $\mathcal{R}(C)$ is the $\ell_{1,2}$-norm regularizer, $\mathcal{L}(C)$ is the negative log likelihood and λ is a positive constant controlling the tradeoff between likelihood and regularization term.

Equation 4.13 is a convex optimization problem composed of a quadratic loss function and a convex nondifferentiable regularizer. It thus shares similarities with the optimization problems discussed in Polonsky and Zibulevsky (2004); Malioutov et al. (2005); Ding and He (2008); Haufe et al. (2008); Ou et al. (2009); Wipf and Nagarajan (2009) and Bolstad et al. (2009). In the majority of these papers, the cost function is reformulated as an instance of second-order cone programming (SOCP, Lobo et al., 1998). The proposed interior-point SOCP solvers are, however, only applicable to small- to medium-sized problems not exceeding several ten thousands of variables. For this reason, some authors perform a dimensionality reduction step prior to the estimation in order to reduce the number of variables and/or the number of observations (Malioutov et al., 2005; Ou et al., 2009). Here, we make use of a more recent advance in numerical optimization that enables us to

solve S-FLEX instances involving millions of model parameters. The proposed algorithm is based on deriving the Fenchel dual of the optimization problem and applying the augmented Lagrangian technique (Tomioka and Sugiyama, 2009). It has thus been termed dual augmented Lagrangian (DAL). We use the reference implementation of DAL, which is provided as open source software[1]. Unfortunately, sparsity of *linearly transformed* variables is not handled by DAL, for which reason our previously-described approach FVR cannot benefit from it.

The dictionary

The idea behind enforcing smoothness and focality at the same time is to avoid the scattering of activity found for many purely focal inverse approaches, while at the same time to maintain high spatial resolution and the associated ability to distinguish multiple sources. In FVR, a combination of two penalties is used to achieve this effect. Here, the task is addressed by designing an appropriate basis function dictionary. We consider an expansion into spherical Gaussian basis functions, which are smooth but also well localized due to their unimodal structure and exponentially decaying tails. Due to the latter, sparse combinations of Gaussians give rise to good spatial separation of sources that are simulataneously active. Using a redundant dictionary containing Gaussians of different length scales, we further achieve reconstruction of sources with arbitrary shape using only few basis elements. Formally, the elements of the basis function dictionary are defined as

$$g_{\mathbf{u}_n,\varsigma}(\mathbf{u}) = \left(\sqrt{2\pi}\varsigma\right)^{-3} \exp\left(-\tfrac{1}{2}\left\|\mathbf{u}-\mathbf{u}_n\right\|_2^2 \varsigma^{-2}\right) , \tag{4.14}$$

where $\mathbf{u}_n$, $n = 1, \ldots, N$ are the center locations and ς_s, $s = 1, \ldots, S$ are the widths of the Gaussian functions (see Figure 4.2 (a) for examples).

The proposed regularization aims at selecting the smallest possible number of basis fields necessary to explain the measurement. This can, however, only be achieved approximately, since not the number of nonzero coefficient vectors but the sum of their magnitudes is effectively penalized by the $\ell_{1,2}$-norm. It is therefore important to normalize the basis functions in order not to prefer some of them a-priori due to low magnitudes. Let G_{ς_s} be the $N \times N$ matrix containing the evaluations of all basis function with standard deviation ς_s. The large matrix

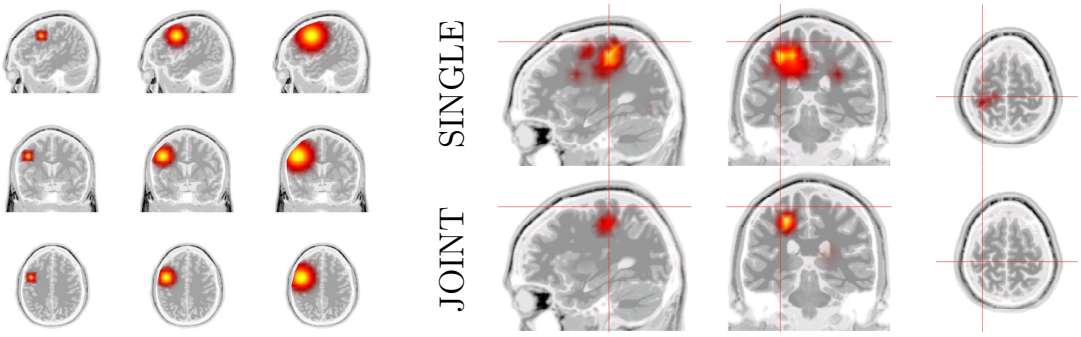
$$G = \left(\frac{G_{\varsigma_1}}{\left\| \mathbf{vec}\left(G_{\varsigma_1}\right) \right\|_1}, \ldots, \frac{G_{\varsigma_S}}{\left\| \mathbf{vec}\left(G_{\varsigma_S}\right) \right\|_1} \right) \in \mathbb{R}^{N \times SN} \tag{4.15}$$

is constructed using normalized G_{ς_s}. By this means, no length scale is preferred a-priori.

Extension to multiple measurements

While (4.13) deals only with single field patterns, we extend S-FLEX here to the localization of multiple measurements, where the goal is to estimate T current densities $\mathbf{s}_n(t)$ based on the corresponding EEG patterns $\mathbf{x}(t)$. Let $X = (\mathbf{x}(1), \ldots, \mathbf{x}(T)) \in \mathbb{R}^{M \times T}$ and let $\mathbf{c}_l(t) \in \mathbb{R}^3$ be the coefficient vector describing the contribution of the l-th basis field to the t-th pattern. Defining $\widetilde{C}(t) =$

[1] http://mloss.org/software/view/183/



(a) (b)

Figure 4.2: (a) Examples of Gaussian basis functions with spatial standard deviations $\varsigma_1 = 0.5\,\text{cm}$, $\varsigma_2 = 1\,\text{cm}$ and $\varsigma_3 = 1.5\,\text{cm}$. (b) Comparison of individual (SINGLE) and joint localization of ten simulated noisy measurements using sparse basis field expansions. The location of the true simulated source is indicated by a red cross-hair.

$(\mathbf{c}_1(t), \ldots, \mathbf{c}_L(t))^\top \in \mathbb{R}^{L\times 3}$, $\widetilde{C} = (\mathbf{vec}(\widetilde{C}(1)), \ldots, \mathbf{vec}(\widetilde{C}(T)))$ and $\widetilde{\mathbf{c}}_l = (\mathbf{c}_l(1)^\top, \ldots, \mathbf{c}_l(T)^\top)^\top \in \mathbb{R}^{3T}$, the multiple-measurement S-FLEX estimate reads

$$\begin{aligned} \widehat{\widetilde{C}} &= \arg\min_{\widetilde{C}} \widetilde{\mathcal{R}}(\widetilde{C}) + \lambda \widetilde{\mathcal{L}}(\widetilde{C}) \\ \widetilde{\mathcal{R}}(\widetilde{C}) &= \sum_{l=1}^{L} \|\widetilde{\mathbf{c}}_l\|_2 \\ \widetilde{\mathcal{L}}(\widetilde{C}) &= \|\mathbf{vec}\left(X - \Gamma\widetilde{C}\right)\|_2^2 . \end{aligned} \tag{4.16}$$

Note that the solution is equivalent to (4.13) for $T = 1$. However, for $T > 1$ it is not equivalent to solving T problems of type (4.13) separately as, in our case, the $3T$ coefficients related to a single basis function are tied under a common ℓ_2-norm and can only be pruned to zero at the same time. Thus, the selection of basis functions which contribute coherently to several patterns is facilitated, while at the same time orientations and amplitudes of the estimated source fields may differ per pattern. Such joint (or co-) localization has also been suggested in previous work. The idea originates from Polonsky and Zibulevsky (2004) and appears also in Malioutov et al. (2005), Wipf and Rao (2007), Ou et al. (2009) and Bolstad et al. (2009). Malioutov et al., Ou et al. and Bolstad et al. use the technique for spatio-temporal source localization, where the ℓ_2-norm in temporal domain ensures that there are no artificial jumps in the estimated source time courses. Similar approaches using sparsity in spatio-temporal time-frequency dictionaries are proposed in Trujillo-Barreto et al. (2008) and Gramfort et al. (2011b), while Gramfort et al. (2011a) suggest a graph-theoretic approach to imposing spatio-temporal regularity constraints. Malioutov et al. and Ou et al. suggest that joint localization leads to a better noise suppression compared to the single-timepoint estimator. A similar effect has been reported to improve BCI classification (van Gerven et al., 2009).

4.3 The earth mover's distance

For assessing the quality of source reconstructions we propose a measure of the disagreement of the simulated and the estimated dipole amplitudes, which is called the *earth mover's distance* (EMD). This quantity is suitable for comparing functions with possibly nonoverlapping support, for which a metric between points in the domain is available. In the case of EEG inverse solutions, the Euclidean distance between dipole locations provides such a metric. The EMD has been introduced in computer vision for comparing images (Rubner et al., 2000).

To understand what the EMD does, consider that for a given source distribution the amplitude at each voxel is divided into a huge number of "units" with tiny and fixed amplitude. Two source distributions have the same total number of units. One can now transform the first source into the second source by moving the units of the first source to match those of the second source. The average distance the units should be transported depends on the specific transformation we choose. The minimum average distance (averaged over all units and minimized over all possible transformations) defines the EMD.

The idea of using this metric in this context is that it provides a meaningful measure for arbitrary types of source distributions. We can, e. g., compare a few-dipoles solution with highly distributed sources without having to worry which local maximum corresponds to which dipole, or we can compare a 3-dipole solution with a 2-dipole solution in a meaningful way.

4.4 Reconstruction of simulated sources

4.4.1 Assessing single-measurement localization performance

Validation of methods for inverse source reconstruction is generally difficult due to the lack of a "ground truth". The measurements $\mathbf{x}$ do not provide such a truth, since the main goal in inverse imaging is not to find an accurate representation for the EEG, but to approximate the underlying current density $\mathbf{s}$, which is unknown, as good as possible. Therefore, a standard way of evaluating inverse methods is to assess their ability to reconstruct known functions. This is performed here by reconstructing simulated current sources, which are generated as follows. We use the realistic "Montreal" head model (Holmes et al., 1998) with $N = 2\,142$ dipoles arranged at positions $\mathbf{u}_n$, $n = 1, \ldots, N$ in a cubic grid with 1 cm side length. The lead field matrix A is constructed for 118 electrode positions defined in the extended 10-20 positioning system according to Nolte and Dassios (2005). Current vectors $\mathbf{s}_n$, $n = 1, \ldots, N$ are sampled from a multivariate standard normal distribution. These vectors are smoothed componentwise using a Gaussian spatial lowpass filter with standard deviation 2.5 cm. Denoting by p_k the k-th percentile of the current moment vector lengths $\|\mathbf{s}_n\|_2$, each $\mathbf{s}_n$ is re-scaled to a length of $\max(\|\mathbf{s}_n\|_2 - p_{90}, 0)$. That is, only the largest 10 % of the currents are retained. Source distributions obtained by this procedure typically exhibit a small number of active patches with small to medium extents and smoothly varying magnitudes and orientations (see Figure 4.3 (SIM) for an example).

Localization is carried out using LORETA, the minimum-current (MC) estimate and our proposed focal vectorfield reconstruction (FVR) and sparse basis field expansion (S-FLEX) techniques. These methods cover the full spectrum from smooth, spread-out, solutions (LORETA) to sparse solutions (MC). Note that other linear methods such as the weighted minimum-norm estimate

perform very similar to LORETA (Haufe et al., 2008). We use a variant of MC, in which the original depth compensation matrix is replaced by the matrix proposed in Section 4.1. As the data is simulated without noise, perfect reconstruction is required for all methods. Basis functions with standard deviations $\varsigma_1 = 0.5$ cm, $\varsigma_2 = 1$ cm, $\varsigma_3 = 1.5$ cm are used for S-FLEX. The tradeoff parameter α of FVR is set to 0.01.

We simulate five current densities and compute the respective pseudo EEG measurements. For each measurement and method, a 5×5-fold cross-validation is conducted. That is, the EEG electrodes are randomly partitioned into five groups of approximately equal size. Each union of four electrode groups gives rise to a "training set", while the remaining channel group is the "test set". The procedure is carried out five times with different randomizations, yielding 25 training sets with corresponding test sets. Inverse reconstructions are conducted based on the "training sets". In each of the 25 cross-validation runs, three criteria are evaluated. Most importantly the *reconstruction error*, defined as

$$\mathrm{REC} = \left\| \frac{\mathbf{s}}{\|\mathbf{s}\|_2} - \frac{\widehat{\mathbf{s}}^{\mathrm{tr}}}{\|\widehat{\mathbf{s}}^{\mathrm{tr}}\|_2} \right\|_2 , \tag{4.17}$$

is considered, where $\widehat{\mathbf{s}}^{\mathrm{tr}}$ are the vector field outputs estimated only from the training electrodes. Apart from this pointwise reconstruction criterion, we also consider the earth mover's distance between true and estimated current densities, which measures the effort needed to tranform one density into the other. A third quantity of interest is the *generalization error*, i. e., the error in predicting the activity at test electrodes from the sources that are estimated from the training electrodes. This is defined as

$$\mathrm{GEN} = \left\| \mathbf{x}^{\mathrm{te}} - A^{\mathrm{te}} \widehat{\mathbf{s}}^{\mathrm{tr}} \right\|_2^2 , \tag{4.18}$$

where $\mathbf{x}^{\mathrm{te}}$ and A^{te} are the parts of $\mathbf{x}$ and A belonging to the test set electrodes, respectively. Statistical significance of performance differences between methods are assessed by means of t-tests based on the estimated means and standard errors across experimental repetitions and cross-validation folds. These quantities are reported in Table 4.1.

4.4.2 Effect of joint localization

To illustrate the effect of co-localization, we perform the following experiment. A single dipolar source is placed in a cortical region of the brain and the resulting field pattern is computed. Ten different phase-shifted versions of the pattern are constructed by multiplying the original pattern with a random unit-length complex number $\exp(i\phi)$. Each of the resulting patterns is superimposed by equal amounts of measurement and biological noise. Note that in this scenario, the SNR cannot be increased by averaging, since both signal and noise are zero-mean complex-valued quantities. Source localization is carried out using both the single- and multiple-measurement variants of S-FLEX, where the regularization constant is set to match the exact SNR. The source estimates of all patterns are averaged to yield the estimated dipole amplitude per voxel. The obtained source amplitude map is compared to the true map, in which only one dipolar source is active, using the earth mover's distance. The experiment is repeated 100 times.

4.4.3 Results

Figure 4.3 shows a simulated current density along with reconstructions according to LORETA, MC, FVR and S-FLEX. Apparently, LORETA and MC do not approximate the true current density well. While the LORETA solution is too blurry to reflect the shape of the true sources, the MC estimate exhibits a large number of spikes, which could be misinterpreted as different sources. The FVR and S-FLEX estimates approximately recover the shape of the sources. The improvement of FVR and S-FLEX over LORETA and MC in terms of the current density reconstruction (REC) and the earth mover's distance (EMD) is statistically significant, as quantified in Table 4.1. Moreover, both FVR and S-FLEX generalize better than LORETA and the MC estimate, although insignificantly.

Joint- as compared to single-measurement localization leads to significantly better reconstruction of a single dipole, as indicated by a lower EMD (3.9 ± 0.1 as compared to 4.6 ± 0.1). An example is shown in Figure 4.2 (b). While both estimated source distributions peak similarly close to the true source location (indicated by a red cross-hair), the multiple-measurement approach has the advantage of being less spread-out. This demonstrates that joint localization effectively removes the noise-induced spatial instability seen in single-trial estimates.

	REC	EMD	GEN
LORETA	1.00 ± 0.01	2.43 ± 0.03	2.87 ± 0.78
MC	1.21 ± 0.01	1.91 ± 0.04	1.86 ± 0.57
FVR	0.95 ± 0.02	0.86 ± 0.02	1.21 ± 1.00
S-FLEX	**0.71 ± 0.04**	**0.75 ± 0.02**	0.92 ± 0.28

Table 4.1: Ability of low-resolution tomography (LORETA), the minimum-current (MC) estimate, focal vectorfield reconstruction (FVR) and sparse basis field expansions (S-FLEX) to recover simulated currents according to the pointwise reconstruction criterion (REC) and the earth mover's distance (EMD). Also shown is the generalization performance (GEN) with respect to the EEG measurements. Entries with significant superior score are highlighted.

4.5 Localization of N20 event-related potentials

4.5.1 Setting

In order to provide a real world example with known ground-truth, we recorded 113-channel EEG of one healthy subject (male, 26 years) during electrical median nerve stimulation. The EEG electrodes were placed according to the extended international 10-20 positioning system. The exact positions were obtained using a 3D digitizer and mapped onto the surface of the modeled skin shell. Electroencephalographic data were recorded with sampling frequency of 2 500 Hz and digitally bandpass-filtered between 15 Hz and 450 Hz. Left and right median nerves were stimulated in separate blocks by applying constant square 0.2 ms current pulses to the respective thenars. Current pulses had intensities above the motor threshold (approx. 9 mA), inducing twitches of the thumbs. The interstimulus interval varied randomly between 500 ms and 700 ms. About 1 100 trials were recorded for each hand.

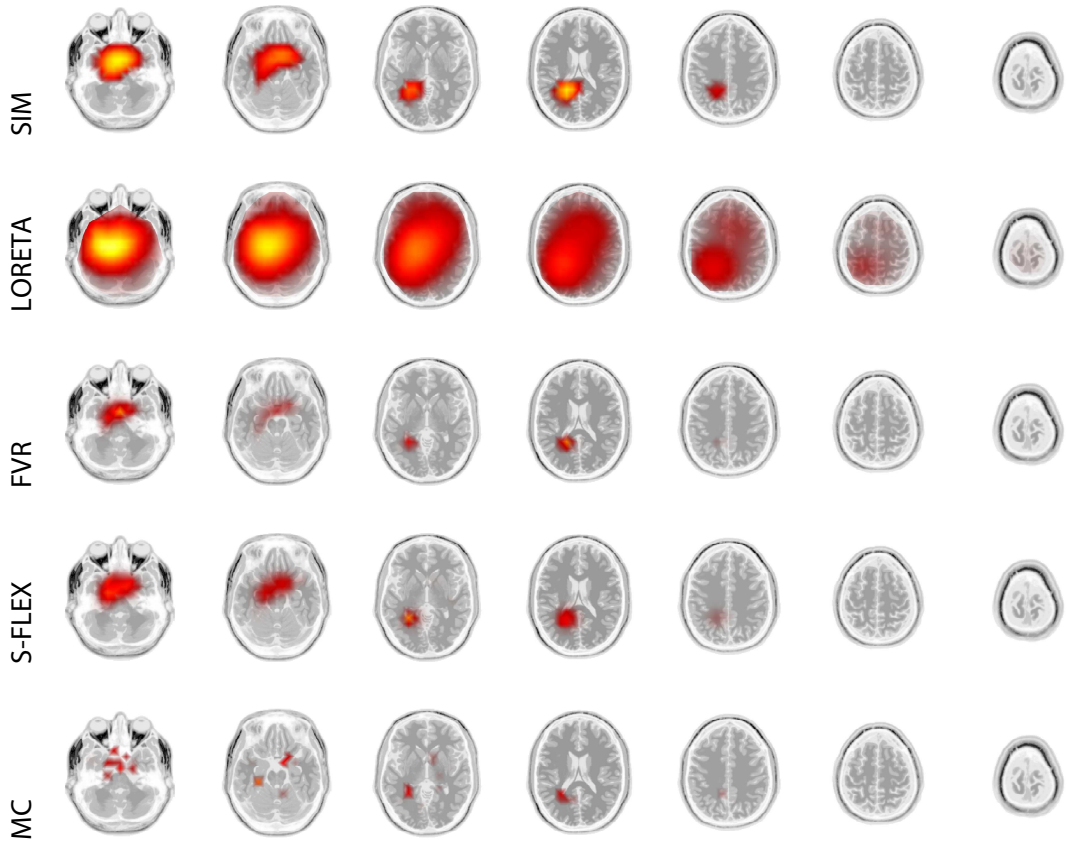


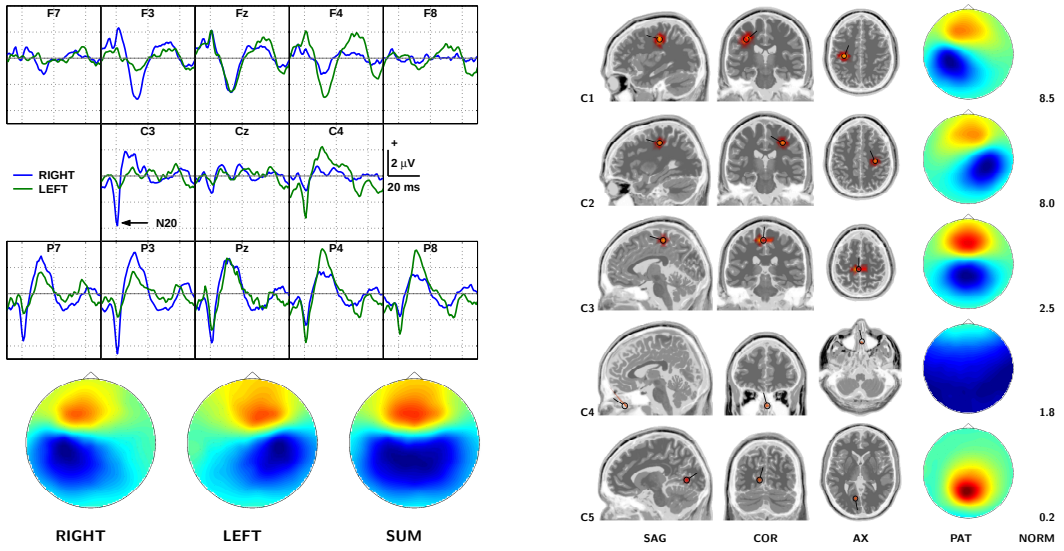
Figure 4.3: Simulated current density (SIM) and reconstructions according to low-resolution tomography (LORETA), the minimum-current estimate (MC), focal vectorfield reconstruction (FVR) and sparse basis field expansions (S-FLEX). Brighter colors encode larger dipole magnitudes.

Data analysis is conducted as follows. Trials and electrodes showing artifactual activity are excluded. For the remaining 1946 trials (973 left hand, 973 right hand), baseline correction is performed by subtracting the mean amplitude in the pre-stimulus interval from −100 ms to −10 ms. Finally, a single measurement vector is constructed by averaging the EEG amplitudes of all trials recorded in the same condition at 21 ms post-stimulus. By means of this, the N20 response to somatosensory stimulation at the hands is aggregated with high signal-to-noise ratio (SNR). Figure 4.4 (a) shows the average EEG time courses in both conditions as well as the spatial distribution of average EEG potentials at this point in time. The combined pattern is created by arithmetic summation of the patterns related to left and right hand stimulation.

We conduct inverse source reconstruction of the EEG field patterns related to single left and single right stimulation, as well as the summed pattern with the aim of revealing the distinct generators of the left and right N20 event-related components. The methods tested are LORETA, MC, FVR and S-FLEX. The regularization parameters of all four methods are adjusted to achieve

equal goodness-of-fit (GOF) scores. We use the estimated noise level to adjust the GOF. Its desired value is defined as $\|\sqrt{(973)}\boldsymbol{\sigma}_{\mathbf{x}}\|_2/\|\boldsymbol{\mu}_{\mathbf{x}}\|_2$, where $\boldsymbol{\mu}_{\mathbf{x}}$ and $\boldsymbol{\sigma}_{\mathbf{x}}$ are the electrode-wise empirical means and standard deviations of the N20 pattern as estimated from the 973 experimental repetitions.

When the presence of more than one source is assumed (as it is the case in the summed pattern, which containts responses to both left and right median nerve stimulation), an automatic decomposition of the estimated current density into individual source components is desirable. For the sparse solution case, such a decomposition is easily obtained by computing the connected components (with respect to the grid neighborhood relation) of the set of dipoles having nonzero estimated amplitude. Such analysis is exemplarily conducted for FVR here. Note that the specific properties of purely smooth inverse solutions (as, for example, provided by LORETA) prohibits connected component analysis of the sources, since there are no voxels with zero estimated activity. While a partitioning based on watersheds (Roerdink and Meijster, 2001) is generally conceivable, it is not applicable to the current experiment, in which the current amplitude distribution according to LORETA exhibits only one local maximum. The MC estimate consists of more than 20 components in this experiment.



(a) (b)

Figure 4.4: (a) Somatosensory evoked N20 potentials after left and right median nerve stimulation. Upper panel: average time series between 10 ms and 70 ms post-stimulus. Lower panel: Average scalp patterns at 21 ms post-stimulus and sum of left and right pattern. (b) Connected component analysis of the focal vectorfield reconstruction (FVR) inverse solution for the summed N20 pattern. Components are sorted from top to bottom according to the ℓ_2-norm (NORM) of their corresponding scalp patterns (PAT). For each source component, the average dipole amplitudes in sagittal (SAG), coronal (COR) and axial (AX) slices of 1 cm thickness around the source gravity center is shown (red and yellow colors, different scale for each component). Additionally, a single dipole (black) representing the mean orientation of the source is drawn at the gravity center.

4.5.2 Results

The estimated left and right N20 generators as deduced from inverse source reconstruction of the summed N20 field patterns are shown in Figure 4.5. The inverse solutions according to FVR and S-FLEX are almost indistinguishable. Both show activity concentrated in two major patches in contralateral central areas. This is in good agreement with the locations in which the hands are known to be represented within the brain, which are two distinct contralateral regions in the sensorimotor cortex (e. g., Komssi et al., 2004; Huttunen et al., 2006). Low-resolution tomography estimates only one active region, which is spread over the whole central area. The maximum of the current density lies exactly in between the hand areas. The MC estimate consists of several spikes scattered across the whole somatosensory area. Note that it is hardly possible to estimate the correct number of active sources from the MC solution.

The results of the connected components analysis of the FVR inverse solution for the summed N20 pattern is depicted in Figure 4.4 (b). The source distribution reveales five distinct sources, which are shown along with their individual scalp patterns as obtained from forward calculations. For each component a dipole representing the mean moment vector is drawn at the gravity center. The components are named C1 to C5 according to their decreasing impact on the EEG as measured by the ℓ_2-norm of their field patterns. The two strongest sources are the assumed N20 generators, the respective field patterns of which closely resemble the single-hand stimulation patterns shown in Figure 4.4 (a). The third component is likely to aggregate further signal parts belonging to C1 and C2, between which it is located. Judging from its frontal pattern, component C4 may correspond to eyeblink artifacts, which is in line with its estimated location close to the eyes. Component C5 has a negligible influence on the EEG measurement and may just represent non-stimulus-related brain activity. In summary, this analysis shows that estimated source distributions, which are smooth and sparse at the same time, can be easily decomposed into individual underlying sources based on purely spatial criteria. This may facilitate source connectivity analysis using such methods.

Our analysis also demonstrates that FVR and S-FLEX are capable of recovering multiple brain sources under real-world conditions. A further demonstration of the applicability of S-FLEX on real data is given in Haufe et al. (2011a), where S-FLEX is used to identify brain sites that differ in band-power between two motor imagery conditions of a brain-computer interface experiment. Here, the co-localization approach is used to map Fourier coefficients related to different frequencies within a band-of-interest to a common set of sources.

4.6 Reconstruction of simulated source connectivity using S-FLEX

4.6.1 Setting

While FVR is only applicable to individual EEG measurements, S-FLEX source estimates can be obtained for entire time series using the co-localization approach. Due to the fact that S-FLEX assumes a set of common sources, which are active throughout the whole recording, it is additionally well suited for the subsequent analysis of interactions between the sources time seris. To demonstrate this, we apply S-FLEX to the simulated dataset introduced in Section 3.9 to evaluate the weighted minimum-norm estimate. The experimental setup considered there involves two dipolar sources below C3 and C4 and information flow from the left to the right source.

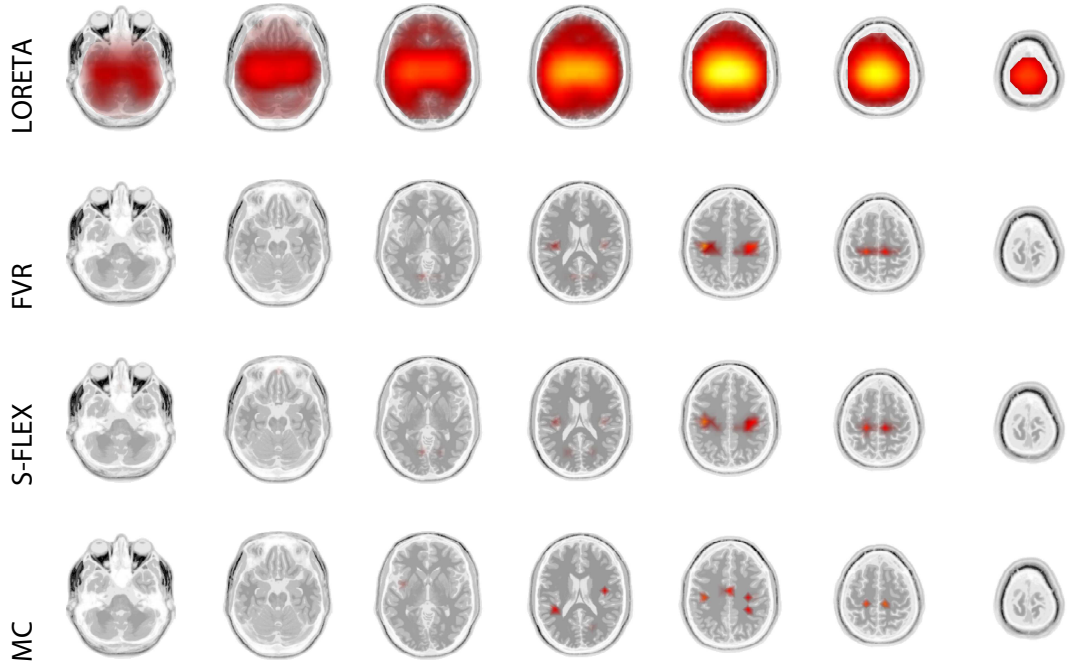


Figure 4.5: Localization of somatosensory evoked N20 generators according to low-resolution tomography (LORETA), the minimum-current (MC) estimate, focal vectorfield reconstruction (FVR) and sparse basis field expansions (S-FLEX). Brighter colors encode larger current magnitudes.

Using a linear inverse such as WMN, the source distribution is easily acquired even for time series data by means of a matrix multiplication. This is different for S-FLEX, which requires estimating all source variables (the coefficients of the basis field expansion related to all measurements) jointly using nonlinear optimization. In our simulation scenario, this amounts to estimating 192 780 000 source variables (coefficients related to $N = 2\,142$ voxels, $S = 3$ choices of ς, $T = 10\,000$ samples and 3 spatial dimensions) jointly based on 590 000 observations (related to $M = 59$ EEG electrodes and $T = 10\,000$ samples). This is prohibitive even for S-FLEX due to excessive memory requirements. Therefore, we adopt a two-step procedure, which restricts the number of variables involved in each step. In the first step, S-FLEX is applied to 100 randomly-selected samples. Using only the "active" basis functions characterized by nonzero estimated coefficients in this step, the second estimation is performed for all time samples. Typically, between 10 and 50 basis functions are selected as a result of the first step, which means that the number of variables in the second step is reduced to less than 1 million in most cases. We apply S-FLEX using Gaussian basis functions with spatial standard deviations $\varsigma_1 = 0.75$, $\varsigma_2 = 1$ and $\varsigma_3 = 1.25$. The regularization parameter in both steps is adjusted such that the S-FLEX solution achieves the same goodness-of-fit as the corresponding cross-validated WMN estimate. As in Section 3.9, 100 datasets are localized and subjected to subsequent connectivity analysis using Granger causality, partial directed coherence and the phase-slope index.

4.6.2 Results

The results of source connectivity analysis based on S-FLEX source estimates are presented in Figure 4.6, which is to be compared to Figure 3.5 depicting results of an equivalent analysis based on WMN inverse source estimates. The average estimated source activity shown in the upper panel of both figures clearly confirms the previously-noted superiority of S-FLEX in terms of localization accuracy and the ability to distinguish multiple sources. Unlike WMN, S-FLEX estimates two distinct prominent patches of strong activity. As can be seen in the source-amplitude map provided in the upper left panel of Figure 4.6, each of these patches is close to one of the true simulated interacting sources, although both are estimated slightly deeper than the true simulated ones. This problem, however, disappears if the SNR is increased by setting $\gamma = 0.75$ (results not shown here). The upper right panel of Figure 4.6 depicts the average activity in source-space regions-of-interest associated with the overlying EEG electrodes. As can be seen in this plot, the largest amount of activity falls into the ROIs associated with C_3 and C_4, which are exactly the electrodes under which the simulated sources are placed. All three measures of effective connectivity considered indicate significant information flow ($z > 4.0$) from the source under C_3 to the source under C_4, and no additional significant connections. This is the correct result, which could not have been obtained by WMN inverse source reconstruction preprocessing. Taken together, the results obtained here demonstrate that connectivity analysis of S-FLEX source estimates has numerous advantages over sensor-space analysis, as well as the analysis of source estimates provided by linear inverse methods. Using S-FLEX one way of avoiding neurophysiological misinterpretation, which is possible in sensor space due to the fact that the mixing patterns of the underlying sources are unknown, as is the contribution of the driving/receiving sources to the reference electrode. Furthermore, S-FLEX is the only preprocessing besides various BSS approaches that recovers sufficiently unmixed time courses of the underlying sources in our simulated example, and the only one besides TDSEP that succeeds in the case of Gaussian MVAR innovation terms. Note, however, that, while TDSEP achieves the correct demixing *despite* its inappropriate assumption of zero mutual time-lagged correlation of the sources, S-FLEX merely utilizes prior knowledge on the spatial separability of the sources. As the majority of inverse methods, S-FLEX is not affected by the distribution of the innovations of the source MVAR process. Rather, it assumes a Gaussian distribution of the *sources*' time courses, which is approximately true even for non-Gaussian innovations due to the temporal filtering induced by MVAR modeling. As a result of the successful demixing provided by S-FLEX, not only PSI but also GC and PDC yield correct results on the estimated sources.

4.7 Discussion (priors for the inverse reconstruction of multiple connected sources)

Computation of distributed EEG inverses heavily relies on regularization, since the physical model alone does not uniquely determine the sources. Within the last years the field has seen tremendous progress, in that the proposed regularization penalties increasingly represent neurophysiologically meaningful prior knowledge. In this chapter, we have subsequently derived an inverse methodology that properly addresses a number of remaining challenges, which are crucial with regard to the neurophysiological plausibility of the estimated sources as well as the feasibility to perform source

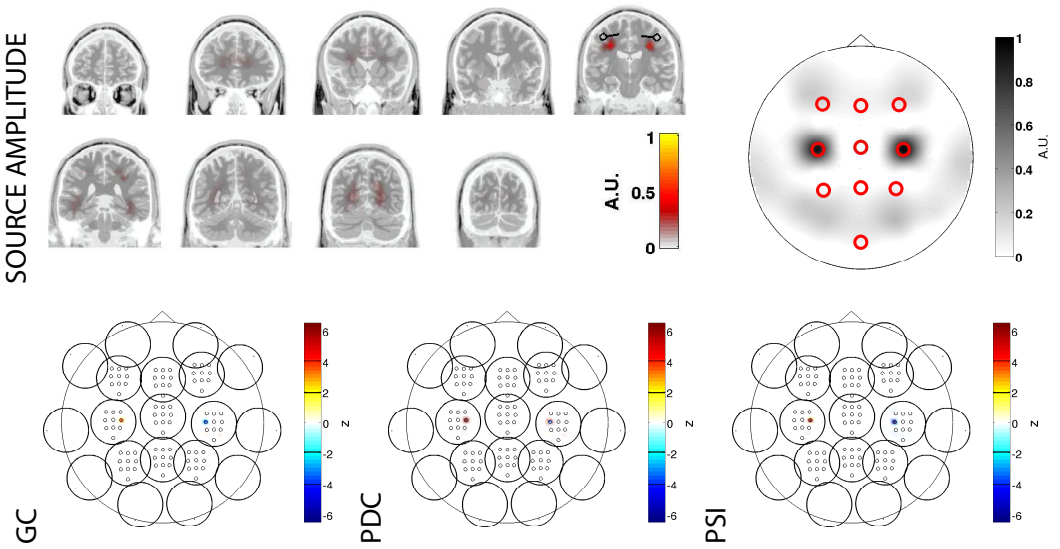


Figure 4.6: Comparison of effective connectivity of simulated EEG as estimated by Granger causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI) after application of sparse basis field expansion (S-FLEX) inverse source reconstruction. Two source dipoles with tangential orientations are modeled 3 cm below the C3/4 electrodes. Information flow from the left (C3) to the right (C4) source is modeled by means of a bivariate AR process. The simulated EEG is superimposed by non-interacting biological and sensor noise (SNR = 1). Upper left panel: average source strength (dipole amplitude) per voxel. Upper right panel: average source strength in regions-of-interest associated with the closest EEG electrode. Lower panel: Connectivity between regions-of-interest in source space, which are defined based on the nearest EEG electrode. The significance of estimated interactions between regions is measured in terms of z-scores and visualized as head-in-head plots, where red and yellow colors ($z > 0$) stand for information outflow and blue and cyan colors ($z < 0$) stand for information inflow. The Bonferroni-corrected significance level is indicated by a thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line.

connectivity estimation on the inverse solutions. An important step towards plausible source estimates has been the insight that the spatial spread of the sources should be controlled and ideally be tuned automatically. We propose two alternative ways here to achieve this. While focal vectorfield reconstruction uses a combination of smoothness- and sparsity-enforcing penalties, the relative influence of each of which may be optimized using cross-validation, inverse source reconstruction via sparse basis field expansions provides a built-in automatic scale selection using a large dictionary of suitable spatial basis functions. A general challenge when dealing with minimum-norm penalties is to balance the influence of a source's location on the regularized cost function. While this has been done previously using column-norm depth weighting, we propose a novel weighting matrix inspired by sLORETA, which has been found to improve localization results (Haufe et al., 2008). A further important factor influencing the plausibility of the estimated sources when using sparsity penalties is the proper handling of the vectorial nature of the dipole moments to be estimated. This is achieved here by means of grouped variable selection using the

$\ell_{1,2}$-norm penalty, which enjoys favorable invariance properties. Finally, if several interrelated field maps are to be localized by sparse estimators, their mutual information should be exploited. Here we suggest suitably enlarging the groups of source variables that are tied under a common ℓ_2-norm penalty, which amounts to modeling that the sets of active brain sites have substantial overlap across EEG patterns. This assumption is certainly fulfilled for repeated measurements recorded under the same experimental condition. The same holds true for time series data that are to be localized for the purpose of source connectivity analysis. Here, spatial consistency is important for achieving stationary behaviour of the source time series, which is a precondition for connectivity analysis.

In addition to the development of inverse methods, we propose the earth mover's distance as a metric for comparing arbitrary source estimates that drastically facilitates the evaluation of inverse solutions. Using the EMD between simulated and estimated sources as a performance measure, the superior source localization accuracy of our approaches in comparison to the state-of-the-art was shown in a simulated setting (Section 4.4), as well as on real-world event-related potential (Section 4.5) and event-related desynchronization Haufe et al. (2010b, 2011a) data. We also investigated the benefit of our improved source modeling for EEG source connectivity estimation. Interestingly, it turns out that the application of S-FLEX preprocessing yields a substantial qualitative improvement over standard approaches.

5 Blind recovery of sources and their connectivity

While inverse source reconstruction is a powerful approach to the estimation of underlying brain sources from EEG measurements, it relies on the availability of a physical model of the head. Once this head model is provided, inverse source reconstruction is concerned about inverting the model equations. In the case of distributed inverse imaging, this boils down to solving the underdetermined linear system $X = AS$ for the sources S under physiologically-motivated constraints.

Blind source separation (BSS) techniques are not restricted to a specific application domain, for which the forward equations must be known. Rather, their idea is to simultaneously estimate the mixing matrix A and the source time series S of the decomposition. Estimating the lead field matrix in the EEG context is necessary when no forward model is known, but it can also be beneficial when the available model is inaccurate (e. g., not matching the individual head). The increased freedom that comes with estimating A, however, makes parameter estimation mathematically more demanding. In general, A and S are not identifiable without contraints, which have to be chosen carefully in order to reflect the assumed properties of the unknown true factors in the targeted application domain. Inappropriate constraints may easily lead to inexact factorizations. For example, orthogonality constraints imposed by PCA on the columns of A prevent physiologically meaningful EEG sources (except for, perhaps, the strongest source) from being found, because orthogonal mixing patterns are incompatible with the physical process of EEG generation. A further example is independent component analysis, which makes the assumption of independence of the source times series, which is inappropriate when investigating source connectivity.

In this chapter, we develop blind source separation techniques that are, unlike PCA and ICA, tailored to the discovery of mutually interacting sources and thereby compete with methods like MVARICA and CICAAR. In accordance with the rest of this thesis, we here assume that interactions are characterized by time-lagged influences of the driving on the receiving sources. Our key to indentifying these sources and their spatial mixing patterns is the assumption that the source time series follow a multivariate AR process. As has been pointed out previously, MVAR models are the foundation of Granger-causal measures of effective connectivity. Our previous simulations have shown that such analyses are seriously hampered by insufficiently demixed EEG signals, while they can be sucessful if the sources are well separated. The idea pursued in this chapter is to integrate the assumption of potential interactions between sources directly into a BSS methodology, which amounts to taking the opposite approach as in ICA.

The chapter starts with a section on sparse Granger-causal modeling (see Section 5.1). Here we outline two approaches to obtaining sparse Granger-causal graphs, which are shown to outperform existing approaches in terms of detecting significant interactions. In Section 5.2, we extend these approaches by an additional mixing matrix, which gives rise to a methodology for the blind source separation of EEG signals that decomposes the data into correlated sources following a multivariate autoregressive model. Two variants are discussed. Connected components analysis (CSA) imposes

no restriction on the MVAR coefficients. This leads to a model that is equivalent to a convolutive ICA model of the source MVAR innovation terms. Motivated by the effectiveness of sparse Granger-causal modeling that is demonstrated in the first section, sparsely-connected sources analysis (SCSA) employs additional $\ell_{1,2}$-norm regularization of the source MVAR coefficients in order to effectively prune connections between sources. Since the regularization parameter can be adapted to the data, SCSA is able to model any degree of source interaction comprising global interaction (as assumed in CSA) as well as completely independent sources (in the ICA sense) as special cases. While CSA and SCSA do not constrain the mixing matrix A in any way, this is done in a more general approach, which is outlined in Section 5.3. Here, we utilize a physical model to obtain a factorization of A into the known lead field matrix and a matrix of source current distributions. Physiological plausability of the mixing patterns is ensured by enforcing that the associated source current distributions have simple spatial signatures in terms of the sparsity of a decomposition into spatial basis fields. Thus, the outlined approach, SCSA+FLEX, fuses SCSA blind source separation with S-FLEX inverse source reconstruction and thereby combines the advantages of all methods developed in the course of this thesis. While the evaluation of SCSA+FLEX is still ongoing, an application of SCSA to the simulated interacting sources example introduced in Chapter 3 is presented in Section 5.4.

5.1 Methods for sparse Granger-causal discovery

In this section, we are not yet concerned with blind source separation. Rather, we aim to improve existing Granger-causal methods, which would be applicable on the source level. As mentioned previously, the fundamental idea underlying Granger causality is that a time series z_j "causally" influences another time series z_i, if knowledge of the past of z_j improves the prediction of the presence of z_i, compared to knowing only the past of z_i. This is assessed by the conventional Granger score, which compares the residual errors of two multivariate AR models (see Section 2.6). However, a causal dependence of z_j on z_i in the Granger sense is already evidenced, if any of the P coefficients $b_{i,j}(p)$ of a single full MVAR model is significantly different from zero. This observation is exploited, for example, by the directed transfer function (DTF) and by partial directed coherence (PDC).

Testing whether MVAR coefficients or (net) PDC/DTF scores are significantly different from zero can be conducted by fitting MVAR models to multiple i.i.d. data chunks and using standard statistical tests. This is the general approach pursued in this thesis. However, when only few data points are available, it is desirable to estimate the presence of effective connections from a single MVAR model. In this section, we describe two approaches for doing so, which are adopted from Haufe et al. (2010a). The first approach uses ridge regression for MVAR estimation. To assess the statistical significance of the estimated coefficients, we here apply the post-hoc multiple statistical testing procedure of Hothorn et al. (2008), which saves us from having to acquire multiple repeated measurements. The second approach is based on using sparse regularization for direct estimation of sparse connectivity graphs. Our approach here accounts for the fact that the absence of a causal relation between z_i and z_j requires all AR coefficients belonging to that pair of time series to be jointly zero, which can be accomplished using an $\ell_{1,2}$-norm penalty. In the following, we provide the details regarding the groupwise sparsity and the alternative testing strategy, respectively.

5.1.1 Ridge regression with post-hoc statistical testing

Under the assumption of Gaussian innovations and additional noise, it is natural to estimate AR coefficients using a regularized least squares approach. Probably the most straightforward way to do so is to use ridge regression, which adds an ℓ_2-norm penalty to the negative log-likelihood. Adopting the notation of Section 2.5, the multivariate AR model reads $\mathbf{x}(t) = \sum_{p=1}^{P} B(p)\mathbf{x}(t-p) + \boldsymbol{\varepsilon}(t)$. Setting $\widetilde{B} = (B(1), \ldots, B(P))$, $X = (\mathbf{x}(P+1), \ldots, \mathbf{x}(T))$, $\widetilde{X} = (X_1, \ldots, X_P)^\top$ and $X_p = (\mathbf{x}(P+1-p), \ldots, \mathbf{x}(T-p))^\top$, the ridge regression estimate is given by

$$\begin{aligned} \widehat{\widetilde{B}}^{\text{ridge}} &= \arg\min_{\widetilde{B}} \left\| \mathbf{vec}\left(X - \widetilde{B}\widetilde{X}\right) \right\|_2^2 + \lambda \left\| \mathbf{vec}\left(\widetilde{B}\right) \right\|_2^2 \\ &= (\widetilde{B}\widetilde{B}^\top + \lambda I_{MP})^{-1} \widetilde{B} X^\top \,, \end{aligned} \tag{5.1}$$

where $\lambda \geq 0$. Due to the ℓ_2-norm penalty, (5.1) delivers solutions with small coefficients, which are, however, in general never exactly zero. In the strict sense of Granger, this corresponds to a fully-connected effective connectivity graph, rendering a single ridge regression incapable of selecting a subset of interacting source pairs. However, sparsification can be performed by means of statistical testing, which returns only the significant connections. While conventional testing procedures require multiple independently-acquired measurements, our hereby proposed approach is capable of deriving p-values directly from a single MVAR estimate. This is possible using the linearity property of the ridge regression estimator. From (5.1), it is apparent that the MVAR estimation can be carried out independently for each row of $\widetilde{B}$, and so can the testing. Let therefore $\boldsymbol{\beta}_k$ denote the k-th row of $\widetilde{B}$ and let $\mathbf{x}_k = (x_k(P+1), \ldots, x_k(T))^\top$. As the ridge regression coefficients depend linearly on X, we can conclude that under the null-hypothesis $H_0 : \boldsymbol{\beta}_k = 0$ the estimated coefficients are distributed according to

$$\widehat{\boldsymbol{\beta}}_k \sim \mathcal{N}(\mathbf{0}, \varsigma_k^2 \Sigma) \quad \text{with} \quad \Sigma = (\widetilde{X}\widetilde{X}^\top + \lambda I_{MP})^{-1} \widetilde{X}\widetilde{X}^\top (\widetilde{X}\widetilde{X}^\top + \lambda I_{MP})^{-1} \,. \tag{5.2}$$

Furthermore, setting $H = \widetilde{X}^\top (\widetilde{X}\widetilde{X}^\top + \lambda I_{MP})^{-1} \widetilde{X}$, an estimate of the model variance ς_k^2 is given by

$$\widehat{\varsigma}_k^2 = \frac{\|\mathbf{y}_k - H\mathbf{y}_k\|^2}{\mathrm{Tr}\{(I_{T-P} - H)\,(I_{T-P} - H^\top)\}} \,. \tag{5.3}$$

Using (5.3), we can construct normalized test statistics $\widetilde{\beta}_{i,k} = \widehat{\beta}_{i,k} / \sqrt{(\varsigma_k^2 \sigma_{i,i})}$, which are jointly normally distributed with $\widetilde{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{0}, R)$ and $r_{i,j} := \sigma_{i,j} / \sqrt{(\sigma_{i,i}\sigma_{j,j})}$. Suppose we want to test all individual hypotheses $H_{0;i,k} : \beta_{i,k} = 0$ simultaneously, then, according to Hothorn et al. (2008), the adjusted p-values are $p_{i,k} = 1 - g(R, |\widetilde{\beta}_{i,k}|)$. Here,

$$g(R,t) = P\left(\max_i |\widetilde{\beta}_{i,k}| \leq t\right) = \int_{-t}^{t} \ldots \int_{-t}^{t} \phi(\beta_{i,1}, \ldots, \beta_{i,MP})\, \mathrm{d}\beta_{i,1} \cdots \mathrm{d}\beta_{i,MP} \tag{5.4}$$

and $\phi(\boldsymbol{\beta})$ is the probability density function of the multivariate normal distribution $\mathcal{N}(\mathbf{0}, R)$. We reject a hypothesis, if the p-value is below the predefined significance level γ.

5.1.2 Groupwise sparsity approach

Sparse approaches to MVAR estimation provide a direct way of assessing the presence of time-lagged effective connectivity without having to conduct additional statistical testing. At the same time, sparsity acts as a regularizer when it is already enforced within the estimation. By this means the model may be prevented from overfitting the data. In contrast to this, the ridge-regression-based approach presented above is a two-step procedure, which may possibly suffer from the aggregation of assumptions made in each step. Along these lines, Valdés-Sosa et al. (2005) propose estimating sparse MVAR models of order $P = 1$ using ℓ_1-norm regularization of the AR coefficients. The resulting model characterizes time series with sparse Granger-causal structure. However, for AR models of order $P > 1$, the structure of the connectivity graph estimated this way may be different for each time lag $p \in \{1, \ldots, P\}$.

Here we propose a novel sparse approach which accounts for the fact that the absence of a causal relation between x_i and x_j requires *all* AR coefficients $b_{i,j}(1), \ldots, b_{i,j}(P)$ related to that certain pair of time series to be *jointly zero*. This corresponds to introducing the prior belief that causal influences between time series is not restricted to only one particular time lag. Technically, the proposed sparsity scheme is implemented using the $\ell_{1,2}$-norm regularizer, i. e., by introducing the penalty $\lambda \sum_{i \neq j} \| (b_{i,j}(1), \ldots, b_{i,j}(P))^\top \|_2$, which groups all coefficients related to the pair (i, j) of time series together, such that they may only be jointly pruned. It is also possible to add an extra term $\| (b_{1,1}(1), \ldots, b_{M,M}(P)) \|_2$ for regularizing the diagonal (auto-regressive) coefficients. The latter variant leads to the estimate

$$\begin{aligned} \widehat{\widetilde{B}}^{\mathrm{GL}} \;=\; & \arg\min_{\widetilde{B}} \left\| \mathbf{vec}\left(X - \widetilde{B}\widetilde{X}\right) \right\|_2^2 \\ & + \lambda \left(\| (b_{1,1}(1), \ldots, b_{M,M}(P)) \|_2 + \sum_{i \neq j} \left\| \left(b_{i,j}(1), \ldots, b_{i,j}(P)\right) \right\|_2 \right) . \end{aligned} \tag{5.5}$$

This defines a standard group lasso problem, which can be solved using a variety of algorithms (Sturm, 1999; Roth and Fischer, 2008; Tomioka and Sugiyama, 2009). It is moreover possible to split the estimation of B into M subproblems (as performed in the combined ridge regression/statistical testing approach), which is beneficial in large-scale scenarios.

Based on the group-lasso-regularized estimate, we define the effective connectivity graph according to *group lasso Granger causality* (GLGC) as

$$\mathrm{glgc}_{i,j} = \left\| \left(b_{i,j}(1), \ldots, b_{i,j}(P)\right)^\top \right\|_2 . \tag{5.6}$$

While this is the easiest way to define the presence of interactions based on the $\ell_{1,2}$-norm regularized MVAR coefficients, it is also possible to calculate partial directed coherence, the directed transfer function and/or statistics on top of the sparse MVAR coefficients. In this regard, it is of interest to note that the sparsity pattern of the GLGC connectivity graph is inherited by PDC, but not by DTF, which is generally non-sparse.

5.1.3 Application to simulated data

Setting

We conduct a series of experiments, in which the effective connectivity of simulated data representing *unmixed* brain source activity is to be recovered. For each experiment, we simulate a multivariate time series with parameters $M = 7$ and $T = 1000$, which is generated by a random MVAR process of order $P = 5$. The innovations terms $\boldsymbol{\varepsilon}(t)$ are drawn from the standard normal distribution. We simulate ten interactions by setting the coefficients for all but ten randomly chosen pairs of time series to zero. The non-zero coefficients are drawn independently from $\mathcal{N}(0, 0.04)$. Each set of MVAR coefficients is tested for the stability of its induced dynamical system by looking at the eigenvalues of the corresponding transition matrix. Only coefficients leading to stable systems are accepted. We consider the following three types of problems, for each of which we create ten instances: i) no noise is added to the data generated by the VAR model, ii) the data is superimposed by Gaussian noise of approximately the same strength, which is uncorrelated (white) both across time points and sensors and iii) the data is superimposed by mixed noise of approximately the same strength, which is generated as a random instantaneous mixture of M stable univariate AR processes of order 20 with random coefficients. Note that in none of these cases the noise time series contains time-lagged interaction that would superimpose the true interaction structure.

For measuring the performance of the Granger-causal discovery, we consider receiver operating characteristics (ROC) curves, which objectively assess the performance according to different regimes. More precisely, the ROC curve plots the sensitivity (true positive rate) of a predictor as a function of the specifity (true negative rate). As an additional measure of absolute performance we also calculate the area under the ROC curve (AUC). We compute average ROC curves and AUC scores across the ten problem instances, and standard errors for the AUC score. We compare the proposed groupwise sparse (group lasso) and ridge regression with multiple testing approaches with ℓ_1-norm regularization (lasso) and conventional Granger causality. All four approaches are applied both with and without knowledge of the true model order. In the latter case, $P = 10$ is chosen for the reconstruction. For all methods considered, it would also be possible to estimate the model order, e. g., via cross-validation.

We use the solver provided by Roth and Fischer (2008) for computing the lasso and group lasso estimates and our own implementation for computing Granger causality as well as performing ridge regression and multiple testing. The MVAR models used in Granger causality are fitted using the BIOSIG toolbox (Vidaurre et al., 2011)[1]. The regularization parameter λ in ridge regression is chosen via ten-fold cross-validation. For this value of λ, we derive the test statistics $\beta_{i,k}$. The multidimensional integrals in (5.4) are computed using Monte Carlo sampling according to Genz (1992)[2]. Receiver operating characteristics curves are constructed by varying the significance level γ. Similarly, for Granger causality, ROC curves are obtained by thresholding the Granger score at different values. For lasso and group lasso, solutions ranging from completely sparse to completely dense are obtained through variation of the regularizing constant λ.

[1] http://biosig.sourceforge.net/
[2] http://www.math.wsu.edu/faculty/genz/software/matlab/qsimvnv.m

Results

Table 5.1 summarizes the AUC scores obtained in the experiments described above. The complementing ROC curves are shown in Figure 5.1. In short, the group lasso and ridge regression approaches outperform their competitors in all scenarios, although not always significantly. While ridge regression with additional testing performs slightly better than the group lasso in the noiseless condition, group lasso has a clearly visible yet insignificant advantage over all methods in the white noise setting. Under the influence of mixed noise, ridge regression and group lasso are on par. Note that the ROC curve for lasso is strictly below the ROC curve of group lasso, which demonstrates that connectivity graphs estimated by lasso tend to be too dense. Interestingly, knowledge of the true model order hardly provides any significant advantage in our simulations.

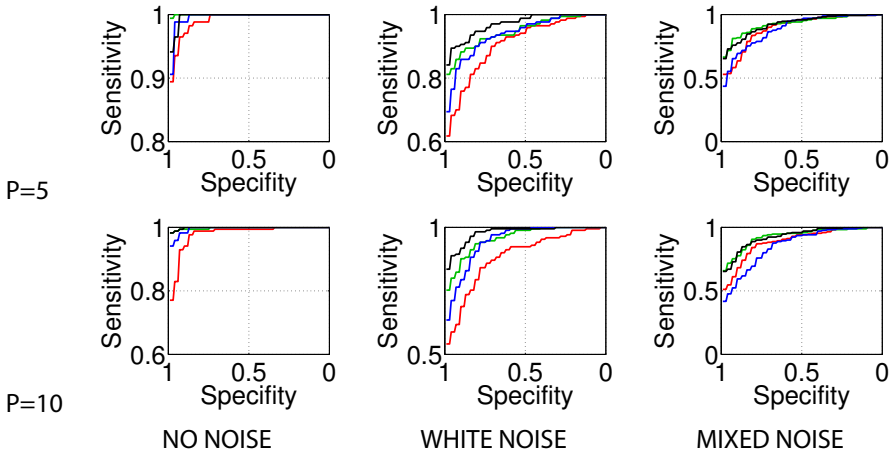


Figure 5.1: Average receiver operating characteristics (ROC) curves of Granger causality (red), ridge regression in combination with multiple statistical testing (green), lasso (blue) and group lasso (black) in three different noise conditions and for two different model orders.

		GC	RIDGE	LASSO	GLASSO
P=5	NO NOISE	0.991 ± 0.004	**1.000 ± 0.000**	0.996 ± 0.002	0.997 ± 0.002
	WHITE NOISE	0.910 ± 0.023	0.948 ± 0.020	0.941 ± 0.021	0.971 ± 0.016
	MIXED NOISE	0.896 ± 0.012	**0.928 ± 0.010**	0.889 ± 0.011	**0.926 ± 0.012**
P=10	NO NOISE	0.980 ± 0.005	**0.998 ± 0.002**	**0.996 ± 0.002**	**0.999 ± 0.001**
	WHITE NOISE	0.885 ± 0.019	0.958 ± 0.012	0.948 ± 0.013	**0.979 ± 0.005**
	MIXED NOISE	0.893 ± 0.013	**0.931 ± 0.015**	0.861 ± 0.014	**0.931 ± 0.007**

Table 5.1: Average area under the receiver operating characteristics curve (AUC) scores and standard errors of Granger causality, ridge regression with multiple statistical testing, lasso and group lasso in three different noise conditions and for two different model orders. Entries with significant superior AUC score are highlighted.

5.2 Sparsely-connected sources analysis (SCSA)

In Chapter 3, we presented simulations demonstrating that the estimation of source connectivity from EEG data using partial directed coherence and conventional Granger causality is severely hampered if performed on insufficiently demixed EEG time series or even directly on sensor data. The same must be expected for the novel sparse connectivity approaches introduced in the previous section, which are similarly grounded on MVAR estimation. While these new approaches have been found to outperform conventional Granger causality and ℓ_1-norm regularized MVAR estimation in terms of detecting the Granger-causal structure of a multivariate time series, this result has only been achieved on the "source level", where no mixing of the interacting signals is present. Such a mixing must, however, be expected in any real EEG dataset due to volume conduction in the head.

In this section, we aim at utilizing sparse Granger-causal approaches for EEG source connectivity analysis without having to worry about a suitable preprocessing. To this end, we present a blind source separation technique called sparsely-connected sources analysis (SCSA), by which the source time series and their respective field patterns are jointly estimated under the assumption of sparsity of the source connectivity graph. To do so, we extend the previously described $\ell_{1,2}$-norm regularized approach to MVAR estimation by an additional instantaneous mixing, which represents volume conduction. Note that this is substantially different to merely applying a sparse connectivity estimator to demixed time series, because the information that (some of) the demixed sources are interacting is accounted for in the estimation of the mixing parameters. Sparsely-connected sources analysis was introduced in Haufe et al. (2010c), along with an unregularized variant called connected sources analysis (CSA). In the following, we review the details of CSA and SCSA from Haufe et al. (2010c). Starting from the formulation of the correlated sources model underlying our methods, we go on to discuss theoretical and practical issues of parameter identification, as well as optimization strategies and the relation to other BSS approaches. The section ends with an exhaustive empirical evaluation of BSS methods in terms of source effective connectivity estimation.

5.2.1 Correlated sources model

We consider a noiseless version of the general EEG model (2.2), which assumes that the EEG sensor measurement is generated as a linear instantaneous mixture of source activity. The source dynamics is assumed to be generated by an MVAR process. This leads to the following model of interconnected sources

$$\mathbf{x}(t) = A\mathbf{s}(t) \tag{5.7}$$

$$\mathbf{s}(t) = \sum_{p=1}^{P} B(p)\mathbf{s}(t-p) + \boldsymbol{\varepsilon}(t)\ . \tag{5.8}$$

Here, $\mathbf{x}(t)$ is the M-dimensional EEG signal at time t, A is a $M \times M$ mixing matrix representing the volume conduction effect and $\mathbf{s}(t)$ is the demixed (source) signal. The sources at time t are modeled as a linear combination of their P past values plus an innovation term $\boldsymbol{\varepsilon}(t)$, according to an MVAR model with coefficient matrices $B(p)$. For simplicity, we deal with the case that the

numbers of sensors and sources are equal and the mixing matrix A is invertible. When there exist less sources than sensors, the problem falls into the current setting after being preprocessed using principal component analysis. The innovation sequence can be obtained by finite impulse response (FIR) filtering of the observation, i. e.,

$$\begin{aligned} \boldsymbol{\varepsilon}(t) &= A^{-1}\mathbf{x}(t) - \sum_{p=1}^{P} B(p)A^{-1}\mathbf{x}(t-p) \\ &= \sum_{p=0}^{P} W(p)\mathbf{x}(t-p)\,, \end{aligned} \tag{5.9}$$

where the filter coefficients are determined by the mixing matrix A and the MVAR parameters $\{B(p)\}$ according to

$$W(p) = \begin{cases} A^{-1} & p = 0 \\ -B(p)A^{-1} & p > 0 \end{cases}\,. \tag{5.10}$$

Note that (5.9) corresponds to a noiseless version of the convolutive ICA model. However, compared to the standard formulation (see Section 2.8), measurements and underlying convolutive sources (innovations) are interchanged here.

5.2.2 Identification by convolutive ICA

In conventional MVAR analysis, the innovation term $\boldsymbol{\varepsilon}(t)$ is a temporally- and spatially-uncorrelated sequence of standard normally distributed vectors. In contrast, we assume here that the components are i.i.d. and subject to non-Gaussian distributions. Under these weak assumptions, the model parameters A and $\{B(p)\}$ can be identified using standard approaches to temporal-domain convolutive ICA. For EEG data, a super-Gaussian is more suitable than a sub-Gaussian distribution, if we assume that ongoing activity of brain networks is triggered by spontaneous local bursts. We here adopt the super-Gaussian hyperbolic secant (sech) distribution that is proposed in Dyrholm et al. (2007). The likelihood of the data under this model is

$$p(\{\mathbf{x}(t)\}_{t=P+1}^{T}|\{W(p)\}) = |W(0)|^{T-P} \prod_{t=P+1}^{T}\prod_{m=1}^{M} \tfrac{1}{\pi}\operatorname{sech}\left(\varepsilon_m(t)\right)\,, \tag{5.11}$$

where T is the number of available time samples. The cost function to be minimized is the negative log-likelihood, which is given by

$$\mathcal{L}^{\mathrm{CSA}}(\{W(p)\}) = (P-T)\log|W(0)| = -\sum_{t=P+1}^{T}\sum_{m=1}^{M}\log\left(\tfrac{1}{\pi}\operatorname{sech}\left(\varepsilon_m(t)\right)\right)\,. \tag{5.12}$$

The minimization of (5.12) leads to estimators $\widehat{A}$ and $\{\widehat{B}(p)\}$ of the mixing matrix and the MVAR coefficients, respectively, via (5.10). We call this procedure *connected sources analysis* (CSA).

5.2.3 Sparse connectivity as regularization

In practice, the number of parameters in CSA may become very large, because the number of AR coefficients grows quadratically with the number of EEG sensors (or PCA components). Maximum-likelihood estimation may lead to overfitting, especially if the number of observations T is small. For this reason, it is advisable to adopt a regularization scheme. In the previous section, we have pointed out that, by using an $\ell_{1,2}$-norm penalty, entire effective connections between time series can be pruned at once. From the practical standpoint this is very appealing, since fewer connections are far easier to interpret. But assuming sparse connectivity is also justified by studies of the numerical characteristics of network connectivity in fMRI data (see Valdés-Sosa et al., 2005, and the references therein). This reasoning also applies to EEG data. Besides the penalty-based approach, post-hoc sparsification of dense esimates by means of statistical testing is another viable option (see our ridge regression based approach presented in Section 5.1). However, due to the compelling built-in regularization, we here adopt $\ell_{1,2}$-norm sparsification analogous to the group lasso approach to Granger-causal discovery described in Section 5.1.

Sparsity can only reasonably be assumed for the MVAR coefficients $\{B(p)\}$ but not for the $W(p)$ matrices which factorize into MVAR coefficients and the (necessarily non-sparse) instantaneous demixing. Hence, in order to apply $\ell_{1,2}$-norm regularization, we have to split the parameters into demixing and MVAR parts, as per the original model given by (5.7) and (5.8). Defining $\mathbf{s}(t) = A^{-1}\mathbf{x}(t)$ and $\widetilde{\mathbf{s}}(t) = \sum_{p=1}^{P} B(p)\mathbf{s}(t-p)$, the regularized cost function reads

$$\begin{aligned}\mathcal{L}^{\mathrm{SCSA}}(A, \{B(p)\}) \quad = \quad & (P-T)\log|A| - \sum_{t=P+1}^{T}\sum_{m=1}^{M} \log\left(\tfrac{1}{\pi}\operatorname{sech}\left(s_m(t) - \widetilde{s}_m(t)\right)\right) \\ & + \lambda \Bigg(\left\| \left(b_{1,1}(1), \ldots, \ldots, b_{M,M}(P)\right)^{\top} \right\|_2 \\ & + \sum_{m \neq f} \left\| \left(b_{m,f}(1), \ldots, b_{m,f}(P)\right)^{\top} \right\|_2 \Bigg) \,, \end{aligned} \tag{5.13}$$

λ being a positive constant. The minimizer of (5.13) for a choice of λ is called the *sparsely-connected sources analysis* (SCSA) estimate.

5.2.4 Relation to other BSS methods

Sparsely-connected sources analysis extends our previously suggested group lasso approach to sparse Granger-causal discovery (see Section 5.1) by a linear demixing. By doing so, we obtain a method that is – unlike all previous Granger-causal approaches discussed in this thesis – suited for EEG-based connectivity analysis. Our method compares with MVARICA (i. e., MVAR+ICA, Gómez-Herrero et al., 2008), which assumes the same model but estimates its parameters differently. Precisely, the authors of MVARICA suggest initially fitting an MVAR model in sensor space. The demixing is then obtained by performing instantaneous ICA on the MVAR innovations, i. e., a dedicated contrast function is used to model independence of the innovations. The obtained sources follow an MVAR model with time-lagged effects (interactions), as in CSA/SCSA.

As mentioned, our model is very similar to the convolutive ICA model, the only difference being that (5.9) employs a finite impulse response (FIR) filter to extract the innovations, while

an infinite response filter (IIR) is usually used in the CICA literature (see, e. g., Dyrholm et al., 2007). This discrepancy is explained by the different philosophies that are associated with both methods. While in our approach, the innovations $\boldsymbol{\varepsilon}(t)$ arise as residuals of a finite-length source-MVAR model, CICA understands them as sources of a finite-length convolutional (forward) mixture. Nevertheless, our unregularized cost function can be regarded as a maximum-likelihood approach to an IIR version of convolutive ICA. This leads to a new view of convolutive ICA as performing an instantaneous demixing into correlated sources. Hence, it is possible to conduct source connectivity analysis using CICA.

Compared to MVARICA and time-domain implementations of convolutive ICA such as CICAAR, our formulation has the advantage that sparse connectivity can easily be modeled by an additional penalty. This is not possible for CICAAR, because CICAAR only indirectly estimates the MVAR coefficients through their inverse filters. However, these are generally nonsparse, even if the true connectivity structure is sparse. Inverting the inverse coefficients is also generally not possible (recall that convolutive ICA is equivalent to an infinite-length source-MVAR model). It is furthermore not possible to introduce a sparse regularization for MVARICA, since this method carries out the MVAR-estimation step in sensor space, where no sparsity can be assumed.

By variation of the regularization parameter, our method is able to cover all possibilities between the extremes of a fully-correlated source model (similar to convolutive ICA) and a model which allows no cross-talk between sources. Interestingly, the latter extreme can be seen as a variant of traditional instantaneous ICA, in which independence is measured in terms of mutual predictability with a Granger-type criterion.

5.2.5 Optimization

The minimizations of (5.12) and (5.13) cannot be cast in the framework of convex optimization for instance using the group lasso or ridge regression, for which reason we here give detailed descriptions of our optimization procedures.

CSA

The gradient of the unregularized (CSA) cost function (5.12) is obtained as

$$\begin{aligned}\frac{\partial \mathcal{L}^{\mathrm{CSA}}}{\partial W_m(p)} &= \delta(p)\left((P-T)W(p)^{-\top}\mathbf{e}_m\right) \\ &\quad + \sum_{t=P+1}^{T} \tanh\left(\sum_{p=0}^{P} W_m(p)^{\top}\mathbf{x}(t-p)\right)\mathbf{x}(t-p)\,, \end{aligned} \tag{5.14}$$

where $W_m(p) := W(p)^{\top}\mathbf{e}_m$, i. e., the m-th column vector of $W(p)^{\top}$. Using (5.14), CSA can be readily solved using limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization. Note that the CSA objective has obvious indeterminacies due to permutations and sign flips. However, once we fix a rule to chose one from all candidates, the cost function can be considered convex.

SCSA via a modified L-BFGS algorithm

Using sparse regularization, two difficulties emerge in contrast to when using the unregularized cost function. First, the factorization of $\{W(p)\}$ into A and $\{B(p)\}$ is likely to introduce local minima of the cost function, which might be found instead of the true global minimum. Furthermore, (5.13) is not differentiable, when one of the terms $\|(b_{m,f}(1),\ldots,b_{m,f}(P))^\top\|_2, m \neq f$ or $\|(b_{1,1}(1),\ldots,\ldots,b_{M,M}(P))^\top\|_2$ is zero, which is expected to be the case at the optimum. For approaching these difficulties, we here propose using a modified version of the L-BFGS algorithm, which allows joint nonlinear optimization of A and $\{B(p)\}$, while taking special care of the nondifferentiability of the regularizer. The gradient of (5.13) for the case $m \neq f$ is obtained as

$$\begin{aligned}\frac{\partial \mathcal{L}^{\mathrm{SCSA}}}{\partial b_{m,f}(p)} &= -\sum_{t=P+1}^{T}\left(\tanh\left(s_m(t)-\widetilde{s}_m(t)\right)\ s_f(t-p)\right)\\ &\quad +\lambda\,\frac{b_{m,f}(p)}{\left\|\left(b_{m,f}(1),\ldots,b_{m,f}(P)\right)^\top\right\|_2}\end{aligned} \tag{5.15}$$

(analogously for $m \equiv f$) and

$$\begin{aligned}\frac{\partial \mathcal{L}^{\mathrm{SCSA}}}{\partial A_m} &= (P-T)A^{-\top}\mathbf{e}_m\\ &\quad +\sum_{t=P+1}^{T}\sum_{m=1}^{M}\left\{\tanh\left(s_m(t)-\widetilde{s}_m(t)\right)\times\left(\mathbf{x}(t)-\sum_{p=1}^{P}x_m(t-p)\mathbf{b}_m(p)\right)\right\}.\end{aligned} \tag{5.16}$$

Our modified L-BFGS algorithm checks before each gradient evaluation, which of the terms $\|(b_{1,1}(1),\ldots,\ldots,b_{M,M}(P))^\top\|_2$ and $\|(b_{m,f}(1),\ldots,b_{m,f}(P))^\top\|_2, m \neq f$ are (close to) zero. If any of these terms equals zero, the gradient is not defined uniquely but is a set-valued quantity called the subdifferential. Nevertheless, it is straightforward to compute the element of the subdifferential with the minimum norm, whose sign inversion is always a descent direction. Care must be taken, because in practice we would not find any of the above terms exactly equal to zero. Thus, we truncate the elements of B corresponding to the terms with small norms below some threshold to zero before computing the minimum norm subgradient. If the minimum is indeed attained at the truncated point, the minimum norm subgradient will be zero. Otherwise the subgradient will drive the solution away from zero. Further care must be taken in practice to prevent the solution from oscillating towards and away from zero.

We find that, using the outlined optimization procedure, sparse solutions can be found in shorter time, if the solution of the unregularized cost function is used as the initializer. The starting point may be obtained using the inverse transformation of (5.10).

SCSA via expectation maximization

Using joint optimization of A and $\{B(p)\}$, the heuristic pruning of connections might in some cases lead to suboptimal solutions regarding the composite loss function. For this reason, we present an alternative optimization scheme, which does not require any heuristic step. The idea here is to alternate between the estimation of both unknowns. Doing so can be justified as an application

of the expectation maximization (EM) algorithm (Neal and Hinton, 1998). Estimation of A given $\{B(p)\}$ (here called E-step) amounts to solving an unconstrained nonlinear optimization problem. This problem is convex, in contrast to the joint approach to SCSA parameter fitting. The convexity follows from the concavity of $\log|X|$ and $\log(\operatorname{sech}(ax))$ for constant a (and from the fact that the sum of convex functions is convex). The advantage of convex problems is that they feature a unique (global) minimum. In our case, the objective is smooth, which means that the minimum is guaranteed to be found, e. g., by the L-BFGS algorithm, making use of the gradient (5.16).

Optimization with respect to $\{B(p)\}$ for fixed A (M-step) is more involved due to the nondifferentiability of the $\ell_{1,2}$-norm regularizer. Gradient-based solvers such as L-BFGS are incapable of finding the exact solution. However, this problem is not as difficult as the joint optimization problem, since the cost function for constant A is convex. This can be seen from the fact that the objective is composed of a sum of $-\log(\operatorname{sech}(ax))$ terms (loss function) and the $\ell_{1,2}$-norm term (regularizer), which is a sum of ℓ_2-norms and thus convex. Hence we can solve this problem using the dual augmented Lagrangian (DAL) procedure (Tomioka and Sugiyama, 2009), which is a method for minimizing arbitrary convex loss functions with additional sparsity penalties. The DAL technique has been introduced in the context of minimizing the S-FLEX cost function in Section 4.2. In S-FLEX, a Gaussian noise model is assumed, leading to the quadratic loss function which is implemented as part of the freely available DAL software package. Here, we describe how to integrate support for the sech noise distribution into DAL, which requires explicit formulas for the loss function and its gradient, the convex conjugate (Legendre transform) of the loss function, as well as the gradient and the Hessian of the conjugate loss. Let $\mathbf{s}(t) = A^{-1}\mathbf{x}(t)$ be the demixed sources and $\widetilde{\mathbf{s}}(t) = \sum_{p=1}^{P} B(p)\mathbf{s}(t-p)$ be their autoregressive approximations. The loss function in terms of $\widetilde{\mathbf{s}}$ is defined as

$$\mathcal{L}^{\mathrm{A}}(\widetilde{\mathbf{s}}) \quad = \quad -\sum_{t=P+1}^{T}\sum_{m=1}^{M} \log\left(\tfrac{1}{\pi}\operatorname{sech}\left(\widetilde{s}_m(t) - s_m(t)\right)\right) , \tag{5.17}$$

while the gradient is

$$\frac{\partial \mathcal{L}^{\mathrm{A}}}{\partial \widetilde{s}_m(t)} \quad = \quad \tanh(\widetilde{s}_m(t) - s_m(t)) . \tag{5.18}$$

Let $h_m(t), m = 1,\ldots,M, t = P+1,\ldots,T$ denote the dual variables associated with the Legendre transform. The conjugate loss function is defined on the interval $[-1,1]$ and evaluates to

$$\begin{aligned}
\mathcal{D}^{\mathrm{A}}(\mathbf{h}) \quad &= \quad \sum_{t=P+1}^{T}\sum_{m=1}^{M} \sup_{\widetilde{s}_m(t)} \left(-h_m(t)\widetilde{s}_m(t) + \log\frac{\operatorname{sech}\left(\widetilde{s}_m(t) - s_m(t)\right)}{\pi}\right) \\
&= \quad \sum_{t=P+1}^{T}\sum_{m=1}^{M}\Bigg(\frac{1-h_m(t)}{2}\log\frac{1-h_m(t)}{2} \\
&\quad + \quad \frac{1+h_m(t)}{2}\log\frac{1+h_m(t)}{2} - h_m(t)s_m(t) + \log\tfrac{2}{\pi}\Bigg) .
\end{aligned} \tag{5.19}$$

The gradient of the conjugate loss is given by

$$\frac{\partial \mathcal{D}^{\mathrm{A}}(\mathbf{h})}{\partial h_m(t)} \quad = \quad \tfrac{1}{2}\log\frac{1+h_m(t)}{1-h_m(t)} - s_m(t) . \tag{5.20}$$

The Hessian is diagonal with elements

$$\frac{\partial^2 \mathcal{D}^{\mathrm{A}}(\mathbf{h})}{\partial h_m(t)^2} = \frac{1}{2(1-h_m^2(t))} . \tag{5.21}$$

Having defined the E- and M-steps, we have turned a nonconvex estimation problem into a sequence of two convex problems, which can both be solved exactly. A final estimate of the model parameters is obtained by alternating between E- and M-steps until convergence.

5.2.6 Performance under realistic conditions

Setting

We conduct a number of simulations in order to assess the efficacy of (sparsely-) connected sources analysis in terms of decomposing EEG data arising from interacting brain sources compared to existing BSS approaches. Here, we perform a quantitative evaluation, while a qualitative assessment of CSA and SCSA using the simulated example introduced in Chapter 3 is provided in Section 5.4.

We simulate seven time series (pseudo-sources) of length $T = 2\,000$ according to a stable MVAR model of order $P = 4$ with random parameters. Seven out of the 42 possible interactions are modeled by setting the corresponding off-diagonal MVAR coefficients $b_{m,f}(p), m \neq f, 1 \leq p \leq P$ to nonzero values. The innovations are drawn from the sech distribution. The pseudo-sources are mapped to 118 EEG channels defined in the extended 10-20 positioning system using the theoretical spread of seven randomly placed dipoles. The spread is computed using the realistic forward model (Nolte and Dassios, 2005), which is built based on anatomical MR images of the "Montreal head" (Holmes et al., 1998).

In reality, measurements are never noise-free and the familiar model $\mathbf{x}(t) = A\mathbf{s}(t) + \boldsymbol{\varepsilon}(t)$ is more appropriate than (5.7). Since none of the methods compared here (see below) explicitly models a noise term, it is important to evaluate their robustness to model violation. To this end, we construct additional variants of the pseudo-EEG dataset by adding six different types of noise $\boldsymbol{\varepsilon}(t)$. These six variants (N1–N6), which are summarized in Table 5.2, differ in their degree of spatial and temporal correlation as follows. In variants N1 and N4, noise terms $\varepsilon_m(t), m = 1, \ldots, M$ are drawn independently for each *sensor*, i. e., have no spatial correlation. For variants N2 and N5, noise terms $\varepsilon^*_m(t), m = 1, \ldots, M$ are drawn independently for each *source*. In this case, sources and noise contributions to the EEG share the same covariance given by the mixing matrix A, i. e., $\mathbf{x}(t) = A(\mathbf{s}(t) + \boldsymbol{\varepsilon}^*(t))$. For the remaining variants N3 and N6, spatially independent noise sources are simulated at all positions of a grid covering the whole brain. This yields the model $\mathbf{x}(t) = A\mathbf{s}(t) + A^*\boldsymbol{\varepsilon}^*(t)$, in which, in contrast to the previous model, noise contributions are not collinear to the sources. We further distinguish between noise sources with and without temporal structure. In variants (N1–N3), noise terms are drawn i.i.d. from a standard normal distribution at each time instant t. In variants N4–N6, the temporal structure of each noise source is determined by a stable univariate AR model of order 20 with random parameters, i. e., $\varepsilon_m(t) = \sum_{p=1}^{20} b^*(p)\varepsilon_m(t-p) + \xi(t)$ for noise type N4. Note that, since no time-delayed dependencies between noise sources are modeled, no additional time-lagged effective connections are introduced by the noise. We use a signal-to-noise ratio of SNR = 2 in all experiments, where the SNR is defined as $\mathrm{SNR} = \|\mathbf{vec}\,(AS)\|_2 / \|\mathbf{vec}\,(E)\|_2$. One-hundred datasets with different realisations of MVAR coefficients, innovations and noise are constructed for each category.

We perform two additional experiments (100 repetitions each) in order to investigate the performance of the various methods under variation of the connectivity structure of the underlying sources, as well as the SNR. Seven degrees of connectedness (linearly interpolated from 0 % to 100 %) and seven choices of the SNR (linearly interpolated from 1 to 4) are considered. These ranges include the parameters used in the previous experiments (17 % = 7/42 of all possible interactions present, SNR = 2). The effect of SNR variation is investigated using white sensor noise without temporal structure (N1), while the effect of the degree of connectivity is studied for the noiseless case (N0).

In all experiments, PCA dimensionality reduction is applied to the pseudo-EEG by retaining the strongest signal components. Since our evaluation scheme relies on a one-to-one mapping between estimated and true components (see Section 3.10), we use here exactly as many dimensions as original sources, i. e., $M = 7$. In practice, this information is, of course, not avalaible and the number of dimensions can be chosen such that, for instance, 99 % of the variation in the EEG is explained. Alternative ways to choose the number of components are discussed in (Hansen et al., 1999). Our experience shows that taking too many dimensions is generally less harmful than taking too few, since spare dimensions can always be used to just dump noise.

	independent in time	correlated in time
independent in sensors	N1	N4
correlated in sensors[a]	N2	N5
correlated in sensors[b]	N3	N6

Table 5.2: The six types of noise used in the simulations. Noise with temporal correlation structure is created using univariate AR models of order $P = 20$. Spatial correlation is introduced using a realistic forward model. We distinguish between the case, where noise sources coincide with the true dipoles ([a]), and the case in which noise from all brain sites contributes to the measurements ([b]).

We test the ability of ICA, MVARICA, CICAAR and the two proposed methods CSA and SCSA to reconstruct the seven sources and their effective connectivity graph. The ICA variant used here is TDSEP, which is implemented using fast Frobenius-norm joint diagonalization (Ziehe et al., 2004). The number of temporal lags used in TDSEP is set to 100. Note that, although the goal of instantaneous ICA in general is fundamentally different from source connectivity analysis, it is included here because it is occasionally used in the literature (Beckmann et al., 2005; Meinecke et al., 2005; Astolfi et al., 2007). Moreover, TDSEP has been shown to perform reasonably well in the simulated example investigated in Section 3.10. We apply MVARICA, CICAAR, CSA and SCSA with $P \in \{1, 2, \ldots, 7\}$ temporal lags, where four is the true MVAR model order for CSA, SCSA and MVARICA. CICAAR has the disadvantage here that it may generally require extended temporal filters for reconstructing sources following (5.8). However, due to computational time constraints, $P = 7$ is also taken as the maximum lag for this method. For MVARICA and CICAAR, we use implementations provided by the respective authors. These implementations adopt the Bayesian information criterion (BIC) for selecting the appropriate number of time lags. The same criterion is used to select the model order in CSA. For SCSA, the same model order as in CSA is used. The regularization constant λ of SCSA is determined by means of 5-fold cross-validation,

i. e., by evaluating the likelihood on test data. Estimates of $\{B(p)\}$ and A according to SCSA are obtained either jointly using the modified L-BFGS algorithm or alternatingly using 20 additional EM steps. These variants are named SCSA and SCSA_EM here, respectively. We use an L-BFGS implementation by Naoaki Okazaki[3] and DAL for computing the SCSA estimates.

The most important evaluation criterion is the reconstruction of the mixing matrix, since all other relevant quantities can be derived therefrom. All considered methods provide an estimate of the 7×7 demixing matrix in PCA space, which can be inverted and multiplied by the precomputed PCA eigenvectors related to the seven strongest eigenvalues to yield a 118×7 mixing matrix estimate $\widehat{A}$. The columns of $\widehat{A}$ correspond to spatial field patterns of the estimated sources, but unfortunately these patterns can generally only be determined up to sign, scale and order. For this reason, optimal pairing of true and estimated patterns as described in Section 3.10 is performed. Having found the optimal pairing, the columns of $\widehat{A}$ are permuted and scaled to approximate A as well as possible using the optimal regression coefficients. The goodness-of-fit with respect to the whole matrix A is used to evaluate the quality of the various decompositions. Additionally, using the optimally-matched mixing patterns, dipole scans are conducted. That is, for each discrete location in the brain (5 mm grid size), a dipolar current source is fitted and the location of the dipole that best explains the EEG pattern is determined. The deviation of these locations from the true locations is measured in terms of the Euclidean distance. A typical example of a mixing pattern estimated by SCSA and the corresponding reconstructed dipole is shown in Figure 5.2.

Eventually, causal discovery is carried out on the demixed sources. We use MVAR estimation by ridge regression and subsequent multiple testing (see Section 5.1) for this task (an application of PSI would be also suitable). A significant influence from s_i to s_j is assumed, if the p-value of one of the coefficients $b_{i,j}(p), p = 1, \ldots, P$ falls below the critical value. We compute the area under the curve (AUC) score measuring the correctness of the effective connectivity estimation as a third As a third performance criterion besides the mixing matrix approximation error and the dipole localization error. The AUC is calculated by varying the significance threshold and comparing estimated and true binary connectivity matrices for each threshold. Note that this way of assessing connectivity is pursued here, because only some BSS methods provide built-in connectivity estimates. For SCSA, however, interaction could as well be read off directly from the sparsity pattern of the estimated MVAR coefficients as it has been performed in the evaluation of sparse Granger-causal measures in Section 5.1.

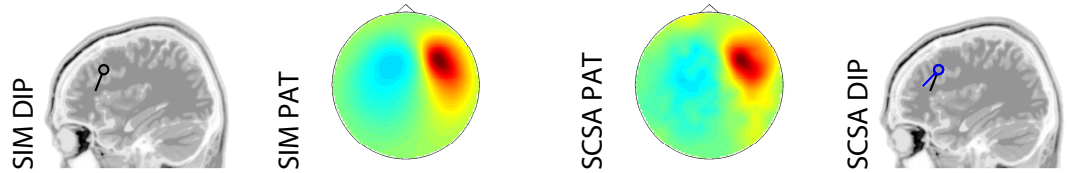


Figure 5.2: Example of simulated data (noise type N1) and corresponding reconstruction by SCSA. SIM DIP: simulated dipole. SIM PAT: field pattern describing the simulated dipole's influence on the EEG (one column of A). SCSA PAT: field pattern as estimated by SCSA from a noisy EEG time series. SCSA DIP: reconstructed dipole, obtained from the estimated pattern via inverse source reconstruction.

[3] http://www.chokkan.org/software/liblbfgs/

Results

Figure 5.3 (a) depicts the extent to which the mixing matrix is approximated by the various BSS approaches. One boxplot is drawn for the noiseless case (No) and each of the six noisy variants (N1–N6, see Table 5.2). The plots display the median performance achieved in 100 repetitions, as well as the lower and upper quartiles and the extremal values. Non-overlapping notches of two boxes indicate that the two medians differ at the 5 % significance level. Outliers (red crosses) do not enter the statistics.

As a result of the simulations, SCSA typically achieves the smallest reconstruction error, followed by CSA, CICAAR, MVARICA and TDSEP independent component analysis. In many cases, the differences are also significant. Correct mixing matrix estimation affects both the localization error achievable by applying inverse source reconstruction to the estimated patterns and the detection error of effective connectivity analyses of the demixed sources. As a result of approximating the mixing matrix well, SCSA achieves smaller dipole localization errors than all the other methods, except in scenario N4 (see Figure 5.3 (b)). The same situation occurs when it comes to estimating the connectivity between sources (Figure 5.3 (c)). Notably, there is a consistent performance gap between CSA, SCSA and CICAAR on one hand, and MVARICA and TDSEP on the other hand according to all three metrics. In this light, the successful separation of two interacting sources reported in Section 3.10 for MVARICA in the case of non-Gaussian innovations and for TDSEP regardless of the distribution of the innovations seems to be a result that cannot be reproduced for larger numbers of sources and respective interconnections.

In the presence of noise, the relative degradation of performance is the same for all methods. Generally, noise that is collinear to the sources (N2/N5) leads to higher performance than noise that is uncorrelated across sensors (N1/N4) and noise with arbitrary spatial correlation structure (N3/N6). These differences are partially explained by the effectiveness of the PCA step performed for dimensionality reduction. The data in the PCA subspace on average accounts for 96 % of the total variation in the data for the noise types N1/N4 and 81 % for the noise types N3/N6, while it is 100 % for the collinear noise N2/N5. That is, in the presence of strong noise sources that are not collinear to the interacting sources, the PCA step is likely to remove decrease the signal-to-noise ratio. As shown in the right panel of Figure 5.3 (d), the performance of all methods decays with decreasing SNR, while the differences between the methods remains rather stable across differing SNR levels. The left panel of the figure demonstrates that the improvement of CSA and SCSA over CICAAR, MVARICA and ICA is stable even under variation of the degree of connectedness. Only, the SCSA variants seem to lose their advantage over CSA in the case of very dense connectivity structures.

The average time consumed by each method for processing one dataset is shown in Figure 5.3 (e). Most methods finish within a rather short time, while the EM implementation of SCSA falls into the intermediate range and the application of CICAAR requires the most time. However, for SCSA there is still room for improvement, since the regularization parameter of this method is currently selected by the cross-validation procedure, which could be changed.

In sum, the success of SCSA, CSA and CICAAR noted here demonstrates that the correlated source model is an appropriate choice for the blind decomposition of interacting sources. If, moreover, the connectivity graph has a sparse structure, this can be effectively exploited by SCSA.

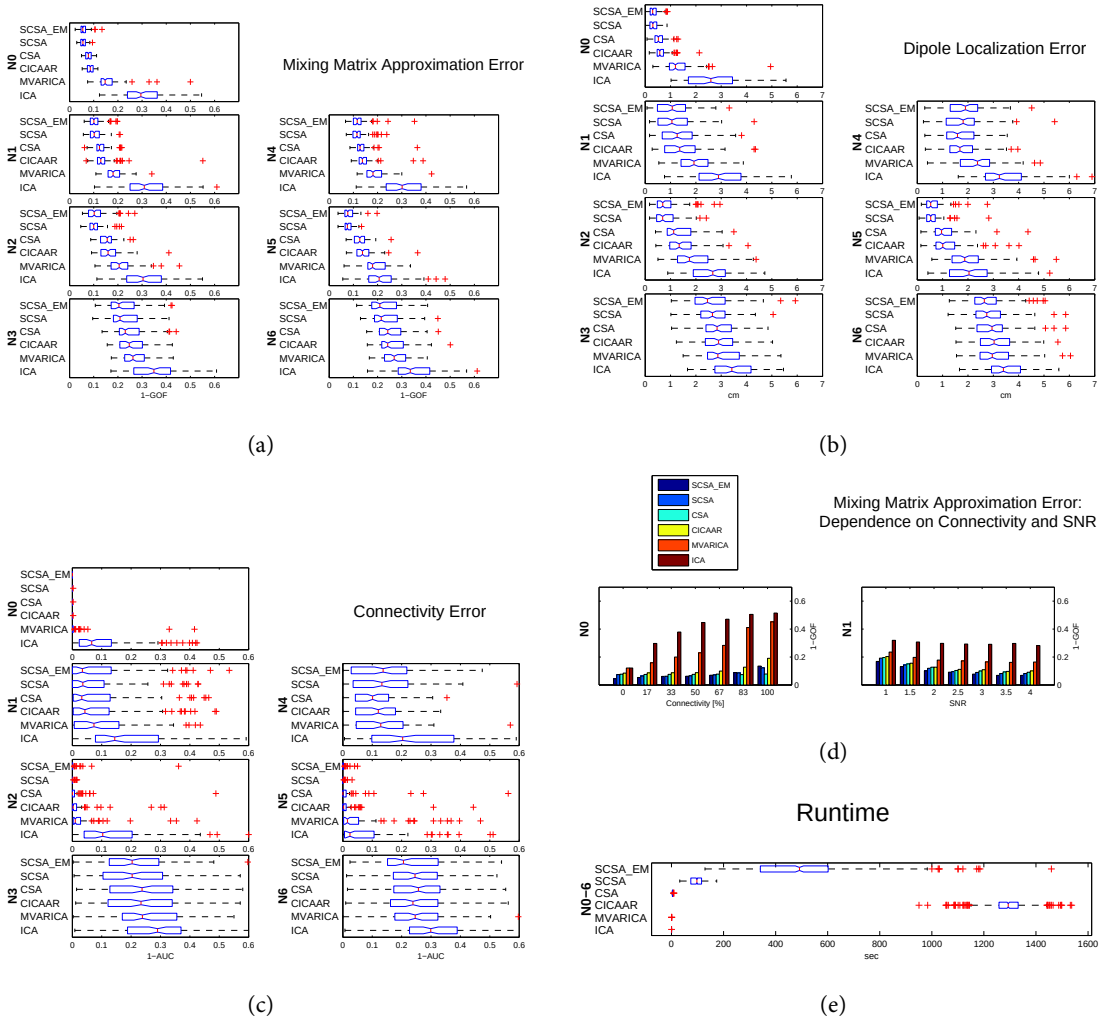


Figure 5.3: Results of the evaluation study for the proposed (sparsely-) connected sources analysis variants (SCSA_EM, SCSA, CSA) as well as convolutive ICA with an auto-regressive inverse model (CICAAR), multivariate autoregression + independent component analysis (MVARICA) and temporal decorrelation source separation independent component analysis (ICA). Different subfigures depict the methods' performance in the noiseless case (N0), as well as in the presence of different types of noise (N1-N6, see Table 5.2). (a) Estimation errors of the mixing matrix according to the goodness-of-fit (GOF) criterion. (b) Localization errors of dipole fits conducted on the estimated mixing field patterns. (c) Estimation errors regarding the source effective connectivity structure as measured by fitting an MVAR model subsequently to the demixed sources using ridge regression and testing the obtained coefficients for significant interaction. The performance measure reported is the area under the curve (AUC) score obtained by varying the significance level. (d) Mixing matrix approximation performance under variation of the degree of connectedness (left side) and SNR (right side). The performance at different noise levels is investigated for white sensor noise without temporal structure (N1), while the influence of connectedness is studied in the noiseless case (N0). (e) Average runtime taken over all experiments conducted for this study.

5.3 Integrating S-FLEX inverse source reconstruction into SCSA

The mixing matrix estimate A that is provided by SCSA and CSA contains the EEG field patterns of brain sources obeying the specific dynamics assumed by these methods. These patterns can be subjected to post hoc inverse source reconstruction. In our empirical evaluation in Section 5.2, we fit single dipoles, but in general any inverse method may be applied, including the distributed inverse imaging approaches proposed in this thesis. Performing source localization amounts to solving the underdetermined system $A = \widetilde{A}F$, where $\widetilde{A}$ is the known lead field matrix and F are the unknown primary current source distributions of the estimated source components. Estimation of the correlated sources model in terms of F instead of A is conceptually similar to CSA/SCSA, while optimization over source distributions bears several advantages compared to estimating field patterns. This is the idea pursued in this section. Most importantly, estimating F amounts to integrating source localization directly into the blind source separation procedure. This enables us to integrate the prior knowledge that the source mixing patterns must correspond to brain activity directly into the estimation of A and $\{B(p)\}$. Moreover, it is possible to impose much more specific assumptions on source distributions than on sensor-space field maps, which enables us to perform regularization of the mixing patterns in a physiologically meaningful way. In particular, we can impose sparsity penalties on the source distributions, which would be pointless in sensor space. We here consider regularization of F by means of sparse basis field decompositions according to S-FLEX. The resulting approach, which we tentatively call SCSA+FLEX, is in preparation for publication (Kawanabe et al., 2011). While a numerical evaluation is not yet available, we here outline the model equations, discuss a suitable optimization scheme and provide a proof-of-concept of the method's applicability on simulated data.

5.3.1 Model specification

We consider the model

$$\mathbf{x}(t) = AW^{-1}(G \otimes I_3)C\,\mathbf{s}(t) + \boldsymbol{\eta}(t) \tag{5.22}$$

$$\mathbf{s}(t) = \sum_{p=1}^{P} B(p)\mathbf{s}(t-p) + \boldsymbol{\varepsilon}(t)\,, \tag{5.23}$$

where A is the $M \times 3N$ lead field matrix, $\mathbf{s}(t), t = 1, \ldots T$ are the K-dimensional source time series and $F = W^{-1}(G \otimes I_3)C$ is the $3N \times K$ matrix of source current distributions, which is factorized into the constant $3N \times 3N$ inverse depth compensation matrix W^{-1}, the constant $3N \times 3L$ spatial basis function matrix $G \otimes I_3$ and the unknown $3L \times K$ coefficient matrix $C = (\mathbf{vec}(\widetilde{C}(1)), \ldots, \mathbf{vec}(\widetilde{C}(K)))$ as in S-FLEX (see Section 4.2), where $\widetilde{C}(k) = (\mathbf{c}_1(k), \ldots, \mathbf{c}_L(k))^\top$. Note that the number of sources K can be lower than the number of electrodes M here. Another difference from the SCSA model is that we introduce measurement noise $\boldsymbol{\eta}(t)$, which is assumed to be i.i.d. Gaussian according to $\mathcal{N}(\mathbf{0}, (\beta\Sigma)^{-1})$. The noise covariance Σ is assumed to be known, e. g., from a baseline measurement. If no estimate of Σ exists, the identity matrix may be used, which leads to the standard squared loss function used in FVR and S-FLEX. To ensure identifiability of the model, the innovations $\boldsymbol{\varepsilon}$ of the source MVAR process (5.23) are assumed to be non-Gaussian. We here employ the hyperbolic secant distribution as in (5.25). Using the shortcut $\Gamma = AW^{-1}(G \otimes I_3)$,

the conditional models are expressed as

$$p(X|S,C,\beta) \;=\; \frac{\beta^{M/2}|\Sigma|^{½}}{(2\pi)^{M/2}} \exp\left(-\tfrac{1}{2}\beta(X-\Gamma CS)^{\top}\Sigma(X-\Gamma CS)\right) \tag{5.24}$$

$$p\left(\{\mathbf{s}(t)\}_{t=1}^{T}|\{B(p)\}\right) = \prod_{t=1}^{T}\prod_{k=1}^{K}\tfrac{1}{\pi}\,\mathrm{sech}\left(s_k(t)-\widetilde{s}_k(t)\right)\;, \tag{5.25}$$

where $\widetilde{\mathbf{s}}(t) = \sum_{p=1}^{P} B(p)\mathbf{s}(t-p)$ and $\mathbf{s}(t)$ for $t \le 0$ is assumed to be zero. If the number of observations M is equal to the number of sources K, $\Gamma = I_M$, and if there is no noise (i. e., $\beta = \infty$), our model is equivalent to SCSA.

5.3.2 Parameter estimation

We are interested in maximizing the posterior probability of the model parameters $\mathbf{s}(t)$, C, $\{B(p)\}$ and β given the observations $\mathbf{x}(t)$ under prior expectations regarding the sparsity of C and $\{B(p)\}$ according to SCSA and S-FLEX. The MAP estimate is obtained by minimizing the penalized negative log-likelihood

$$\begin{aligned}
&\mathcal{L}^{\text{SCSA+FLEX}}(C,\{B(p)\},\{\mathbf{s}(t)\}_{t=1}^{T},\beta)\\
&\quad= \tfrac{1}{2}\beta\left(\sum_{t=1}^{T}\left(\mathbf{x}(t)-\Gamma C\mathbf{s}(t)\right)^{\top}\Sigma\left(\mathbf{x}(t)-\Gamma C\mathbf{s}(t)\right)\right) - \tfrac{MT}{2}\log\beta\\
&\qquad-\sum_{t=1}^{T}\sum_{k=1}^{K}\log\left(\tfrac{1}{\pi}\,\mathrm{sech}\left(s_k(t)-\widetilde{s}_k(t)\right)\right) + \lambda_C\sum_{l=1}^{L}\sum_{k=1}^{K}\|\mathbf{c}_l(k)\|_2\\
&\qquad+\lambda_B\left(\left\|\left(b_{1,1}(1),\ldots,\ldots,b_{K,K}(P)\right)^{\top}\right\|_2 + \sum_{k\neq f}\left\|\left(b_{k,f}(1),\ldots,b_{k,f}(P)\right)^{\top}\right\|_2\right)\;,
\end{aligned} \tag{5.26}$$

in which the parameters λ_B and λ_C control the sparsity of the source MVAR coefficients $\{B(p)\}$ and the source basis field expansion coefficients C, respectively.

As for SCSA, minimization of $\mathcal{L}^{\text{SCSA+FLEX}}$ can be carried out using an EM type algorithm, i. e., by optimizing one variable (group) at a time. Interestingly, optimization with respect to $\{B(p)\}$ for fixed $\mathbf{s}$, C and β leads to exactly the same optimization problem that occurs in SCSA. Optimization with respect to C amounts to minimizing a quadratic function with additional $\ell_{1,2}$-norm terms, which can be performed in the same way as for S-FLEX using dual augmented Lagrangians. Optimization with respect to $\mathbf{s}$ is a smooth nonlinear but nonconvex problem, which can be solved using conventional L-BFGS optimization. We here use an implementation by Mark Schmidt[4]. The gradient of the loss function with respect to $\mathbf{s}$ is given by

$$\begin{aligned}
\frac{\partial\mathcal{L}^{\text{SCSA+FLEX}}}{\partial\mathbf{s}(t)} \;&=\; -\beta\,C^{\top}\Gamma^{\top}\Sigma\left(\mathbf{x}(t)-\Gamma C\,\mathbf{s}(t)\right) + \tanh\left(\mathbf{s}(t)-\widetilde{\mathbf{s}}(t)\right)\\
&\qquad-\sum_{p=1}^{P}(B(p))^{\top}\tanh\left(\mathbf{s}(t+p)-\widetilde{\mathbf{s}}(t+p)\right)\;.
\end{aligned} \tag{5.27}$$

[4] http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html

Finally, optimization with respect to β is a quadratic programming problem. By setting the gradient

$$\frac{\partial \mathcal{L}^{\text{SCSA+FLEX}}}{\partial \beta} = \frac{1}{2}\left(\sum_{t=1}^{T}\left(\mathbf{x}(t) - \Gamma C\mathbf{s}(t)\right)^{\top}\Sigma\left(\mathbf{x}(t) - \Gamma C\mathbf{s}(t)\right)\right) - \frac{MT}{2\beta} \tag{5.28}$$

to zero, we obtain the analytic update rule

$$\beta = \left\{\frac{1}{MT}\sum_{t=1}^{T}\left(\mathbf{x}(t) - \Gamma C\mathbf{s}(t)\right)^{\top}\Sigma\left(\mathbf{x}(t) - \Gamma C\mathbf{s}(t)\right)\right\}^{-1}. \tag{5.29}$$

By alternating between optimization of the different variable groups, SCSA+FLEX is guaranteed to find a local minimum of the overall cost function. Notably, the subproblems related to minimizing $\{B(p)\}$, C and β are convex, which means that the unique minimizer given the current estimate of $\mathbf{s}(t)$ is guaranteed to be found in each iteration of the EM algorithm. Upon convergence of the procedure, the matrix of estimated source distributions can be calculated via $\widehat{F} = W^{-1}(G \otimes I_3)\widehat{C}$.

5.3.3 Simulated example

Setting

Since this section contains work in progress, an exhaustive numerical evaluation as presented for CSA and SCSA in Section 5.2 is still to be conducted for SCSA+FLEX. Nevertheless, we provide a proof-of-concept here. We generate $K = 3$ source time series z_i, $i = 1, \ldots, 3$ of length $T = 3\,000$ according to a stable multivariate AR model of order $P = 3$. The source innovation time series are drawn from the sech distribution. Only the interaction from z_2 to z_3 is modeled by allowing the respective off-diagonal coefficients to be nonzero. Three dipolar sources are modeled by choosing random locations within the brain shell of the "Montreal" head model and Gaussian-distributed random current moment vectors. The spread function of these dipoles is used to obtain a pseudo-EEG measurement $\mathbf{x}(t)$ at $M = 59$ channels of the extended international 10-20 electrode placement system. The sensor data are superimposed with Gaussian white noise of equal strength (SNR = 1). We apply CSA (see previous section) to the strongest three PCA components of $\mathbf{x}(t)$ for all $P \in \{1, \ldots, 9\}$ and use the Bayesian information criterion to determine the optimal P. Using the selected value $P = 3$, we apply SCSA+FLEX with parameters $M = 3$, $\lambda_C = 10^{-2}$ and $\lambda_H = 10^{-6}$ to the full-rank dataset $\mathbf{x}(t)$ using a source space of 2 142 dipoles that are arranged in a cubic grid of 1 cm side length within the brain shell. We compute PSI on the estimated sources and perform statistical testing of PSI scores according to Nolte et al. (2008).

Results

Figure 5.4 depicts the source distributions estimated by SCSA+FLEX with the true simulated dipoles overlaid as arrows, as well as the spatial field patterns corresponding to the true and estimated sources. The pairs of simulated and estimated source components have been assigned manually here. Apparently, the source reconstruction using SCSA+FLEX is successful, with focal source distributions, the maxima of which are close to the true simulated dipoles. The correctness of the demixing is also indicated by the estimated mixing patterns, which closely resemble the

field patterns of the simulated dipoles. Moreover, B has a sparse structure in the optimum, where two of the six possible effective connections are nonzero, including the correct connection from z_2 to z_3. The analysis of PSI on the estimated sources yields this connection as the only significant one (after correction for three multiple tests, i. e., z > 2.4), which is the correct result.

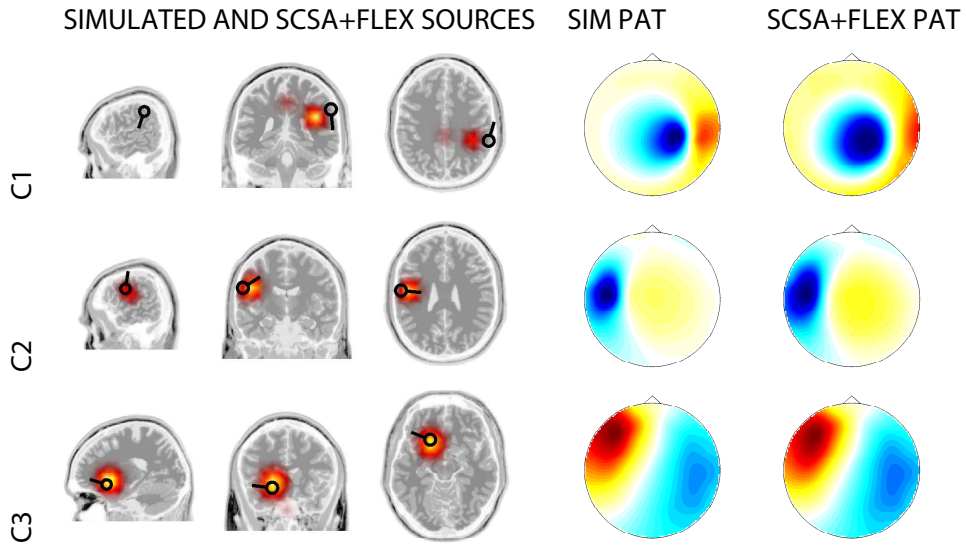


Figure 5.4: An example of simulated interacting sources and their reconstructions according to SCSA+FLEX. Left panel: Three dipolar sources (depicted as black arrows) and corresponding SCSA+FLEX reconstructions (underlaid red and yellow patches) estimated under assumptions on the sparsity of the sources' effective connectivity structure and the simplicity of the sources' current distributions. SIM PAT: field patterns describing the simulated dipoles' spread into the EEG. SCSA+FLEX PAT: EEG field patterns of the current distributions estimated by SCSA+FLEX.

5.3.4 Relation to SCSA

In principle, both CSA/SCSA on one hand and SCSA+FLEX on the other hand are capable of delivering EEG mixing patterns as well as source distributions (see Figures 5.2 and 5.4). Both CSA and SCSA perform unconstrained optimization over sensor-space mixing patterns. These patterns may be subjected to post-hoc inverse source reconstruction, but there is generally no guarantee for the success of this, since the estimated patterns may contain non-physiological noise. In contrast, SCSA+FLEX directly estimates the underlying current distributions in a physiologically-constrained way and thereby solves the EEG inverse problem as part of the global optimization. Calculation of sensor-space mixing patterns from these source distributions is trivially performed by application of the well-defined forward mapping. A further advantage of SCSA+FLEX is that a noise model of the observations is employed, which allows one to model the case $K \leq M$. In CSA and SCSA, $K = M$ must be ensured using PCA dimensionality reduction. The variation of the data in the PCA subspace is to be explained perfectly, resulting in potentially noisy source time series and mixing patterns.

5.4 Reconstruction of simulated source connectivity using SCSA

5.4.1 Setting

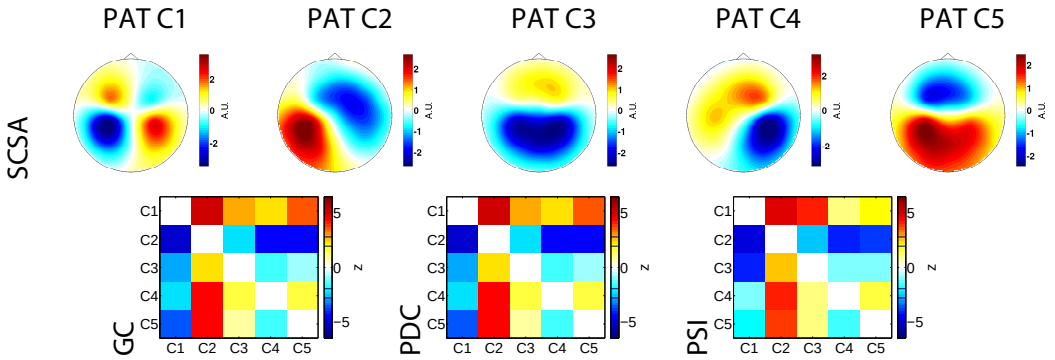
We now return to the simulated scenario used throughout Chapter 3 for benchmarking standard approaches to effective connectivity analysis and in Section 4.6 for assessing S-FLEX inverse source reconstruction preprocessing with respect to subsequent effective connectivity analysis. The simulation comprises two interacting sources that are placed below C_3 and C_4. Information is modeled to flow from the source in the left hemisphere to the source in the right hemisphere. We here analyze the data presented in Section 3.10 using a similar methodology. Sparsely-connected sources analysis is applied to normalized time series associated with the five strongest PCA components using the implementation based on modified L-BFGS optimization. The number of lags P of the source MVAR model is selected from $\{1, \ldots, 9\}$ by evaluating the BIC criterion for the solution of the unregularized (CSA) problem. Using the selected P, the regularization parameter λ is determined by 5-fold cross-validation. A final SCSA fit using optimal values for P and λ is then performed on the complete time series for each dataset. The components found in each of the 100 repetitions of the experiment are paired according to the procedure outlined in Section 3.10 and statistical testing is performed using a one-sample t-test with subsequent transformation of t-scores into z-scores. Analogously to Section 3.10, we consider the case in which the MVAR innovation terms are sampled from the standard normal distribution, as well as the case of hyperbolic-secant-distributed innovations.

5.4.2 Results

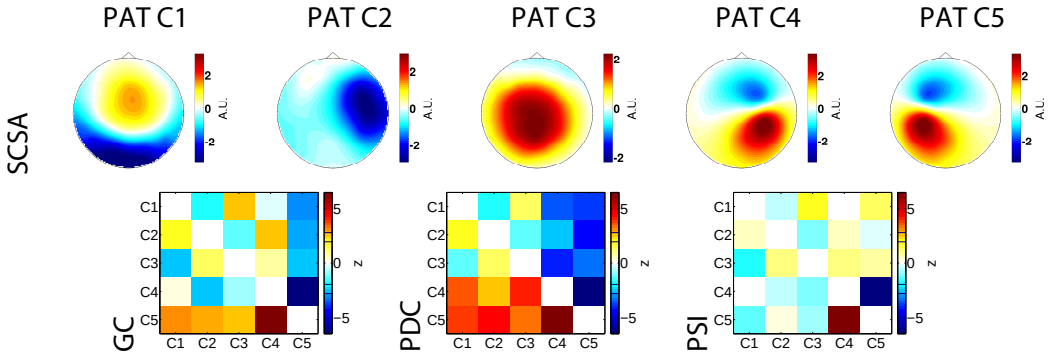
As can be seen from Figure 5.5 (a), SCSA does not perform very well on data generated with Gaussian distributed source MVAR innovation terms. All of the five extracted components contain contributions from both interacting sources, while none of the sources is isolated in a single component. The best approximation of the source in the left hemisphere is achieved by component C_3 (GOF = 0.55 ± 0.02), while the source in the right hemisphere is best approximated by component C_1 (GOF = 0.35 ± 0.02). Several significant interactions are indicated by Granger causality, partial directed coherence and also the phase-slope index, which are, however, not reported here, since they result from an insufficient demixing.

The performance on data that is generated using non-Gaussian (sech-distributed)innovations is depicted in Figure 5.5 (b). Here, SCSA recovers the true source in the left hemisphere as the fifth component (GOF = 0.94 ± 0.01) and the source in the right hemisphere as the fourth component (GOF = 0.94 ± 0.004). Hence, SCSA outperforms all other BSS approaches tested in this example in terms of the accuracy of the source separation. Consequently, significant ($z > 2.8$) information flow from C_5 and C_4 is estimated by all three measures of effective connectivity. While this is the only connection estimated by PSI, GC indicates additional flow from C_5 to C_1, and PDC yields additional interactions from C_5 to C_1, C_2 and C_3, and from C_4 to C_1 and C_3.

The qualitative results obtained here confirm the potential of SCSA blind source separation to recover interacting sources, which has already been demonstrated in the quantitative evaluation study described in Section 5.2. Notably, a non-Gaussian distribution of the source MVAR innovations is a prerequisite for successful blind source separation in the correlated sources model.



(a) Gaussian innovations



(b) super-Gaussian innovations

Figure 5.5: Effective connectivity of simulated EEG as estimated by Granger causality (GC), partial directed coherence (PDC) and the phase-slope index (PSI) after application of blind source separation (BSS) preprocessing according to sparsely-connected sources analysis (SCSA). Two source dipoles with tangential orientations are modeled 3 cm below the $C_3/4$ electrodes. Information flow from the left (C_3) to the right (C_4) source is modeled by means of a bivariate AR process with (a) *Gaussian*-distributed innovations or (b) *non-Gaussian* (sech-distributed) innovations. The simulated EEG is superimposed by non-interacting biological and sensor noise (SNR = 1). BSS techniques are applied to the five strongest PCA components. The significance of estimated interactions between demixed signals is measured in terms of z-scores and visualized as matrices, where entries with red and yellow colors ($z > 0$) stand for information outflow and entries with blue and cyan colors ($z < 0$) stand for information inflow of the source marked in the respective row. The Bonferroni-corrected significance level is indicated by thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line.

5.5 Discussion (fusing physical and dynamical constraints in blind source separation)

Analyzing brain effective connectivity based on EEG data is a challenging problem, since volume conduction negatively affects the interpretability of sensor-space connectivity estimates and may even give rise to spurious results. In this chapter, we have established novel methods for the blind recovery of interacting sources, which overcome these problems in an elegant and numerically appealing manner. Our approaches model the EEG as a mixture of correlated sources, the connectivity of which is accounted for by a source MVAR model. Starting with an unregularized model in CSA, we add assumptions in order to regularize the solution in a way that is driven by neurophysiological considerations. In SCSA, we use an $\ell_{1,2}$-norm penalty to introduce sparsity of the source effective connectivity graph. In this manner, we achieve a data-driven interpolation between two extremes: a source model that has fully correlated sources and one that does not allow for cross-talk between the extracted components.

As it is desirable to relate interacting components found by blind source separation to actual brain anatomy, a subsequent source localization may be performed on the SCSA source mixing patterns. This strategy is not uncommon in the BSS literature (Nolte et al., 2006; Gómez-Herrero et al., 2008; Marzetti et al., 2008; Nolte et al., 2009). Notably, source localization may also be carried out prior to connectivity analysis, which is the approach pursued in Chapter 4 as well as, e. g., Astolfi et al. (2006a) and Supp et al. (2007). Unlike all these two-step approaches, SCSA+FLEX is a single-step fused BSS/inverse source reconstruction technique, which is capable of demixing EEG data into source time series while jointly estimating the sources' current distributions and connectivity structure. Thereby, SCSA+FLEX integrates assumption on the physical properties of the sources' current distributions (smoothness, sparsity, rotational invariance) as imposed by S-FLEX with assumptions on the sources' dynamics (sparsity of the connectivity graph) made by SCSA. Such a combined procedure is compelling, because the dynamical assumptions on the sources affect the estimation of their spatial patterns, while conversely spatial criteria drive the optimization away from solutions that are non-physiological. Using a similar argument, Valdés-Sosa et al. (2009) devise a method that combines inverse source reconstruction with ICA.

Let us finally focus on the dynamical assumptions we impose to estimate individual brain sources and their interactions. While ICA results in a unique decomposition assuming statistical independence, such an assumption is inconsistent when studying brain interactions. However, all neural interactions require a minimum delay well within the temporal resolution of electrophysical measurements of brain activity. Hence, it makes sense to assume independent innovation processes and to model all interactions explicitly using AR matrices. However, that means that our methods can exploit independence only on reduced information contained in the residuals of the model. We emphasize, that BSS methods employing higher order statistics without using temporal information would fail completely if the data were Gaussian distributed. Processes tend to be super-Gaussian if they are not always active, which is a reasonable assumption for brain sources. Here we assume a linear dynamical model and super-Gaussian innovation processes, i. e., the only cause of non-Gaussianity is the innovation process itself. Real brain networks are, of course, more complicated. However, the question whether nonlinear dynamical models may improve results or are even essential for a correct decomposition is beyond the scope of this thesis.

6 Source connectivity analysis of the human alpha rhythm during rest

In his pioneering work, Berger (1938) observed oscillations around 10 Hz in the EEG, which are nowadays known as the *alpha-rhythm*. Activity in the alpha band, which ranges from 8 to 13 Hz, is the strongest signal component that can be recorded by EEG. Within this band, there exist at least two distinct rhythms with distinct functional relevance and slightly differing peak frequencies. The predominant posterior alpha-rhythm has been related to a number of behavioral markers including vigilance (Papadelis et al., 2007; Schmidt et al., 2009; Schubert et al., 2009), fatigue (Simon et al., 2011) and the inhibition of actions (Klimesch et al., 2007). Most notably, alpha power is modulated by the amount of relaxation in the visual system and is strongest when the eyes are closed. This is in line with the "idling hypothesis", which states that observable brain rhythms reflect the default mode of the brain, in which information is exchanged within functional networks in a loop (Palva and Palva, 2007; Sabate et al., 2011). This hypothesis is also valid for the weaker central alpha rhythm, called mu rhythm. Mu-rhythmic activity reflects the level of engagement in the sensorimotor system. There are actually multiple mu rhythms, since every body part is associated with a distinct generator of mu rhythmic activity in the sensorimotor cortex. These individual rhythms have been found to desynchronize during observed, planned, executed and even imagined movements of the respective body parts (Pfurtscheller and Lopez da Silva, 1999). Each of the mu rhythms has a distinct spatial signature in the EEG, which depends on where on the somatosensory cortex the respective body part is represented. This is used, e. g., to distinguish different types of motor imagery in brain-computer interfacing and related disciplines (e. g., Blankertz et al., 2007; Müller et al., 2008; Blankertz et al., 2008, 2010).

Although much is known about the functional relevance of the alpha and mu rhythms and the locations of their generating brain networks, the pathways of neural information transfer within and between these networks are largely unknown. The resting state offers an excellent opportunity for studying these interactions due to the strong intensity and almost stationary behaviour of alpha and mu oscillations in this condition. Currently, even simple questions related to whether information predominantly flows from frontal to occipital or from occipital to frontal brain structures, have not found consistent answers (Ito et al., 2005; Oishi et al., 2007; Buckner and Vincent, 2007; Nolte et al., 2008). In a recent paper by Nolte et al. (2008), this issue has been investigated by applying the phase-slope index (PSI) to EEG sensor-space time series recorded in the "eyes-closed" condition. The results acquired from 88 subjects indicate a clear front-to-back information flow.

In this chapter, we analyze resting state connectivity on the source level using the inverse source reconstruction methodology developed in Chapter 4. We start, however, by applying PSI in sensor space analogously to Nolte et al. (2008) with the aim of establishing consistency of the results obtained on the different datasets used in the two studies. In the second part of the chapter, we perform connectivity analysis using PSI on source estimates obtained from S-FLEX inverse source

reconstruction. We introduce the datasets in Section 6.1, present the sensor- and source-space analyses in Sections 6.2 and 6.3, respectively, and provide a discussion of the neurophysiological implications of our study in Section 6.4. This study is in preparation for publication.

6.1 Datasets

We consider data acquired from 29 right-handed healthy volunteers (20 males, 9 females; mean age 29.2 years; range: 23–49 years) as part of a baseline measurement embedded in an in-car EEG-study on attentional processes (Schmidt et al., 2009). The participants were screened for a variety of exclusion criteria (left-handedness, auditory and visual disabilities and various illnesses), instructed to sleep normally the night before the experiment and to refrain from consuming caffeine in the morning on the day of the experiment. They were compensated for their time with a gift worth approximately € 25. The section of the data we analyze is a 5-minute relaxation measurement recorded prior to the experiment. During this recording, the participants sat relaxedly in the driver's seat with their eyes closed. The engine and all electronic devices apart from the EEG instrumentation were switched off. Electroencephalography was recorded from 128 electrodes (extended 10-20 system, 1 000 Hz sampling rate, low cut-off: 0.016 Hz; high cut-off: 250 Hz, nose reference) using BrainAmp recording hardware (Brainproducts GmbH, Munich). From these electrodes, a subset of 59 electrodes was found to be free of severe technical artifacts in all recordings. This set was considered for grand-average data analysis. Finally, the EEG signal was lowpass-filtered to 50 Hz and down-sampled to 100 Hz. We refer to this dataset as the Schmidt et al. dataset.

Nolte et al. (2008) present similar data of 88 healthy participants, which were recruited randomly from the Swedish population register. During the experiment, which lasted for 15 minutes, the participants were instructed to relax and keep their eyes closed. Every minute, the participants were asked to open their eyes for 5 s. Electroencephalography was measured with a standard 10-20 system consisting of 19 channels using a linked-mastoids reference.

6.2 Sensor-space analyses

6.2.1 Setting

Figure 6.1 (e) is a reproduction of Figure 4 of Nolte et al. (2008), where the color scheme is adjusted to fit the style of this document and average PSI scores reported in there are transformed into z-scores by multiplying them with the square root of the number of datasets analyzed. As in the head-in-head plots presented previously in this thesis, each small circle of Figure 6.1 (e) depicts estimated information transfer of the corresponding EEG electrode with all other electrodes, where blue and cyan colors denote information inflow and red and yellow colors information outflow. The front-to-back flow reported by Nolte et al. is clearly visible.

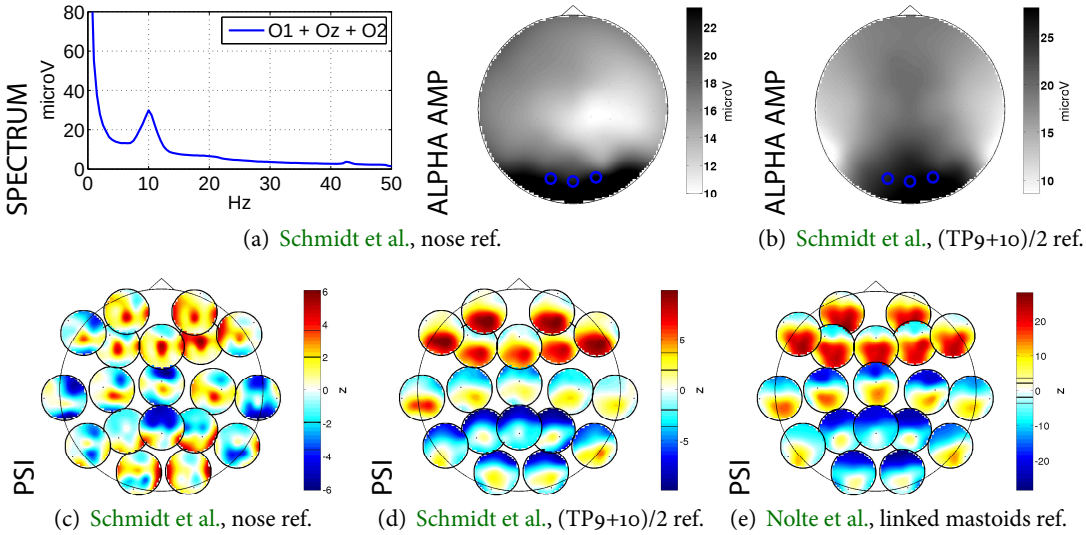
We analyze the data by Schmidt et al. here using the methodology of Nolte et al. as follows. The phase-slope index is calculated for each participant in a 5 Hz band centered at the individual alpha peak frequency, where the peak is determined as the frequency with maximum average spectral power at the electrodes O1, O2 and Oz. By averaging the connectivity scores obtained from

analyzing 2 s intervals and dividing the average by its estimated standard deviation, a standardized PSI score is obtained for each participant, which is treated as a z-score. The z-scores obtained from individual participants are aggregated to yield a grand-average z-score, which is depicted in Figure 6.1 (c). An equivalent analysis is conducted for a re-referenced version of the dataset. Here, the average activity of the temporally-located electrodes TP9 and TP10 is subtracted from each electrode before conducting the computations. To ensure consistency with Nolte et al. (2008), we use only the 19 electrodes defined in the standard 10-20 system to produce these plots. The number of multiple comparisons is hence $(19 \cdot 18)/2 = 171$ and the Bonferroni-corrected significance threshold is $z = 3.6$.

We also compute the grand-average spectrum at occipital channels, as well as the average signal amplitude in the alpha band for all electrodes. The results are shown in Figure 6.1 (a) and (b). The spectrum is computed by applying the Fourier transform in 2 s intervals and by averaging the amplitudes of the Fourier coefficients frequency-bin-wise across repetitions, participants and the three electrodes O1, O2 and Oz. Electrode-wise alpha band amplitudes are calculated per subject both for the original nose-referenced data and the re-referenced data by averaging the amplitudes of the Fourier coefficients across the frequency bins corresponding to the individual alpha range. A grand-average estimate of the alpha amplitude is obtained by averaging the participants' individual amplitudes.

6.2.2 Results

Figure 6.1 (a) depicts the grand-average amplitude spectrum, averaged over O1, O2 and Oz, which clearly exhibits the peak of the alpha rhythm at 10 Hz. Grand-average amplitudes in a 5 Hz band around the alpha peak (shown next to the spectrum as a scalp map) are highest at the occipital channels (O1, O2 and Oz are marked in blue). This does not change substantially, when the data is transformed from the nose reference to an artificial reference simulating the linked-mastoids reference, in which the mean potential at TP9 and TP10 is subtracted from the data (see Figure 6.1 (b)). However, the impact of the re-referenciation on PSI sensor-space connectivity estimates is huge (c. f. Figure 6.1 (c) and (d)). The simulated linked-mastoids reference using TP9 and TP10 as reference electrodes (Figure 6.1 (c)) leads to results that are similar to those obtained by Nolte et al. based on data recorded in genuine linked-mastoids reference (Figure 6.1 (e)). There is significant ($z > 3.6$) interaction in 66 of the 171 electrode pairs, where the flow predominantly passes from frontal to occipital channels. The nose-referenced data yields a substantially different head-in-head plot (Figure 6.1 (d)), in which a general flow from front to back is not so obvious. Rather, the picture is more complex, suggesting that frontal and occipital electrodes are drivers, while central and temporal electrodes are receivers. In total, the analysis of nose-referenced data yields only 22 significant ($z > 3.6$) interactions. Only nine significant connections are shared between nose- and simulated linked-mastoids-referenced data. This result demonstrates, that the ambiguity introduced by the reference electrode, which is noted in the simulation study described in Section 3.6, poses a problem that is of practical relevance in the sensor-space connectivity analysis of real data.



(a) Schmidt et al., nose ref. (b) Schmidt et al., (TP9+10)/2 ref.

(c) Schmidt et al., nose ref. (d) Schmidt et al., (TP9+10)/2 ref. (e) Nolte et al., linked mastoids ref.

Figure 6.1: Grand-average analysis of resting state EEG data in the "eyes-closed" condition. (a) Analysis of nose-referenced data. Left panel: amplitude spectrum, averaged over three occipital electodes. Right panel: alpha band amplitude per electrode with occipital electrodes (O1, O2 and Oz) marked in blue. (b) Alpha band amplitude per electrode for (TP9+10)/2-referenced data, simulating a linked-mastoids reference. (c-d) Estimated sensor-space connectivity maps according to PSI visualized as head-in-head plots, where red and yellow colors ($z > 0$) stand for information outflow and blue and cyan colors ($z < 0$) stand for information inflow. The Bonferroni-corrected significance level is indicated by a thin black line in the colorbar, while the uncorrected significance level is indicated by a thick black line. (c) Analysis of nose-referenced data by Schmidt et al.. (d) Analysis of the same data after re-referencing the data by subtracting the average activity at the TP9 and TP10 electrodes. (e) Re-analysis of data used in Nolte et al., which were recorded using linked-mastoids reference.

6.3 Source connectivity analysis

6.3.1 Setting

In order to obtain meaningful and reference-independent estimates of information flow, we perform source connectivity analysis as proposed in Section 4.6. To this end we apply inverse source reconstruction using sparse basis field expansions (S-FLEX) to the EEG time series as a preprocessing step. Prior to source localization, we apply bandpass-filtering in the individual 5 Hz alpha range to the EEG time series in order to filter out signals of non-neural origin (artifacts), as well as neurophysiological signals that are not of interest here. This is achieved by means of a fifth order digital Butterworth filter.

As in Section 4.6, we consider a two-step approach to localizing the whole time series. In the first step, 100 samples are randomly selected and jointly localized using S-FLEX. Here, we use the familiar "Montreal" head model with a source space of 2 142 dipoles that are arranged in a 1 cm grid, and we assume standard electrode positions according to the extended 10-20 system.

The standard deviations of the spatial basis function are set to $\varsigma_1 = 0.75$, $\varsigma_2 = 1$ and $\varsigma_3 = 1.25$, while the regularization constant λ is adjusted such that the source estimate achieves the same goodness-of-fit as a corresponding cross-validated weighted minimum-norm (WMN) estimate. The second estimation step is performed for all time samples using only the subset of spatial basis functions that are selected in the first step. Here, λ is again set to match the goodness-of-fit of a corresponding cross-validated WMN estimate.

We calculate grand-average voxel-wise source strengths by averaging the amplitudes of the source dipole moment vectors across time and participants. This yields a heat map of source activity in the alpha band (see Figure 6.2), from which we identify twelve active hotspots as the local maxima with strongest source amplitude.

We calculate PSI in the individual alpha band between the twelve local maxima and all other voxels using the same procedure as in sensor space. The resulting $3 \cdot 12 \times 3 \cdot 2\,142$ matrix of net inter- and intra-voxel connectivity scores is reduced to a $12 \times 2\,142$ matrix reflecting only inter-voxel relations by summing over the entries of the 3×3 submatrices related to each pair of voxels. Finally, PSI scores are transformed into z-scores and averaged across participants to yield grand-average statistical maps of source-space connectivity, which are depicted in Figure 6.2. The number of multiple comparisons is$(12 \cdot 11)/2 = 66$ and the Bonferroni-corrected significance threshold is hence $z = 3.4$.

6.3.2 Results

The local maxima of alpha band activity are more or less symmetrically distributed across the two hemispheres. We note two non-lateralized source regions, of which one is frontally (C1) located and one is occipitally (C12) located. The remaining active regions come in lateralized pairs, which are located frontally (C2 and C3), deep (C4 and C5), centrally (C6 and C7), deep occipitally (C8 and C9) and occipitally (C10 and C11). Figure 6.2 (AMP) depicts these source regions, where even numbers always refer to sources in the left hemisphere and odd numbers refer to sources in the right hemisphere. Although the sources' locations are almost symmetric, the mean alpha amplitude is stronger in the right hemisphere for the frontal, central and occipital sources, which cover the sensorimotor and the visual areas (see Section 2.1).

The analysis of PSI between the source regions reveals significant ($z > 3.4$) flow from C1 to C2, C3, C4, C5, C6, C7, C9, C10 and C12, from C2 to C7 and C9, from C3 to C6, C8 and C10, from C4 to C2, C5, C6, C7 and C10, from C5 to C3, C6, C7, C9, C10 and C12, from C6 to C9 and C12, from C7 to C9, C11 and C12, from C8 to C2, C4, C9 and C10, and from C11 to C1, C2, C3, C4, C9 and C12. That is, there indeed appears to be information flow from frontal to (deep) occipital sources but this is by far not the only interaction that is present. Most notably, there is also significant flow in the other direction, as a source in the right occipital cortex (C11) is estimated to send information to six regions, which are all located more frontally.

Strikingly, the estimated information pathways are not exclusively symmetric, i. e., some contralateral source pairs do not share the same connectivity pattern. For example, the source region C10, which is the contralateral analogon of C11, does not send information to any other region.

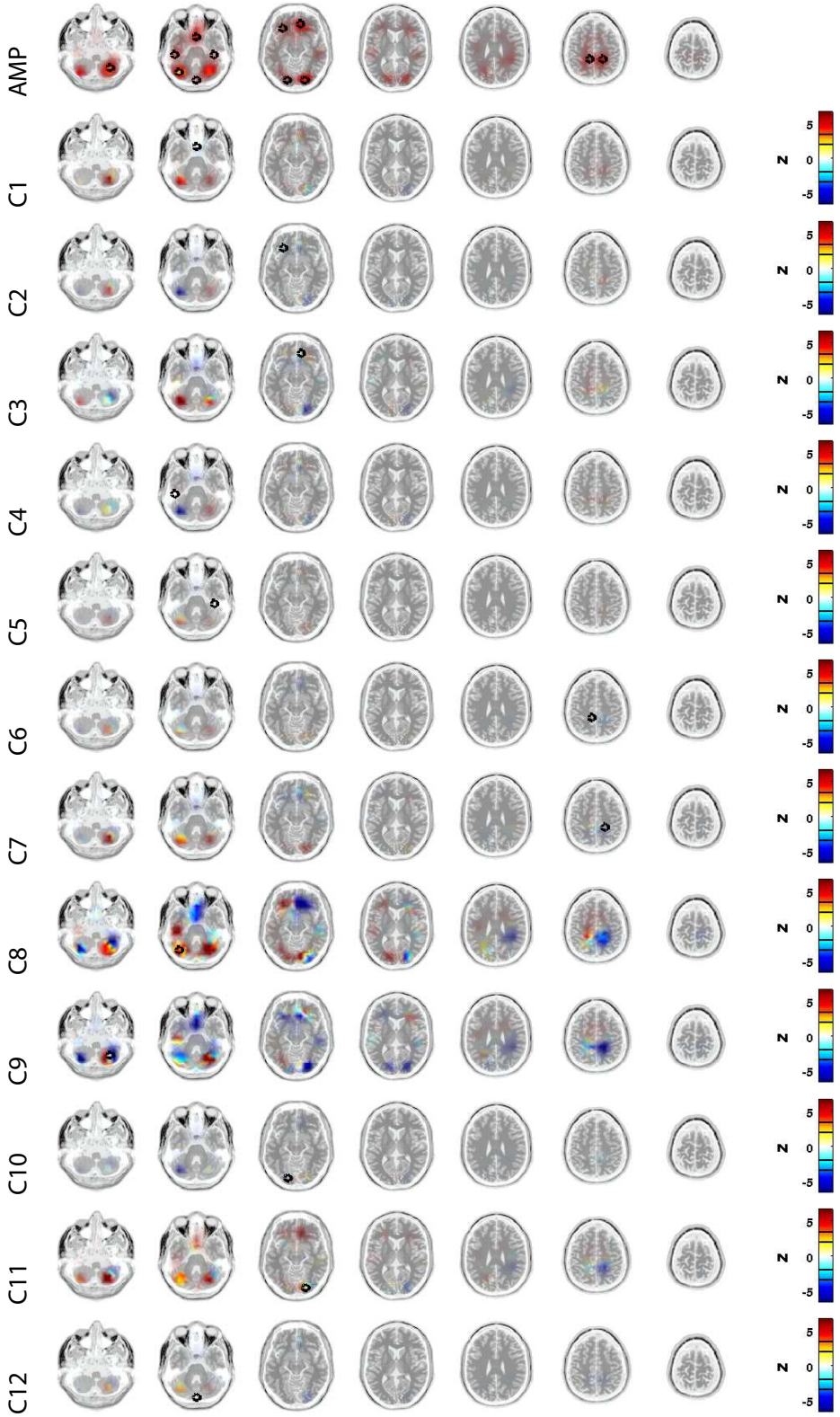
6.4 Discussion (asymmetries in brain default networks)

The results presented in this chapter provide a number of data-analytic as well as neurophysiological insights that are worth mentioning. First of all, our sensor-space analyses demonstrate the influence of the choice of the reference electrode when analyzing PSI (as well as other connectivity measures) on raw EEG time series. Note that it is not possible to estimate the amount of signal the reference electrode picks from the various driving and receiving sources, which would be crucial in order to compensate for its influence, without performing explicit inverse calculations. To overcome the reference problem, the use of reference-free derivations such as, e. g., the scalp Laplacian is conceivable but the interpretation of head-in-head plots obtained from Laplace-transformed data is also difficult. Genuine inverse source reconstruction preprocessing, on the other hand, not only provides reference-independent source estimates but also maps the sources activity to realistic brain anatomy, which facilitates neurophysiological interpretation.

As was demonstrated in Chapter 4, the use of sparse basis field expansions (S-FLEX) is highly advantageous regarding subsequent source connectivity estimation, since S-FLEX is able to spatially distinguish between multiple possibly correlated sources. The application of S-FLEX to resting state data here indeed indicates a multitude of source regions that are active in the alpha range. Notably, alpha activity is found to stronger in the right hemisphere than in the left hemisphere. With respect to the idling hypothesis, this finding indicates a potentially higher relaxation state of the systems represented in the right hemisphere, which are exactly the left body parts (Pfurtscheller and Lopez da Silva, 1999). Put differently, it is the left hemisphere (responsible for the right body half) that maintains vigilance during the rest condition, which is in line with the right hand being dominant for all subjects.

Note that we refrain here from discussing anatomical issues in more exact terms such as "occipital" and "frontal". The reason for this is the fact that our analysis is based on the "Montreal" head model, which has been obtained from a single head that is not representative of the population studied here. Indeed, this imprecision leads to the result that some of the estimated source regions are localized in rather improbable brain regions. For example, the source regions C8 and C9 are located slightly below the visual cortex – where they would have been expected – in the cerebellum. In absence of individual head models of the participants included in our study, the present analysis therefore has to bee seen as an intermediate step on the way towards understanding the human alpha rhythm in terms of its interacting subsystems. Nevertheless, it is possible to investigate connectivity of gross brain regions such as the main brain lobes using the present results, which is a clear improvement over sensor-space analysis.

Figure 6.2 *(following page)*: Analysis of alpha band grand-average source-space connectivity using sparse basis field expansion (S-FLEX) inverse preprocessing and the phase-slope index (PSI). AMP: Alpha band source amplitudes obtained by localizing bandpass-filtered EEG time series using S-FLEX. Brighter colors encode larger amplitudes. The twelve strongest local maxima of the amplitude distribution are marked by black circles. C1–C12: Effective connectivity between twelve active brain regions and all other regions as estimated by applying PSI to source time series estimated by S-FLEX. Red and yellow colors ($z > 0$) stand for information outflow and blue and cyan colors ($z < 0$) stand for information inflow according to PSI. The Bonferroni-corrected significance level is indicated by thin black line in the colorbar, while the z-score corresponding to an uncorrected p-value of 0.05 is indicated by a thick black line.


AMP
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
z
5
0
-5

The estimated information flow between active regions reveals highly interesting patterns, the complex structure of which by no means have been reproduced using mere sensor-space analysis, or source estimates obtained from linear inverse source reconstruction. While we leave it to a future publication to relate our findings to existing results on resting state neurophysiology, we do note that there appears to exist a densely-interconnected network especially in the occipital part of the brain, where the visual cortex resides. In this region, the roles of effective (net) drivers and receivers occasionally change within centimeters. Importantly, there are a number of asymmetries regarding the estimated information flow between source regions. While we investigated only the "eyes-closed" condition here, it will be interesting to repeat our analysis for similar data recorded in the "eyes-open" condition, and to possibly relate the connectivity patterns to the phenomenon of ocular dominance (Porac and Coren, 1975; Chaurasia and Mathur, 1976).

The results presented here were obtained using an inverse source reconstruction preprocessing methodology. It may be equally insightful to study the data using blind-source separation (BSS) or integrative BSS/inverse source reconstruction approaches such as those outlined in Chapter 5, which is the subject of our ongoing work. The challenge to meet here in order to provide a similar grand-average assessment of source connectivity is to meaningfully combine results obtained from multiple subjects, which is more difficult than in a simulated scenario due to the high inter-subject variability of the BSS components.

7 Summary and Conclusions

With this piece of work we have extended the field of EEG-based brain effective connectivity analysis with methods which model EEG data in terms of underlying interacting brain *sources*. We followed two general directions: inverse source reconstruction and blind source separation. In both of these disciplines we presented novel source estimation algorithms, which are well-suited for the purpose of conducting source connectivity analysis. Our methods explicitly account for the influence of volume conduction on sensor data as well as for dependencies between the underlying brain sources. As a result, they outperform the state-of-the-art in terms of reconstructing interacting brain sources, which is a crucial requirement for successful subsequent source connectivity analysis. In this regard, our efforts can be regarded as an important step *towards EEG source connectivity analysis*.

After giving an overview of state-of-the-art approaches to the EEG-based study of brain connectivity in Chapter 2, we presented an exhaustive evaluation of these approaches on simple yet realistically simulated EEG data involving two interacting sources in Chapter 3. As a result of the study, we identified a number of factors that are not related to actual time-delayed interaction of the sources but influence the results of common measures of effective connectivity. A fundamental difficulty for all measures is the fact that volume conduction mixes the underlying sources linearly into the EEG. For the particular class of Granger-causal measures, we noted that this mixing, in combination with asymmetries in the SNR, gives rise to effects that are not related to time-delayed interaction. In addition, Granger-causal measures derived directly from the coefficients of a multivariate AR model were found to suffer from asymmetries in the scaling and correlation structure of the data.

We proposed a test based on time inversion, by means of which the presence of spurious results related to such asymmetries can be identified. Moreover, we noted that even the phase-slope index, which is robust to all of these asymmetries, is hard to interpret when applied in sensor space, since not only the location and orientation of the underlying current sources but also the signal-to-noise ratio and the choice of the reference electrode affects sensor-space connectivity estimates in a way that is difficult to assess without performing explicit inverse calculations. On the other hand, all connectivity measures considered are found to perform well on unmixed sources. Thus, the problem of source connectivity estimation is closely tied to the problem of finding the correct source demixing.

Importantly, our study suggests that there are only a few demixing methods available yet, which are both theoretically and practically suited for source connectivity analysis. We demonstrated that the Laplace transform is incapable of demixing deep and tangentially-oriented sources, which limits its use in EEG connectivity studies. For inverse source reconstruction, we found that smooth (linear) approaches are unable to sufficiently separate the underlying sources, which causes many of the problems experienced in sensor space to remain in source space. Sparse estimators, on the other hand, do not currently reflect certain fundamental assumptions on current source distributions such as spatial continuity and rotational invariance and are not yet available for

the analysis of long time series. Regarding blind source separation methods, we noted that the most popular approaches PCA and ICA are unsuitable for source connectivity analysis owing to their assumptions of orthogonality and independence, respectively. Methods that (implicitly or explicitly) allow for dependencies of the underlying sources, such as MVARICA and CICAAR, are better suited and indeed outperform PCA and ICA approaches in terms of demixing two interacting sources if certain statistical assumptions are fulfilled. In summary, the results of our simulation study show that EEG-based connectivity analysis poses a challenging problem even in the presence of only two interacting sources.

The central contribution of this thesis is the development of source demixing algorithms that both theoretically qualify for source connectivity analysis and practically feasible. Such methods are described in Chapters 4 and 5.

In Chapter 4, we have introduced an inverse source reconstruction methodology, focal vector-field reconstruction (FVR), which is able to recover and distinguish multiple sources, the spatial distributions of which are characterized by smoothness and focality at the same time. Unlike previous approaches involving sparsity constraints, FVR ensures rotational invariance of the solution. Moreover, it employs a novel depth-weighting strategy for compensating the location bias common to previous approaches. Due to these properties, FVR outperforms existing approaches in its ability to reconstruct two to three dipolar sources of arbitrary depth (Haufe et al., 2008). We quantified the superior reconstruction quality by means of the earth mover's distance (EMD), which we suggest as a metric for comparing arbitrary current distributions.

In the course of Chapter 4, we extended our inverse source reconstruction methodology by introducing sparse basis field expansions (S-FLEX), which employs the same depth weighting as FVR and achieves rotational invariance in a similar way but is able to recover sources of arbitrary shape without requiring one to explicitly adjust the relative importance of smoothness and focality, respectively. We provide an optimization scheme, by means of which it is possible to compute S-FLEX source estimates for multiple measurements. The S-FLEX approach hence facilitates the reconstruction of whole source time series for subsequent connectivity analysis.

The favorable localization properties of S-FLEX and FVR compared to state-of-the-art approaches were demonstrated on simulated data involving multiple sources of various shapes and depths using the EMD as a performance metric. In addition, both approaches were shown to be able to identify and separate the active brain regions related to the processing of sensory stimulation at the left and right hands, unlike reference methods. Applied to the analysis of simulated EEG time series reflecting the activity of interacting brain sources, S-FLEX correctly localizes and distinguishes the two generating source regions, which is a clear improvement over linear inverses. As a result of the successful demixing, the direction of information flow between S-FLEX sources is correctly estimated by all of the measures of effective connectivity considered.

Chapter 5 started with the development of Granger-causal measures of effective connectivity, which facilitate the estimation of parsimonious effective connectivity graphs either through optimization under appropriate sparsity constraints or through subsequent multiple statistical testing. Since sparse connectivity is an essential feature of the brain, at least if it is analyzed on a macroscopic level, we derived a BSS method called sparsely-connected sources analysis (SCSA), which utilizes this assumption to recover the sources, their mixing patterns and a sparse connectivity graph in one step. An interesting theoretical result is that a completely dense source connectivity graph corresponds to the convolutive ICA model, while a graph without connections

between sources models independence of the source components in a similar way as variants of instantaneous ICA such as TDSEP do. These two extremes are special cases of SCSA, which correspond to specific choices of an adjustable hyperparameter of SCSA that effectively constrains the number of connections.

As for MVARICA and CICAAR, successful demixing using SCSA depends on whether assumptions regarding the non-Gaussianity of the data are fulfilled. We here observed that if the data indeed fulfills the non-Gaussianity requirement, SCSA outperforms MVARICA and CICAAR and all other BSS methods tested in terms of demixing two interacting sources. Similar results were obtained for more than two sources and for more than one connection regardless of the type of noise and the signal-to-noise ratio. Importantly, our simulations also show that the accuracy of the demixing has a crucial influence on the correctness of the source connectivity estimation as well as the localization of the source components. In both terms, SCSA was found to outperform competing BSS methods.

To summarize, we presented two solutions to the problem of demixing interacting EEG sources, which differ conceptually in the type of assumptions used. Meanwhile, we are working on a novel approach called SCSA+FLEX, which combines both types of assumptions in a hybrid BSS/inverse source reconstruction methodology. The advantage of such a hybrid approach is that the fusion of assumptions reduces the ambiguity inherent to BSS and inverse source reconstruction problems.

Technically, our approaches make extensive use of groupwise sparsity via $\ell_{1,2}$-norm regularization, which is used as a versatile tool for encoding different types of prior expectations. In sparse Granger-causal modeling, as well as in SCSA and SCSA+FLEX, MVAR coefficients related to the same pair of times series are grouped together in order to achieve sparsity of the connectivity graph. In FVR and S-FLEX, sparsity is used as a means to obtain current densities with simple and therefore physiologically plausible spatial structure. The grouping of the three variables related to a spatial basis function here naturally encodes the physiologically-motivated requirement of rotational invariance. In multi-measurement localization, these groups are furthermore naturally extended to encode the assumption that the set of active source regions is constant over time. Except FVR, all of our methods apply groupwise sparsification directly to the variables which are to be optimized. In this case, the dual augmented Lagrangian (DAL) technique can be used to solve large-scale problem instances with arbitrary convex loss functions. While not all of our proposed loss functions are convex, it is always possible to split the variables into groups, which are optimized one after another in a loop, similarly to the expectation propagation algorithm. In this way, we obtain subproblems that are either smooth or convex and can be solved efficiently.

Our methods are suited to be applied to real world problem instances, which was demonstrated for the S-FLEX-based approach in Chapter 6. Our analysis of effective connectivity between the sources of the human alpha rhythm yields a number of novel insights which could not have been obtained from mere sensor-space analysis. We identified a number of alpha band sources, which are mostly symmetrically arranged on the cortex. Importantly, the corresponding interaction patterns are *not* exclusively symmetric, which supports the hypothesis of a consistent dominant laterality in humans. While our findings still remain to be affirmed by alternative methodologies such as SCSA and SCSA+FLEX, we are confident that they can contribute to a better understanding of the neurodynamics underlying the human alpha rhythm. We also expect that future EEG studies will benefit from the novel type of source connectivity analysis proposed in this work.

Bibliography

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans Automat Contr 19, 716–723.

Anemüller, J., Sejnowski, T. J., Makeig, S., 2003. Complex independent component analysis of frequency-domain electroencephalographic data. Neural Networks 16, 1311–1323.

Astolfi, L., Bakardjian, H., Cincotti, F., Mattia, D., Marciani, M. G., De Vico Fallani, F., Colosimo, A., Salinari, S., Miwakeichi, F., Yamaguchi, Y., Martinez, P., Cichocki, A., Tocci, A., Babiloni, F., 2007. Estimate of causality between independent cortical spatial patterns during movement volition in spinal cord injured patients. Brain Topogr 19, 107–123.

Astolfi, L., Cincotti, F., Mattia, D., De Vico Fallani, F., Salinari, S., Ursino, M., Zavaglia, M., Marciani, M. G., Babiloni, F., 2006a. Estimation of the cortical connectivity patterns during the intention of limb movements. IEEE Eng Med Biol 25, 32–38.

Astolfi, L., Cincotti, F., Mattia, D., Marciani, M. G., Baccalà, L. A., de Vico Fallani, F., Salinari, S., Ursino, M., Zavaglia, M., Babiloni, F., 2006b. Assessing cortical functional connectivity by partial directed coherence: simulations and application to real data. IEEE Trans Biomed Eng 53, 1802–1812.

Astolfi, L., Cincotti, F., Mattia, D., Salinari, S., Babiloni, C., Basilisco, A., Rossini, P. M., Ding, L., Ni, Y., He, B., Marciani, M. G., Babiloni, F., 2004. Estimation of the effective and functional human cortical connectivity with structural equation modeling and directed transfer function applied to high-resolution EEG. Magn Reson Imaging 22, 1457–1470.

Attias, H., Schreiner, C. E., 1998. Blind source separation and deconvolution: the dynamic component analysis algorithm. Neural Comput 10, 1373–1424.

Babiloni, C., Ferri, R., Moretti, D. V., Strambi, A., Binetti, G., Dal Forno, G., Ferreri, F., Lanuzza, B., Bonato, C., Nobili, F., Rodriguez, G., Salinari, S., Passero, S., Rocchi, R., Stam, C. J., Rossini, P. M., 2004. Abnormal fronto-parietal coupling of brain rhythms in mild Alzheimer's disease: a multicentric EEG study. Eur J Neurosci 19, 2583–2590.

Babiloni, F., Cincotti, F., Babiloni, C., Carducci, F., Mattia, D., Astolfi, L., Basilisco, A., Rossini, P. M., Ding, L., Ni, Y., Cheng, J., Christine, K., Sweeney, J., He, B., 2005. Estimation of the cortical functional connectivity with the multimodal integration of high-resolution EEG and fMRI data by directed transfer function. NeuroImage 24, 118–131.

Baccalá, L. A., Sameshima, K., 2001. Partial directed coherence: a new concept in neural structure determination. Biol Cybern 84, 463–474.

Baillet, S., Mosher, J. C., Leahy, R. M., 2001. Electromagnetic brain mapping. IEEE Signal Proc Mag 18, 14–30.

Beckmann, C. F., DeLuca, M., Devlin, J. T., Smith, S. M., 2005. Investigations into resting-state connectivity using independent component analysis. Philos Trans Roy Soc B 360, 1001–1013.

Belouchrani, A., Abed-Meraim, K., Cardoso, J. F., Moulines, E., 1997. A blind source separation technique using second-order statistics. IEEE Trans Signal Proces 45, 434–444.

Berger, H., 1938. Das Elektrenkephalogramm des Menschen. Nova Acta Leopoldina 6, 173–309.

Berger, J. O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd Edition. Springer.

Bishop, C. M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st Edition. Springer.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Curio, G., 2007. The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects. NeuroImage 37, 539–550.

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

Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L. E., Sturm, I., Curio, G., Müller, K.-R., 2010. The Berlin Brain-Computer Interface: Non-medical uses of BCI technology. Front Neuroscience 4, 198.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.-R., 2008. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Proc Mag 25, 41–56.

Blinowska, K., Kus, R., Kaminski, M., Janiszewska, J., 2010. Transmission of brain activity during cognitive task. Brain Topogr 23, 205–213.

Bolstad, A., Van Veen, B., Nowak, R., 2009. Space-time event sparse penalization for magneto-/electroencephalography. NeuroImage 46, 1066–1081.

Bonferroni, C. E., 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8, 3–62.

Brockwell, P. J., Davis, R. A., 1998. Time Series: Theory and Methods (Springer Series in Statistics). Springer.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., Bressler, S. L., 2004. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. PNAS 101, 9849–9854.

Buckner, R. L., Vincent, J. L., 2007. Unrest at rest: default activity and spontaneous network correlations. NeuroImage 37, 1091–1096.

Butterworth, S., 1930. On the Theory of Filter Amplifiers. Wireless Engineer 7, 536–541.

Calhoun, V. D., Adali, T., Pearlson, G. D., Pekar, J. J., 2001. A method for making group inferences from functional MRI data using independent component analysis. Hum Brain Mapp 14, 140–151.

Cardoso, J.-F., Souloumiac, A., 1996. Jacobi angles for simultaneous diagonalization. SIAM J Matrix Anal Appl 17, 161–164.

Carlsson, A., Lindqvist, M., Magnusson, T., 1957. 3,4-Dihydroxyphenylalanine and 5-hydroxytryptophan as reserpine antagonists. Nature 180, 1200.

Chaurasia, B. D., Mathur, B. B., 1976. Eyedness. Acta Anat 96, 301–305.

Dale, A. M., Sereno, M. I., 1993. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. J Cognitive Neurosci 5, 162–176.

Dale, H. H., 1914. The action of certain esters and ethers of choline, and their relation to muscarine. J Pharmacol Exp Ther 6, 147–190.

Davis, P. A., 1939. Effects of acoustic stimuli on the waking human brain. J Neurophysiol 2, 494–499.

Deiters, O., 1865. Untersuchungen über Gehirn und Rückenmark des Menschen und der Säugethiere. Vieweg.

Ding, L., He, B., 2008. Sparse source imaging in EEG with accurate field modeling. Hum Brain Mapp 29, 1053–1067.

Dyrholm, M., Makeig, S., Hansen, L. K., 2007. Model selection for convolutive ICA with an application to spatiotemporal analysis of EEG. Neural Comput 19, 934–955.

Eccles, J., 1964. The physiology of synapses. Springer.

Eichele, T., Debener, S., Calhoun, V. D., Specht, K., Engel, A. K., Hugdahl, K., von Cramon, D. Y., Ullsperger, M., 2008. Prediction of human errors by maladaptive changes in event-related brain networks. PNAS 105, 6173–6178.

Eichler, M., 2005. A graphical approach for evaluating effective connectivity in neural systems. Philos Trans Roy Soc B 360, 953–967.

Engl, H. W., Hanke, M., Neubauer, A., 2000. Regularization of inverse problems. Kluwer Academic Publishing.

Erlanger, J., Gasser, H. S., 1924. The compound nature of the action current of nerve as disclosed by the cathode ray oscillograph. Am J Physiol 70, 624–666.

Fabiani, M., Gratton, G., Coles, M. G. H., 2000. Event-related brain potentials - methods, theory, and applications. In: Cacioppo, J. T., Tassinary, L. G., Bertson, G. G. (Eds.), Handbook of psychophysiology. Cambridge University Press, pp. 55–83.

Farwell, L. A., Donchin, E., 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr Clin Neurophysiol 70, 510–523.

Friston, K., 2009. Causal modelling and brain connectivity in functional magnetic resonance imaging. PLoS Biol 7, e1000033.

Friston, K. J., 1994. Functional and effective connectivity in neuroimaging: A synthesis. Hum Brain Mapp 2, 56–78.

Friston, K. J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.

Genz, A., 1992. Numerical computation of multivariate normal probabilities. J Comput Graph Stat 1, 141–150.

Golgi, C., 1885. Sulla fina anatomia degli organi centrali del sistema nervoso. Regio Emilia Riv Sper Feniatr 1, 405–425.

Gómez-Herrero, G., Atienza, M., Egiazarian, K., Cantero, J. L., 2008. Measuring directional coupling between EEG sources. NeuroImage 43, 497–508.

Gorodnitsky, I. F., George, J. S., Rao, B. D., 1995. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. Electroencephalogr Clin Neurophysiol 95, 231–251.

Gosset, W. S., 1908. The probable error of a mean. Biometrika 6, 1–25, originally published under the pseudonym "Student".

Gotch, F., 1902. The submaximal electrical response of nerve to a single stimulus. J Physiol (London) 28, 395–416.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., Lin, F. H., 2008. Lexical influences on speech perception: a Granger causality analysis of MEG and EEG source estimates. NeuroImage 43, 614–623.

Gramfort, A., Papadopoulo, T., Baillet, S., Clerc, M., 2011a. Tracking cortical activity from M/EEG using graph cuts with spatiotemporal constraints. NeuroImage 54, 1930–1941.

Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M., Kowalski, M., 2011b. Functional brain imaging with M/EEG using structured sparsity in time-frequency dictionaries. In: Székely, G., Hahn, H. (Eds.), Information Processing in Medical Imaging. Vol. 6801 of Lecture Notes in Computer Science. Springer, pp. 600–611.

Granger, C., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Grave de Peralta-Menendez, R., Gonzalez-Andino, S. L., 1998. A critical analysis of linear inverse solutions to the neuroelectromagnetic inverse problem. IEEE Trans Biomed Eng 45, 440–448.

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

Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V. J., Meuli, R., Thiran, J. P., 2007. Mapping human whole-brain structural networks with diffusion MRI. PLoS ONE 2, e597.

Hämäläinen, M., Ilmoniemi, R., 1994. Interpreting magnetic fields of the brain: minimum norm estimates. Med Biol Eng Comput 32, 35–42.

Hansen, L. K., Larsen, J., Nielsen, F. Å., Strother, S. C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., Paulson, O. B., 1999. Generalizable patterns in neuroimaging: how many principal components? NeuroImage 9, 534–544.

Hansen, P. C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev 34, 561–580.

Hastie, T., Tibshirani, R., Friedman, J. H., 2001. The elements of statistical learning: data mining, inference, and prediction. Springer.

Haufe, S., Nikulin, V., Ziehe, A., Müller, K.-R., Nolte, G., 2008. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. NeuroImage 42, 726–738.

Haufe, S., Nikulin, V. V., Ziehe, A., Müller, K.-R., Nolte, G., 2009. Estimating vector fields using sparse basis field expansions. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21. MIT Press, pp. 617–624.

Haufe, S., Nolte, G., Müller, K.-R., Krämer, N., 2010a. Sparse causal discovery in multivariate time series. In: Guyon, I., Janzing, D., Schölkopf, B. (Eds.), Causality: Objectives and Assessment. Vol. 6 of JMLR W&CP. pp. 97–106.

Haufe, S., Tomioka, R., Dickhaus, T., Sannelli, C., Blankertz, B., Nolte, G., Müller, K.-R., 2010b. Localization of class-related mu-rhythm desynchronization in motor imagery based brain-computer interface sessions. In: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE. pp. 5137–5140.

Haufe, S., Tomioka, R., Dickhaus, T., Sannelli, C., Blankertz, B., Nolte, G., Müller, K.-R., 2011a. Large-scale EEG/MEG source localization with spatial flexibility. NeuroImage 54, 851–859.

Haufe, S., Tomioka, R., Nolte, G., Müller, K.-R., Kawanabe, M., 2010c. Modeling sparse connectivity between underlying brain sources for EEG/MEG. IEEE Trans Biomed Eng 57, 1954–1963.

Haufe, S., Treder, M. S., Gugler, M. F., Sagebaum, M., Curio, G., Blankertz, B., 2011b. EEG potentials predict upcoming emergency brakings during simulated driving. J Neural Eng 8, 056001.

Hesse, W., Möller, E., Arnold, M., Schack, B., 2003. The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. J Neurosci Methods 124, 27–44.

Hodgkin, A. L., Huxley, A. F., 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol (London) 117, 500–544.

Højen-Sørensen, P. A., Winther, O., Hansen, L. K., 2002. Mean-field approaches to independent component analysis. Neural Comput 14, 889–918.

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A., Evans, A. C., 1998. Enhancement of MR images using registration for signal averaging. J Comput Assist Tomogr 22, 324–333.

Horwitz, B., 2003. The elusive concept of brain connectivity. NeuroImage 19, 466–470.

Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous Inference in General Parametric Models. Biometrical J 3, 346–363.

Huang, M.-X., Dale, A. M., Song, T., Halgren, E., Harrington, D. L., Podgorny, I., Canive, J. M., Lewise, S., Lee, R. R., 2006. Vector-based spatial-temporal minimum L1-norm solution for MEG. NeuroImage 31, 1025–1037.

Huttunen, J., Komssi, S., Lauronen, L., 2006. Spatial dynamics of population activities at S1 after median and ulnar nerve stimulation revisited: An MEG study. NeuroImage 32, 1024–1031.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Networks 13, 411–430.

Hyvärinen, A., Zhang, K., Shimizu, S., Hoyer, P. O., 2010. Estimation of a structural vector autoregression model using non-Gaussianity. JMLR 11, 1709–1731.

Ioannides, A. A., Bolton, J. P. R., Clarke, C. J. S., 1990. Continous probabilistic solutions to the biomagnetic inverse problem. Inverse Probl 6, 523–542.

Ito, J., Nikolaev, A. R., van Leeuwen, C., 2005. Spatial and temporal structure of phase synchronization of spontaneous alpha EEG activity. Biol Cybern 92, 54–60.

Janzing, D., Schölkopf, B., 2010. Causal inference using the algorithmic markov condition. IEEE Trans Inform Theory 56, 5168–5194.

Jeffreys, D. A., Axford, J. G., 1972. Source locations of pattern-specific components of human visual evoked potentials. II. Component of extrastriate cortical origin. Exp Brain Res 16, 22–40.

Jeffs, B., Leahy, R. M., Singh, M., 1987. An evaluation of methods for neuromagnetic image reconstruction. IEEE Trans Biomed Eng 34, 713–723.

Jirsa, V., McIntosh, A., 2007. Handbook of brain connectivity. Understanding complex systems. Springer.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., Sejnowski, T. J., 2000a. Removing electroencephalographic artifacts by blind source separation. Psychophysiology 37, 163–178.

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., Sejnowski, T. J., 2000b. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. Clin Neurophysiol 111, 1745–1758.

Kamiński, M., Blinowska, K., Szelenberger, W., 1997. Topographic analysis of coherence and propagation of EEG activity during sleep and wakefulness. Electroencephalogr Clin Neurophysiol 102, 216–227.

Kamiński, M., Ding, M., Truccolo, W. A., Bressler, S. L., 2001. Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. Biol Cybern 85, 145–157.

Kamiński, M. J., Blinowska, K. J., 1991. A new method of the description of the information flow in the brain structures. Biol Cybern 65, 203–210.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., 2000. Principles of Neural Science, 4th Edition. McGraw-Hill Medical.

Kawanabe, M., Winkler, I., Haufe, S., 2011 In preparation.

Kayser, J., Tenke, C. E., 2006. Principal components analysis of Laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks. Clin Neurophysiol 117, 348–368.

Kebabian, J. W., Greengard, P., 1971. Dopamine-sensitive adenyl cyclase: possible role in synaptic transmission. Science 174, 1346–1349.

Kiebel, S. J., Garrido, M. I., Moran, R. J., Friston, K. J., 2008. Dynamic causal modelling for EEG and MEG. Cogn Neurodyn 2, 121–136.

Kim, S.-J., Koh, K., Boyd, S. P., Gorinevsky, D. M., 2009. ℓ_1 trend filtering. SIAM Rev 51, 339–360.

Kincses, W. E., Braun, C., Kaiser, S., Grodd, W., Ackermann, H., Mathiak, K., 2003. Reconstruction of extended cortical sources for EEG and MEG based on a Monte-Carlo-Markov-Chain estimator. Hum Brain Mapp 18, 100–110.

Klem, G. H., Luders, H. O., Jasper, H. H., Elger, C., 1999. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 52, 3–6.

Klimesch, W., Sauseng, P., Hanslmayr, S., 2007. EEG alpha oscillations: the inhibition-timing hypothesis. Brain Res Rev 53, 63–88.

Koldovský, Z., Tichavský, P., Oja, E., 2006. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. IEEE Trans Neural Netw 17, 1265–1277.

Komssi, S., Huttunen, J., Aronen, H. J., Ilmoniemi, R. J., 2004. EEG minimum-norm estimation compared with MEG dipole fitting in the localization of somatosensory sources at S1. Clin Neurophysiol 115, 534–542.

Kornhuber, H. H., Deecke, L., 1965. Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Arch. 284, 1–17.

Kuhn, H. W., 1955. The Hungarian method for the assignment problem. Nav Res Log 2, 83–97.

Lauterbur, P. C., 1973. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. Nature 242, 190–191.

Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. NeuroImage 56, 387–399.

Lobo, M. S., Vandenberghe, L., Boyd, S., Lebret, H., 1998. Applications of second-order cone programming. Lin Alg Appl 284, 193–228.

Loewi, O., 1921. Über humorale Übertragbarkeit der Herznervenwirkung. Pflügers Archiv für die Gesamte Physiologie des Menschen und der Tiere 189, 239–242.

Makeig, S., Jung, T. P., Bell, A. J., Ghahremani, D., Sejnowski, T. J., 1997. Blind separation of auditory event-related brain responses into independent components. PNAS 94, 10979–10984.

Malioutov, D., Çetin, M., Willsky, A. S., 2005. A sparse signal reconstruction perspective for source localization with sensor arrays. IEEE Trans Signal Proces 55, 3010–3022.

Marinazzo, D., Pellicoro, M., Stramaglia, S., 2008. Kernel method for nonlinear Granger Causality. Phys Rev Lett 100, 144103.

Markram, H., 2006. The blue brain project. Nat Rev Neurosci 7, 153–160.

Marzetti, L., Del Gratta, C., Nolte, G., 2008. Understanding brain connectivity from EEG data by identifying systems composed of interacting sources. NeuroImage 42, 87–98.

Marzetti, L., Nolte, G., Perrucci, M. G., Romani, G. L., Del Gratta, C., 2007. The use of standardized infinity reference in EEG coherency studies. NeuroImage 36, 48–63.

Matsuura, K., Okabe, Y., 1995. Selective minimum-norm solution of the biomagnetic inverse problem. IEEE Trans Biomed Eng 42, 608–615.

McKeown, M. J., Hansen, L. K., Sejnowski, T. J., 2003. Independent component analysis of functional MRI: what is signal and what is noise? Curr Opin Neurobiol 13, 620–629.

Meinecke, F. C., Ziehe, A., Kurths, J., Müller, K.-R., 2005. Measuring Phase Synchronization of Superimposed Signals. Phys Rev Lett (8), 084102.

Molgedey, L., Schuster, H. G., 1994. Separation of a mixture of independent signals using time delayed correlations. Phys Rev Lett 72, 3634–3637.

Mørup, M., Hansen, L. K., Arnfred, S. M., Lim, L. H., Madsen, K. H., 2008. Shift-invariant multilinear decomposition of neuroimaging data. NeuroImage 42, 1439–1450.

Mørup, M., Hansen, L. K., Herrmann, C. S., Parnas, J., Arnfred, S. M., 2006. Parallel Factor Analysis as an exploratory tool for wavelet transformed event-related EEG. NeuroImage 29, 938–947.

Moseley, M. E., Cohen, Y., Kucharczyk, J., Mintorovitch, J., Asgari, H. S., Wendland, M. F., Tsuruda, J., Norman, D., 1990. Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system. Radiology 176, 439–445.

Mosher, J. C., Leahy, R. M., 1999. Source localization using recursively applied and projected (RAP) MUSIC. IEEE Trans Signal Proces 47, 332–340.

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., Blankertz, B., 2008. Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. J Neurosci Methods 167, 82–90.

Munkres, J., 1957. Algorithms for the assignment and transportation problems. J Soc Indust Appl Math 5, 32–38.

Murayama, Y., Weber, B., Saleem, K. S., Augath, M., Logothetis, N. K., 2006. Tracing neural circuits in vivo with Mn-enhanced MRI. Magn Reson Imaging 24, 349–358.

Neal, R., Hinton, G. E., 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan, M. I. (Ed.), Learning in Graphical Models. Kluwer Academic Publishers, pp. 355–368.

Neumaier, A., Schneider, T., 2001. Estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Trans Math Software 27, 27–57.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M., 2004. Identifying true brain interaction from EEG data using the imaginary part of coherency. Clin Neurophysiol 115, 2292–2307.

Nolte, G., Dassios, G., 2005. Analytic expansion of the EEG lead field for realistic volume conductors. Phys Med Biol 50, 3807–3823.

Nolte, G., Marzetti, L., Valdes Sosa, P., 2009. Minimum Overlap Component Analysis (MOCA) of EEG/MEG data for more than two sources. J Neurosci Methods 183, 72–76.

Nolte, G., Meinecke, F. C., Ziehe, A., Müller, K.-R., 2006. Identifying interactions in mixed and noisy complex systems. Phys Rev E 73, 051913.

Nolte, G., Müller, K.-R., 2010. Localizing and estimating causal relations of interacting brain rhythms. Front Hum Neurosci 4, 209.

Nolte, G., Ziehe, A., Krämer, N., Popescu, F., Müller, K.-R., 2010. Comparison of Granger causality and phase slope index. In: Guyon, I., Janzing, D., Schölkopf, B. (Eds.), Causality: Objectives and Assessment. Vol. 6 of JMLR W&CP. pp. 267–276.

Nolte, G., Ziehe, A., Nikulin, V. V., Schlögl, A., Krämer, N., Brismar, T., Müller, K. R., 2008. Robustly estimating the flow direction of information in complex physical systems. Phys Rev Lett 100, 234101.

Nunez, P., Srinivasan, R., 2006. Electric fields of the brain: the neurophysics of EEG. Oxford University Press.

Nunez, P. L., Silberstein, R. B., Shi, Z., Carpenter, M. R., Srinivasan, R., Tucker, D. M., Doran, S. M., Cadusch, P. J., Wijesinghe, R. S., 1999. EEG coherency II: experimental comparisons of multiple measures. Clin Neurophysiol 110, 469–486.

Nunez, P. L., Srinivasan, R., Westdorp, A. F., Wijesinghe, R. S., Tucker, D. M., Silberstein, R. B., Cadusch, P. J., 1997. EEG coherency. I: Statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. Electroencephalogr Clin Neurophysiol 103, 499–515.

Ogawa, S., Lee, T. M., Nayak, A. S., Glynn, P., 1990. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magn Reson Med 14, 68–78.

Oishi, N., Mima, T., Ishii, K., Bushara, K. O., Hiraoka, T., Ueki, Y., Fukuyama, H., Hallett, M., 2007. Neural correlates of regional EEG power change. NeuroImage 36, 1301–1312.

Oostenveld, R., Praamstra, P., 2001. The five percent electrode system for high-resolution EEG and ERP measurements. Clin Neurophysiol 112, 713–719.

Ou, W., Hämäläinen, M. S., Golland, P., 2009. A distributed spatio-temporal EEG/MEG inverse solver. NeuroImage 44, 932–946.

Palva, S., Palva, J. M., 2007. New vistas for alpha-frequency band oscillations. Trends Neurosci 30, 150–158.

Papadelis, C., Chen, Z., Kourtidou-Papadeli, C., Bamidis, P. D., Chouvarda, I., Bekiaris, E., Maglaveras, N., 2007. Monitoring sleepiness with on-board electrophysiological recordings for preventing sleep-deprived traffic accidents. Clin Neurophysiol 118, 1906–1922.

Parra, L., Spence, C., 2000. Convolutive blind source separation of non-stationary sources,. IEEE Trans Speech Audio Proc 8, 320–327.

Pascual-Marqui, R., 2002. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Meth Find Exp Clin Pharmacol 24, 5–12.

Pascual-Marqui, R., Michel, C., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int J Psychophysiol 18, 49–65.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos Mag 2, 559–572.

Penfield, W., Boldrey, E., 1937. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain 60, 389–443.

Penfield, W., Jasper, H., McNaughton, F., 1954. Epilepsy and the functional anatomy of the human brain. J. and A. Churchill.

Pfurtscheller, G., Lopez da Silva, F. H., 1999. Event-related desynchronization. Handbook of electroencephalography and clinical neurophysiology. Elsevier.

Polonsky, A., Zibulevsky, M., 2004. MEG/EEG source localization using spatio-temporal sparse representations. In: Puntonet, C. G., Prieto, A. (Eds.), Independent Component Analysis and Blind Signal Separation, Fifth International Conference. Lecture Notes in Computer Science. Springer, pp. 1001–1008.

Porac, C., Coren, S., 1975. Is eye dominance a part of generalized laterality? Percept Motor Skills 40, 763–769.

Pravdich-Neminsky, V. V., 1913. Ein Versuch der Registrierung der elektrischen Gehirnerscheinungen. Zbl Physiol 27, 951–960.

Ramón y Cajal, S., 1904. Textura del sistema nervioso del hombre y de los vertebrados. Moya.

Recht, B., Fazel, M., Parrilo, P. A., 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev 52, 471–501.

Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using Granger causality and fMRI. NeuroImage 25, 230–242.

Roerdink, J. B. T. M., Meijster, A., 2001. The watershed transform: Definition, algorithms and parallelization strategies. Fundamenta Informaticae 41, 187–228.

Rohlfing, T., Zahr, N. M., Sullivan, E. V., Pfefferbaum, A., 2010. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp 31, 798–819.

Roth, V., Fischer, B., 2008. The Group Lasso for Generalized Linear Models: Uniqueness of Solutions and Efficient Algorithms. In: Proceedings of the 25th International Conference on Machine Learning. pp. 848–855.

Roy, C. S., Sherrington, C. S., 1890. On the Regulation of the Blood-supply of the Brain. J Physiol (London) 11, 85–158.

Rubner, Y., Tomasi, C., Guibas, L. J., 2000. The earth mover's distance as a metric for image retrieval. Int J Comput Vision 40, 99–121.

Sabate, M., Llanos, C., Enriquez, E., Gonzalez, B., Rodriguez, M., 2011. Fast modulation of alpha activity during visual processing and motor control. Neuroscience 189, 236–249.

Sannelli, C., Vidaurre, C., Müller, K.-R., Blankertz, B., 2011. Common spatial pattern patches - an optimized filter ensemble for adaptive brain-computer interfaces. J Neural Eng 8, 4351–4354.

Sarvas, J., 1987. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Phys Med Biol 32, 11–22.

Scherg, M., Ebersole, J., 1993. Models of brain sources. Brain Topogr 5, 419–423.

Scherg, M., von Cramon, D., 1986. Evoked dipole source potentials of the human auditory cortex. Electroencephalogr Clin Neurophysiol 65, 344–360.

Schlögl, A., 2006. A comparison of multivariate autoregressive estimators. Signal Processing 86, 2426–2429.

Schlögl, A., Supp, G., 2006. Analyzing event-related EEG data with multivariate autoregressive parameters. Prog Brain Res 159, 135–147.

Schmidt, E. A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., Kincses, W. E., 2009. Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving. Accident Anal Prev 41, 1087–1093.

Schmidt, R. O., 1986. Multiple emitter location and signal parameter estimation. IEEE Trans Antenn Propag 43, 276–280.

Schoffelen, J. M., Gross, J., 2009. Source connectivity analysis with MEG and EEG. Hum Brain Mapp 30, 1857–1865.

Schreuder, M., Blankertz, B., Tangermann, M., 2010. A new auditory multi-class brain-computer interface paradigm: Spatial hearing as an informative cue. PLoS ONE 5, e9813.

Schubert, R., Haufe, S., Blankenburg, F., Villringer, A., Curio, G., 2009. Now you'll feel it - now you won't: EEG rhythms predict the effectiveness of perceptual masking. J Cognitive Neurosci 21, 2407–2419.

Schwarz, G., 1978. Estimating the dimension of a model. Ann Stat 6, 461–464.

Sekihara, K., Sahani, M., Nagarajan, S., 2005. Localization bias and spatial resolution of adaptive and non-adaptive spatial filters for MEG source reconstruction. NeuroImage 25, 1056–1067.

Seth, A. K., 2010. A MATLAB toolbox for Granger causal connectivity analysis. J Neurosci Methods 186, 262–273.

Sharbrough, F., Chatrian, G. E., Lesser, R. P., Lüders, H., Nuwer, M., Picton, T. W., 1991. American Electroencephalographic Society guidelines for standard electrode position nomenclature. J Clin Neurophysiol 8, 200–202.

Silberstein, R. B., 2006. Dynamic sculpting of brain functional connectivity and mental rotation aptitude. Prog Brain Res 159, 63–76.

Simon, M., Schmidt, E. A., Kincses, W. E., Fritzsche, M., Bruns, A., Aufmuth, C., Bogdan, M., Rosenstiel, W., Schrauf, M., 2011. EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions. Clin Neurophysiol 122, 1168–1178.

Spehlmann, R., 1965. The averaged electrical responses to diffuse and to patterned light in the human. Electroencephalogr Clin Neurophysiol 19, 560–569.

Sporns, O., Tononi, G., Kötter, R., 2005. The human connectome: A structural description of the human brain. PLoS Comput Biol 1, e42.

Srinivasan, R., Nunez, P. L., Silberstein, R. B., 1998. Spatial filtering and neocortical dynamics: estimates of EEG coherence. IEEE Trans Biomed Eng 45, 814–826.

Srinivasan, R., Winter, W. R., Ding, J., Nunez, P. L., 2007. EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics. J Neurosci Methods 166, 41–52.

Stahlhut, C., Mørup, M., Winther, O., Hansen, L., 2010. Simultaneous EEG source and forward model reconstruction (SOFOMORE) using a hierarchical Bayesian approach. J Signal Process Sys (in press).

Sturm, J., 1999. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optim Method Softw 11, 625–653.

Sun, X., Janzing, D., Schölkopf, B., 2008. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. Neurocomputing 71, 1248–1256.

Supp, G. G., Schlögl, A., Trujillo-Barreto, N., Müller, M. M., Gruber, T., 2007. Directed cortical information flow during human object recognition: analyzing induced EEG gamma-band responses in brain's source space. PLoS ONE 2, e684.

Sutton, S., Braren, M., Zubin, J., John, E. R., 1965. Evoked-potential correlates of stimulus uncertainty. Science 150, 1187–1188.

Sutton, S., Tueting, P., Zubin, J., John, E. R., 1967. Information delivery and the sensory evoked potential. Science 155, 1436–1439.

Talairach, J., Tournoux, P., 1988. Co-Planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System : An Approach to Cerebral Imaging. Thieme Medical Publishers.

Tangermann, M., Winkler, I., Haufe, S., Blankertz, B., 2009. Classification of artifactual ICA components. Int J Bioelectromagnetism 11, 110–114.

Ter-Pogossian, M. M., Phelps, M. E., Hoffman, E. J., Mullani, N. A., 1975. A positron-emission transaxial tomograph for nuclear imaging (PETT). Radiology 114, 89–98.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J Roy Stat Soc B Meth 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J Roy Stat Soc B Meth 67, 91–108.

Tichavský, P., Koldovský, Z., 2004. Optimal pairing of signal components separated by blind techniques. IEEE Signal Proc Let 11, 119–122.

Tikhonov, A. N., Arsenin, V. Y., 1977. Solutions of Ill-Posed Problems. V. H. Winston & Sons.

Tomioka, R., Sugiyama, M., 2009. Dual augmented lagrangian method for efficient sparse reconstruction. IEEE Signal Proc Let 16, 1067–1070.

Trujillo-Barreto, N. J., Aubert-Vazquez, E., Penny, W. D., 2008. Bayesian M/EEG source reconstruction with spatio-temporal priors. NeuroImage 39, 318–335.

Uutela, K., Hämäläinen, M., Somersalo, E., 1999. Visualization of magnetoencephalographic data using minimum current estimates. NeuroImage 10, 173–180.

Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J., Friston, K., 2011. Effective connectivity: Influence, causality and biophysical modeling. NeuroImage (in press).

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., Canales-Rodríguez, E., 2005. Estimating brain functional connectivity with sparse multivariate autoregression. Philos Trans Roy Soc B 360, 969–981.

Valdés-Sosa, P. A., Vega-Hernández, M., Sánchez-Bornot, J. M., Martínez-Montes, E., Bobes, M. A., 2009. EEG source imaging with spatio-temporal tomographic nonnegative independent component analysis. Hum Brain Mapp 30, 1898–1910.

van Gerven, M., Hesse, C., Jensen, O., Heskes, T., 2009. Interpreting single trial data using groupwise regularisation. NeuroImage 46, 665–676.

Van Veen, B. D., Buckley, K. M., 1988. Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag 5, 4–24.

Van Veen, B. D., van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans Biomed Eng 44, 867–880.

Vega-Hernández, M., Martínez-Montes, E., Sánchez-Bornot, J., Lage-Castellanos, A., Valdés-Sosa, P. A., 2008. Penalized least squares methods for solving the EEG inverse problem. Stat Sinica 18, 1535–1551.

Vidaurre, C., Sander, T. H., Schlögl, A., 2011. BioSig: The free and open source software library for biomedical signal processing. Comput Intell Neurosci 2011, 935364.

von Euler, U. S., 1946. A specific sympathomimetic ergone in adrenergic nerve fibres (sympathin) and its relations to adrenaline and nor-adrenaline. Acta Physiol Scand 12, 73–97.

Winkler, I., Haufe, S., Tangermann, M., 2011. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. Behav Brain Funct 7, 30.

Wipf, D., Nagarajan, S., 2009. A unified Bayesian framework for MEG/EEG source imaging. NeuroImage 44, 947–966.

Wipf, D. P., Owen, J. P., Attias, H. T., Sekihara, K., Nagarajan, S. S., 2010. Robust Bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using MEG. NeuroImage 49, 641–655.

Wipf, D. P., Rao, B. D., 2007. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. IEEE Trans Signal Proces 55, 3704–3716.

Wolters, C., de Munck, J. C., 2007. Volume conduction. Scholarpedia 2, 1738.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J Roy Stat Soc B Meth 68, 49–67.

Ziehe, A., Laskov, P., Nolte, G., Müller, K.-R., 2004. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. JMLR 5, 777–800.

Ziehe, A., Müller, K.-R., 1998. TDSEP–an efficient algorithm for blind separation using time structure. In: Niklasson, L. F. (Ed.), Proc. Int. Conf. on Artificial Neural Networks (ICANN '98). Springer, pp. 675–680.

Ziehe, A., Müller, K.-R., Nolte, G., Mackert, B.-M., Curio, G., 2000. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. IEEE Trans Biomed Eng 47, 75–87.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J Roy Stat Soc B Meth 67, 301–320.

Zwoliński, P., Roszkowski, M., Żygierewicz, J., Haufe, S., Nolte, G., Durka, P. J., 2010. Open database of epileptic EEG with MRI and postoperational assessment of foci—a real world verification for the EEG inverse solutions. Neuroinformatics 8, 285–299.