On the extraction, classification and causal analysis of EEG signal components

vorgelegt von
Diplom-Informatikerin Irene Winkler
geboren in Berlin

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissensschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender	Prof. Dr. Benjamin Blankertz
Gutachter	Prof. Dr. Klaus-Robert Müller
Gutachter	Prof. Dr. Aapo Hyvärinen
Gutachter	Dr. Guido Nolte

Tag der wissenschaftliche Aussprache: 10. Dezember 2015

Berlin, 2016

Abstract

Due to its high temporal resolution and relatively low costs, electroencephalography (EEG) is a widespread tool in neuroscientific research and clinical diagnosis. However, EEG signals suffer from low signal-to-noise ratio and are contaminated by artifacts, which are undesired signals that do not originate from the brain. Therefore, sophisticated data analysis methods are needed to extract information from EEG data.

This cumulative thesis contributes to the development of multivariate methods for the analysis of EEG data in several ways. First, it addresses necessary pre-processing steps for EEG signal analysis. An open-source toolbox that automates the time-consuming process of manually identifying artifactual signal components by EEG practitioners is developed and validated on several data sets. Second, the extraction of oscillatory signal components which explain behavioral variables is addressed. As causal information flow in time series data is often operationalized by the statistical concept of 'Granger causality', a method which directly optimizes this quantity is developed and compared to state-of-the-art methods on simulated and real EEG data. Third, problems of spurious Granger causality in the presence of measurement noise are addressed in a theoretical contribution. Drawing on results from time series analysis, this thesis provides more theoretical guarantees for time-reversed Granger causality, a recently proposed approach which is more robust with respect to noise.

Zusammenfassung

Die Elektroenzephalografie (EEG) ist eine Methode zur nicht-invasive Messung der elektrischen Aktivität des Gehirns, welche in der medizinischen Diagnostik und der neurowissenschaftlichen Forschung angewandt wird. Die Auswertung der EEG Signale wird jedoch durch deren niedriges Signal-Rausch-Verhältnis und deren Kontaminierung durch Artefakte erschwert. Daher werden fortgeschrittene datenanalytische Methoden benötigt.

Diese kumulative Doktorarbeit besteht aus Beiträgen, die sich mit der Entwicklung multivariater Methoden für die Analyse von EEG Signalen beschäftigen. Zuerst geht es um notwendige Vorverarbeitungsschritte. Es wird eine open-source Toolbox zur automatischen Klassifikation von Artefaktkomponenten des EEGs entworfen, implementiert, und auf mehreren großen Datensätzen getestet. Danach beschäftigt sich diese Arbeit mit der Extrahierung von oszillatorischen EEG Komponenten, die Verhaltensvariablen erklären. In der Zeitreihenanalyse wird ein kausaler Informationsfluss häufig durch das statistische Konzept der 'Granger-Kausalität' operationalisiert. In dieser Arbeit wird daher eine Methode entwickelt, welche direkt die Granger-Kausaliät optimiert. Die vorgeschlagene Methode wird in umfangreichen Simulationen und auf mehreren EEG Datensätzen mit state-of-the-art Methoden verglichen. In einem letzten theoretischeren Beitrag geht es um das Problem, welches Rauschen für Granger-Kausalität darstellt. Für das vor kurzem vorgeschlagene, robustere Verfahren 'zeit-invertierte Granger-Kausalität' werden theoretische Garantien bewiesen.

Acknowledgment

First of all, I would like to express my thanks to Klaus-Robert Müller, who provided unreserved support during all my PhD years. These years have been an incredible learning experience, and I thank him for giving me the freedom to follow my own research interests, while always being encouraging and providing scientific advice at the right moment. I am very grateful for the trust he placed in me and for the very good atmosphere in his group.

I would like to thank Aapo Hyvärinen and Guido Nolte for reviewing this thesis.

The thesis would not have been possible without my coworkers and colleagues. Most importantly, I would like to thank all my co-authors –also those whose work did not make it into the thesis– for their inspiring ideas, support and our fruitful cooperation, notably Michael Tangermann, Stefan Haufe, Sven Dähne, Daniel Bartz, Danny Panknin, Anne Porbadnigk, Carsten Allefeld, Stefan Debener, Stephanie Brandl, Franziska Horn, Eric Waldburger, Laura Frølich, Wojciech Samek, Sofie Hansen, Nico Görnitz, Shinichi Nakajima, Nihar Jangle, Mamta Mehra, Vaibhav Sharma and Sarah Favrichon. I am also grateful for all the administrative and technical work by Imke Weitkamp and Dominik Kühne, and for the reliable teaching assistance from Ivana Balažević. Special thanks goes to Andrea Gerdes, not only for all her administrative support but also for lifting my spirits on several occasions.

I would like to thank all my colleagues from the BBCI and Machine Learning group for the private stories, advice and laughs we shared. I like to think that it was this combined (unconscious) effort of every person in the group which made my PhD years enjoyable. I would like to especially mention Daniel Bartz, Sebastian Bach and Nico Görnitz, who deserve special thanks for having been so important for my emotional well-being. Finally, I am indebted to my friends, family and Andrew Dowding, for always supporting me.

Contents

1 Introduction

1.1 Scope of this thesis

According to the state of current neuroscientific research, human information processing is based on the communication between individual brain cells (Bear et al., 2015). This communication is mediated by electrical and chemical signals, the observation of which is very difficult. Electroencephalography (EEG) is a method that makes it possible to non-invasively measure broad activity patterns of the cerebral cortex by recording voltage fluctuations from electrodes on the scalp surface. In humans, EEG signals were first recorded by Hans Berger in the 1920s (Berger, 1929). Since then, EEG has become a standard tool for cognitive neuroscientific research and for the diagnosis of certain neurological conditions, most importantly epilepsy.

The widespread use of EEG in neuroscientific research is due to the following key advantages: it is non-invasive, can be obtained with high temporal resolution and the hardware costs are low compared to most other neuroimaging techniques. However, EEG signals suffer from a low signal-to-noise ratio and are spatially smeared, because the activity measured at a given electrode is a mixture of contributions from several neuronal sources (Baillet et al., 2001; Parra et al., 2005; Nunez and Srinivasan, 2006). EEG signals are also contaminated by artifacts, which are undesired signals that do not originate from the brain. Such artifacts are caused by eye movements or muscle activity, or by external technical sources (Iwasaki et al., 2005; Goncharova et al., 2003; Urigüen and Garcia-Zapirain, 2015).

Therefore, advanced multivariate methods are needed for the data analysis of EEG signals. This cumulative thesis contains methodological contributions, which address several issues: the removal of artifacts from EEG signals, the extraction of neuronal components of interest, and problems that measurement noise poses for causal inference.

Several publications included in this thesis are concerned with the removal of artifacts from EEG signals. One of the most common approaches for artifact removal is the transformation of EEG signals into a space of independent source components (ICs) using Independent Component Analysis (ICA) (Makeig et al., 1996; Vigário, 1997; Jung et al., 2000; Vigário et al., 2000; Ziehe et al., 2000; Hyvärinen and Oja, 2000; Vigário and Oja, 2008; Urigüen and Garcia-Zapirain, 2015). Ideally ICA separates artifactual and neuronal activity into distinct ICs, so that artifactual ICs

can be identified and a cleaner EEG can be constructed without them. However, the identification of artifactual components is a non-trivial task which requires expert knowledge, is often done manually and is therefore time-consuming. In this thesis, we develop an open-source toolbox which automatizes this process, and we evaluate the effectiveness of ICA-based artifact removal in several scenarios.

While ICA is a state-of-the art method for the extraction of neuronal phenomena of interest, a number of alternative methods turned out to be very useful in specific contexts (e.g. (Blankertz et al., 2008b; Gómez-Herrero et al., 2008; Haufe et al., 2010; Hyvärinen et al., 2010a; Nikulin et al., 2011; Dmochowski et al., 2012; Dähne et al., 2014a; Dähne et al., 2014b; Blythe et al., 2014; Fazli et al., 2015)). Many of theses methods focus on the extraction of neuronal oscillations, which have been linked to a wide range of brain functions and whose extraction relies on interesting algorithmic solutions. As the causal effects of oscillatory activity on behavior are a field of intense research (Buzsáki and Draguhn, 2004; Thut and Miniussi, 2009), this thesis explores alternatives of ICA for the extraction of oscillatory activity which displays weak evidence of causal links to behavior. We make use of the statistical concept of 'Granger causality' (Granger, 1969), which is based on the idea that the cause should precede its effect and is a standard approach for causal inference in time series analysis.

A problem for the estimation of directed interaction with Granger causality is that spurious causality can occur due to measurement noise (Nalatore et al., 2007; Nolte et al., 2008), which is especially problematic for the study of interconnected brain areas using EEG (Gómez-Herrero et al., 2008; Schoffelen and Gross, 2009; Haufe et al., 2013). In recent years, several compelling ideas for more robust causality measures have therefore been developed (Nolte et al., 2008; Vicente et al., 2011; Vinck et al., 2015). However, theoretical guarantees for these techniques are scarce. In this thesis, we provide a proof of the correctness of one recently proposed technique (Haufe et al., 2012; Haufe et al., 2013) for a relatively general class of time series models.

1.2 Outline and published work

Following this introduction, we discuss relevant background information in Chapter 2. The main contributions of this cumulative thesis are then presented in three chapters as follows.

Automatic artifact reduction for EEG signals: Chapter 3

[1] Irene Winkler, Stefan Haufe and Michael Tangermann. Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals. *Behavioral and Brain Functions*, 7:30, 2011

Summary: In this paper, we develop an automatic method for the classification of artifactual independent components. This includes a thorough feature selection procedure. We later call this method MARA (Multiple Artifact Rejection Algorithm). The classifier's performance and generalization ability is demonstrated on data of different EEG studies.

Parts of this work were presented at two TOBI (Tools for Brain-Computer Interaction) workshops in Rome, Italy in 2009 and 2010.

[2] Irene Winkler, Stephanie Brandl, Franziska Horn, Eric Waldburger, Carsten Allefeld and Michael Tangermann. Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering*, 11(3), 035013, 2014.

Summary: This paper presents a number of changes to MARA that make it more useful for practitioners. First, we make sure that the method generalizes to different electrode setups. Second, we validate the method on more data sets, notably a data set of 4473 components from a cooperation with Carsten Allefeld from the neuroimaging lab of Prof. Haynes. Third, we investigate the effect of artifact removal on the performance of a Brain-Computer-Interface (BCI) system on data from 101 users and 3 paradigms. Last but not least, we make MARA available as an open-source plug-in for EEGLAB (Delorme and Makeig, 2004), which is an interactive Matlab toolbox used by many EEG practitioners.

Parts of this work were presented at the 5$^{\text{th}}$ International Brain-Computer Interface Meeting 2013 in Asilomar, USA and the 30$^{\text{th}}$ International Congress on Clinical Neurophysiology (ICCN) 2014 in Berlin, Germany.

[3] Irene Winkler, Stefan Debener, Klaus-Robert Müller and Michael Tangermann. On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. *Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE*, 2015. In press.

Summary: Successful ICA-based artifact removal crucially depends on the quality of the obtained ICA decomposition. In turn, the ICA decomposition crucially depends on pre-processing steps, notably high-pass filtering. In this paper we use MARA to systematically evaluate the effects of high-pass filtering at different frequencies, which allows us to give practical recommendations. We also show that artifact reduction based on ICA and MARA outperforms a regression-based artifact removal method on the analyzed data set of 21 participants.

This work was presented at the 37$^{\text{th}}$ conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2015 in Milan, Italy.

[4] Irene Winkler, Stefan Haufe, Klaus-Robert Müller. Removal of muscular artifacts for the analysis of brain oscillations: Comparison between ICA and SSD. *ICML Workshop on Statistics, Machine Learning and Neuroscience*, 2015.

Summary: This paper presents preliminary work, in which we explore an alternative for ICA when we are interested in clean oscillatory EEG activity. We compare ICA with the recently proposed spatio-spectral decomposition (SSD) method (Nikulin et al., 2011). Results indicate that SSD recovers cleaner signals than ICA on a data set of 18 subjects performing self-paced foot movements.

This work was presented at the Workshop on Statistics, Machine Learning and Neuroscience (STAMLINS) of the International Conference of Machine Learning (ICML) in 2015 in Lille, France.

Extracting brain oscillations which Granger cause experimental variables: Chapter 4

[5] Irene Winkler, Stefan Haufe, Anne Porbadnigk, Klaus-Robert Müller, and Sven Dähne. Identifying Granger causal relationships between neural power dynamics and variables of interest. *NeuroImage*, 111: 489-504, 2015.

Summary: In this paper, we are interested in Granger causal links between oscillatory brain activity and behavior. Using both real and simulated EEG data, we compare Granger causal analysis on power dynamics obtained from a) sensor directly, b) state-of the art multivariate methods (e.g. ICA) and c) a novel method that directly optimizes for Granger causality, which we call GrangerCPA (Granger Causal Power Analysis). We find that computing Granger causality on channel-wise spectral power suffers from a poor signal-to-noise ratio, while all analyzed multivariate approaches alleviate this issue. GrangerCPA may or may not yield improvements over ICA, depending on the analyzed data set.

Parts of this work were presented at the BBCI Workshop 2012 in Berlin, Germany and at the Organization of Human Brain Mapping's (OHBM) annual meeting 2014 in Hamburg, Germany.

Time reversal for Granger causality: Chapter 5

[6] Irene Winkler, Danny Panknin, Daniel Bartz, Klaus-Robert Müller and Stefan Haufe. Validity of time reversal for testing Granger causality. Submitted - available as arXiv preprint arXiv:1509.07636.

Summary: To address problems of Granger causality in the presence of measurement noise, (Haufe et al., 2013) suggested to compare causality metrics

obtained from the original time series agains those from time-reversed signals. The intuitive idea is that, if temporal order is crucial to identify cause and effect, causality results should change if the temporal order is reversed. This manuscript develops theory (based on time series analysis/autoregressive modeling) which shows that this is indeed the case. Furthermore, simulations confirm that time-reversed Granger causality testing is able to infer correct directionality with high statistical power while being relatively robust with respect to measurement noise.

Chapter 6 concludes with a summary and a discussion of directions of future work.

Additional publications not included in this thesis

The following list contains all additional publications that I have (co-)authored, but which are not included in this thesis. Items are ordered chronologically. They are all peer-reviewed conference articles.

[7] Michael Tangermann, Irene Winkler, Stefan Haufe, Benjamin Blankertz. Classification of artifactual ICA components. *International Journal of Bioelectromagnetism*, 11(2):110-114, 2009.

[8] Irene Winkler, Mark Jäger, Vojkan Mihajlović and Tsvetomira Tsoneva. Frontal EEG asymmetry based classification of emotional valence using common spatial patterns. *World Academy of Science, Engineering and Technology*, 45:373-378, 2010.

[9] Irene Winkler and Michael Tangermann. Artifact-Insensitivity of CSP in Motor Imagery BCI. *International Journal of Bioelectromagnetism*, 13(2):72-73, 2011.

[10] Sofie Therese Hansen, Irene Winkler, Lars Kai Hansen, Klaus-Robert Müller and Sven Dähne. Fusing Simultaneous EEG and fMRI Using Functional and Anatomical Information. *Workshop on Pattern Recognition in NeuroImaging (PRNI), IEEE*, pages 33-36, 2015.

[11] Irene Winkler, Mamta Mehra, Sarah Favrichon, Vaibhav Sharma and Nihar Jangle. Assessing the applicability of NDVI data for the design of index-based agricultural insurance in Bihar, India. *Annual International Geoscience and Remote Sensing Symposium (IGARSS), IEEE*, 2015. In press.

[12] Laura Frølich, Irene Winkler, Klaus-Robert Müller and Wojciech Samek. Investigating effects of different artefact types on Motor Imagery BCI. *Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE*, 2015. In press.

2 Fundamentals

In this chapter, we provide an overview of the important characteristics of EEG signals (Section 2.1), and cover machine learning methods for Blind Source Separation (Section 2.2) and causal inference (Section 2.3). As we introduce these methods, we also link the contributions in this thesis to the field.

A good introduction to machine learning methods for brain imaging can be found in the tutorial paper by (Lemm et al., 2011). For a comprehensive introduction into the field of machine learning we refer the reader to (Bishop, 2006; Hastie et al., 2009). Introductions to parametric and non-parametric statistics can be found in (Field, 2009; Casella and Berger, 2002; Manly, 2007).

2.1 Electroencephalography (EEG)

Electroencephalography (EEG) records electrical activity from electrodes placed on the scalp. EEG signals are typically measured at several, approximately equidistant electrodes which cover the whole scalp. They are placed at specific locations according to the international 10-20 system (Klem et al., 1999) as visualized in Figure 2.1. The electrodes are also referred to as channels or sensors.

A conductive gel is used between the electrode and the scalp in order to reduce impedance, and the recorded signals go through amplifiers. Resulting EEG signals are about 10 to 100 μV in amplitude. EEG signals are always the difference between the voltages at two electrodes. They are therefore defined with respect to a reference. Typical choices for reference electrodes are the ear-lobes, the nose, the mastoids (the bone behind the ear) or the 'average reference', that is the average over all electrodes.

Typical sampling rates are between one hundred and several thousand Hz. This is very high compared to methods which measure blood flow like functional magnetic resonance imaging (fMRI) and near infrared spectroscopy (NIRS). However, EEG suffers from low spatial resolution. Localizing the source of measured activity is a challenging problem (Baillet et al., 2001; Haufe et al., 2008; Shahbazi et al., 2015).

In the following, we will briefly cover the underlying neurophysiology of EEG signal generation. We then review pre-processing methods for the removal of artifacts from EEG signals. Furthermore, we describe two common neurophysiological signatures that can be extracted from EEG signals, Event-Related Potentials (ERP) and

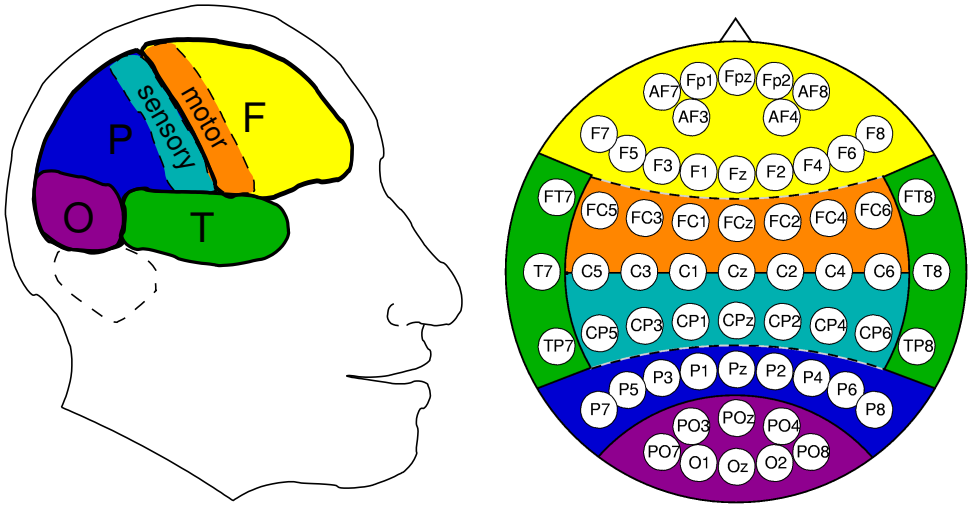


Figure 2.1: (Left) Schematic representation of the major sectors of the cerebral cortex: F is frontal, P is parietal, O is occipital and T temporal. (Right) example electrode montage and electrode labels according to the international 10-20-system. Figures are taken from the lecture 'Brain Computer Interfacing' of Benjamin Blankertz, with permission.

rhythmic activity. Finally, we give a short introduction on how these signatures can be used to control Brain-Computer-Interface (BCI) systems.

2.1.1 Neurons and their electrical activity

The two principal categories of cells in the nervous system are nerve cells, or neurons, and glia cells (Rosenzweig et al., 2002; Bear et al., 2015). Each human being has about 100 billion to 150 billion neurons and about nine times that number of glia cells. Neurons are electrically excitable cells and are recognized to be the basic information processing unit of the nervous system. Glia cells are thought to mainly provide support, protection and nutrients for neurons. In the following, we will give a short overview on a single neuron's structure and function.

Neurons have diverse forms that vary between different parts of the brain and the function they perform. However, most of them have three distinct structural parts illustrated in Figure 2.2: The cell body, dendrites and one axon. Information flows from the dendrites through the cell body to the axon. The majority of neurons have exactly one axon. An axon might be only a few micrometers long, but can reach more than a meter in length, for example in spinal and motor neurons. Towards its end, an axon typically divides into numerous branches. It forms connections to

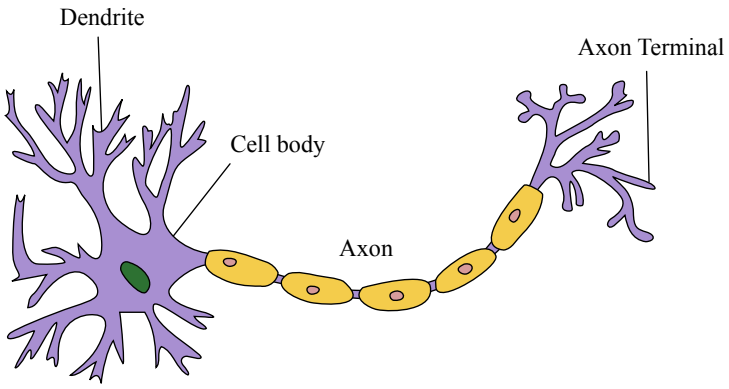


Figure 2.2: The structure of a typical neuron. Taken from (Wikipedia, 2015).

other neurons via specialized structures called synapses.

The information that travels along a neuron is encoded in a flow of electrical current. The basis of this electrical activity is the movement of ions through the neuron's cell membrane. A typical undisturbed human neuron has a membrane potential of -70 mV, that is, the inside of the cell is negatively charged relative to the outside. Neurons communicate with each other via so-called action potentials: brief reversals of the membrane potential that travel rapidly along the axon. When the cell depolarizes to a certain threshold potential of about -50 mV, an action potential is generated. The cell membrane depolarizes, then becomes positive reaching a value of +40 mV and rapidly returns to its resting potential. This process takes only about 1 millisecond. It is caused by voltage-gated ion channels in the cell's membrane which allow a brief, large influx of sodium ions followed by a brief, large efflux of potassium ions.

When an action potential reaches the synapse, chemicals are released which trigger so-called postsynaptic potentials at the next neuron. As the next neuron also has numerous synaptic inputs but only one axon, it integrates the information received. Whether or not the neuron will generate an action potential depends on the strength of the excitatory and inhibitory postsynaptic potentials. Note that the size of the action potential is independent from the amount of current that produced it. It either occurs or does not occur, a characteristic referred to as the all-or-none property. Thus, the information of, for example which color was perceived, cannot be represented by just one action potential. It is rather encoded in the complex interaction of several action potentials.

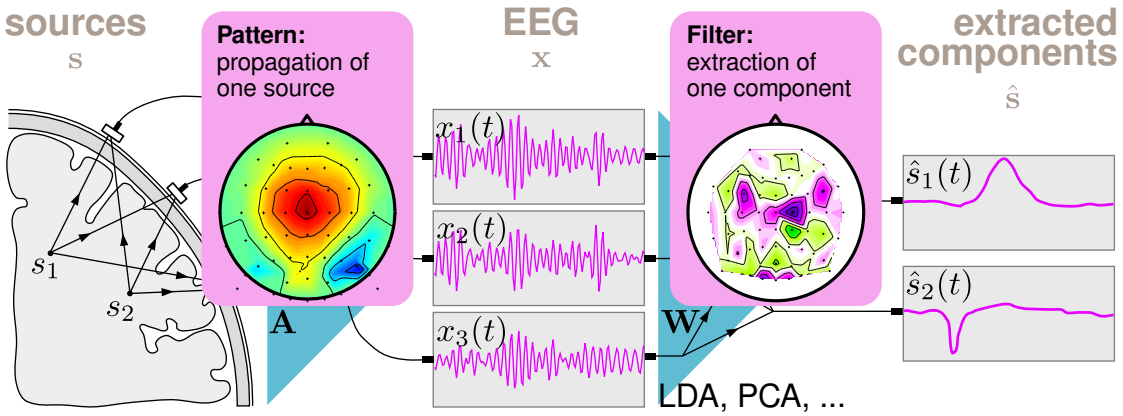


Figure 2.3: Schematic illustration of the generative model of EEG. Electrodes on the scalp record a linear mixture of the source activity. Each source is associated with a spatial pattern, which describes the influence of the source on the recorded signals. Machine learning methods can be used to learn a spatial filter, which gives a weighting of EEG electrodes to recover the sources. In contrast to the spatial pattern, the spatial filter is in general not directly interpretable in terms of the spatial origin of the extracted source. The figure is taken from the lecture 'Brain Computer Interfacing' of Benjamin Blankertz, with permission

2.1.2 EEG signal generation and its generative model

The transmission of electric or magnetic fields from an electric current source through biological tissue towards measurement sensors is termed 'volume conduction'. Due to volume conduction, neuronal signals generated inside the brain are spatially smeared while propagating to the sensors. Because the potentials generated by a single neuron are too small to be detected on the scalp, EEG signals reflect the superposed activity of many neurons with similar spatial location (Baillet et al., 2001; Nunez and Srinivasan, 2006), as illustrated in Figure 2.3. Structured arrangements of cortical pyramidal neurons are believed to be the main EEG signal generator.

Because the superposition of neuronal sources is instantaneous and linear in the sources (Baillet et al., 2001; Parra et al., 2005; Nunez and Srinivasan, 2006), the electrophysics of EEG can be modeled as

$$X = A \cdot S + \eta. \tag{2.1}$$

Here $X \in \mathbb{R}^{M \times T}$ denotes the surface potentials measured at M sensors over T time points, $S \in \mathbb{R}^{K \times T}$ denotes the time-courses of K underlying neuronal sources, and the matrix $A \in \mathbb{R}^{M \times K}$ describes the influence of each source on each sensor. Each

column of A is called a spatial pattern of the respective source. It depends on the spatial location of the source and the conductivity of different brain tissues. Any contribution that is not described by A is summarized in an additive sensor noise term $\eta \in \mathbb{R}^{M \times T}$.

The data sets analyzed in this thesis contain between $M = 50$ and $M = 110$ electrodes. The number of time points T depends on the sampling rate and the length of the recording. If EEG signals are recorded at 200 Hz over one hour, this results in $T = 200 \cdot 60 \cdot 60 = 720000$ data points.

When analyzing a neuronal phenomenon of interest, it is beneficial to try to recover the unknown neuronal signals S from the sensor measurements X. To do so, various multivariate signal processing algorithms have been proposed that linearly combine channels to extract signals of interest $\hat{S} = \hat{W}X$. Here $\hat{W} \in \mathbb{R}^{K \times M}$ denotes the demixing matrix and its rows are called spatial filters. The spatial filters give a weighting of EEG channels to recover the sources. They are in general not directly interpretable in terms of the spatial origin of the extracted sources (Haufe et al., 2014b).

We will describe mathematical methods for the estimation of $\hat{W}$ in Section 2.2.

2.1.3 Artifacts

EEG measurements of brain activity are contaminated by artifacts. Typical artifacts of the EEG are caused either by the non-neuronal physiological activities of the subject or by external technical sources. Eye blinks, eye movements, muscle activity in the vicinity of the head (e.g. face muscles, jaws, tongue, neck), and the heart beat are examples for physiological artifact sources. Swaying cables in the magnetic field of the earth, line humming, power supplies or transformers can be the cause of technical artifacts.

The two most common physiological artifacts are ocular (EOG) and muscle (EMG) artifacts. Figure 2.4 shows example time courses. Electrooculographic (EOG) activity is either caused by rolling of the eyes or by eye blinks which occur about 20 times per minute (Iwasaki et al., 2005). Both result in a low-frequency activity which is most prominent over the frontal head regions, with maximal frequencies below 4 Hz. In contrast, electromyographic (EMG) activity, caused by chewing, swallowing, head or tongue movements, is usually a high-frequency activity (> 20 Hz) (Goncharova et al., 2003).

Artifacts are problematic in clinical applications and in neuropsychological research, because they can be mistaken for brain activity and distort analysis. The removal of artifacts from EEG signals is therefore an important issue for EEG signal processing. In the following, we provide a short overview of the methods used for this purpose. A more extensive review can be found in (Fatourechi et al., 2007;

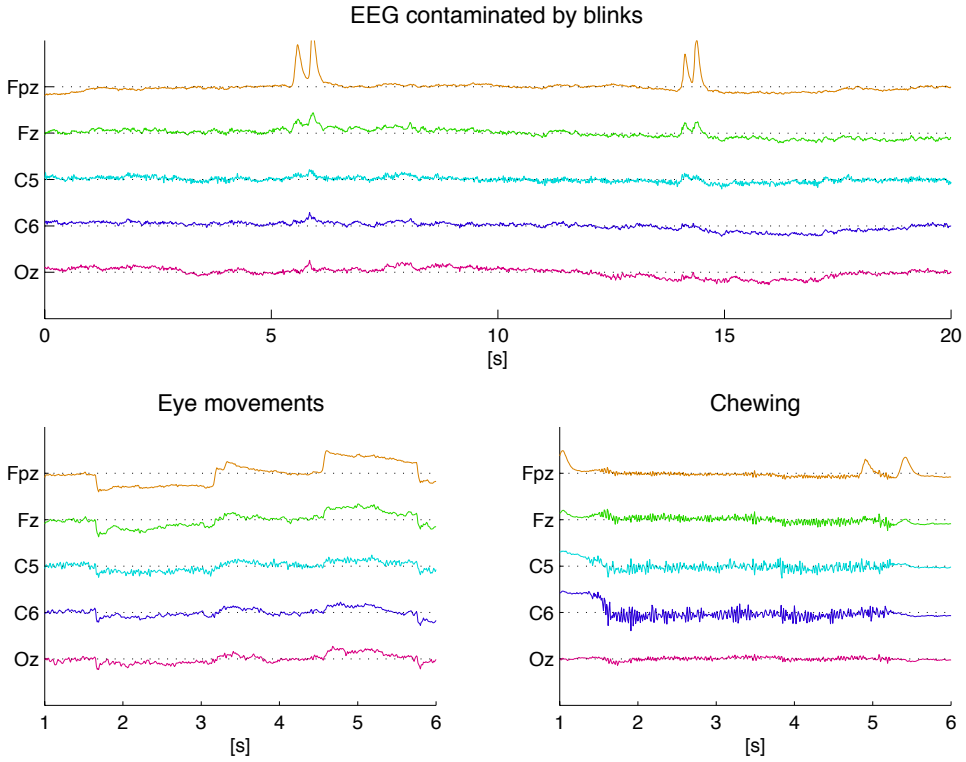


Figure 2.4: Example EEG signals from five electrodes contaminated by (top) blinks, (bottom left) a downward followed by an upward eye movement, and (bottom right) chewing.

Urigüen and Garcia-Zapirain, 2015).

One common solution to the artifact problem is to remove artifactual segments of data. However, this leads to a loss of data. Methods which remove the artifacts while preserving the underlying neuronal activity are often desirable.

Ocular activity can be partially removed by regression-based methods, which subtract a part of the activity measured at additional electrooculogram (EOG) channels from the EEG (see (Croft and Barry, 2000) for a review). Regression-based methods require the reliable recording of additional EOG channels. They are also limited by the fact that the EOG is contaminated by brain activity which is removed as well.

A widely used technique for the removal of general types of artifacts is based on techniques of Blind Source Separation (BSS), most importantly Independent Component Analysis (ICA) (Makeig et al., 1996; Vigário, 1997; Jung et al., 2000; Vigário et al., 2000). ICA linearly transforms EEG signals into independent source components (ICs) (cf. Section 2.2.2). If artifactual and neuronal activity are contained in

separate components, artifactual components can be identified and a cleaner EEG can be reconstructed without them. Even though assumptions for the application of ICA methods are only approximately met in practice, their application usually leads to a good separation of artifactual and neuronal signals (Fitzgibbon et al., 2007; Romero et al., 2008; Crespo-Garcia et al., 2008; McMenamin et al., 2010).

While some studies argued for the use of regression-based methods for the removal of eye artifacts (Wallstrom et al., 2004; Schlögl et al., 2007; Schlögl et al., 2009), cumulating evidence suggests that ICA-based artifact removal yields cleaner signals (Romero et al., 2008; Hoffmann and Falkenstein, 2008; Ghaderi et al., 2014; Winkler et al., 2015a). The state of current research on artifact removal in EEG signals has been nicely summarized in a recent review by (Urigüen and Garcia-Zapirain, 2015):

> During the last decade only a few novel methods have been proposed in the artifact removal area, in addition to classic existing approaches such as regression, ocular artifact correction, filtering or the more widely used blind source separation (BSS) techniques. Rather, the area has evolved with authors either improving on existing algorithms, combining different methods or trying to make the denoising process automatic [...] In our opinion, this seems to indicate that in fact most of the modern artifact removal methods converge in terms of results
>
> [...]
>
> Despite the fact that some contradictory studies exist, ICA-based procedures are the main accepted solution to obtain clean EEG of improved signal quality, even if they do not always completely separate artifactual from cerebral sources.

The typical process chain for ICA artifact rejection consists of the following steps:

1. A rough pre-cleaning of the data by channel rejection and trial rejection may be performed. This step is usually helpful for obtaining a good ICA decomposition. It is possible but not always necessary to run ICA twice, first to reject epochs based on the IC time courses, second to obtain a good ICA decomposition for the cleaner data.

2. As ICA decomposition is known to be sensitive to slow drifts, high pass filtering the data can improve the quality of the decomposition (Hyvärinen et al., 2001; Pignat et al., 2013).

3. The ICA decomposition is computed. For discussions on which ICA algorithm is best suited for artifact removal, we refer the interested reader to (Meinecke et al., 2002; Kierkels et al., 2006; Fitzgibbon et al., 2007; Romero et al., 2008; Delorme et al., 2012).

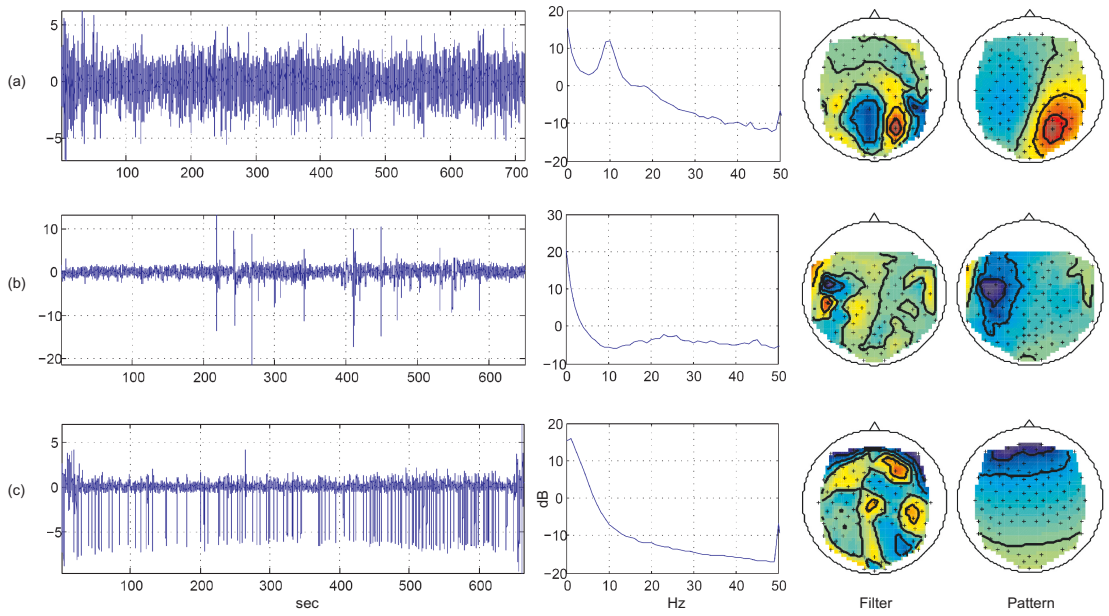


Figure 2.5: Three example independent source components. Time series (first column), spectrum (second column), filter (third column) and pattern (fourth column) of three components. The first row (a) shows a neuronal component. The second row (b) shows a rare muscle artifact component with an increased spectrum in higher frequencies. The third row (c) shows an eye artifact component that appears regularly, has an increases spectrum in lower frequencies and a typical front-back distribution in the pattern. Taken from (Winkler et al., 2011).

It is possible to perform a dimensionality reduction prior to ICA computation. This is typically done with Principal Component Analysis (PCA) (cf. Section 2.2.1). Such a step can reduce the noise level and avoid an unnatural splitting of sources. It also makes ICA computation faster and reduces the number of components that have to be visually inspected.

4. Artifactual ICA components are identified. This identification of artifactual component is a non trivial task and requires time and expert knowledge. To do so, EEG practitioners typically inspect the time course, the spectrum, and the scalp pattern of all independent component (Chaumon et al., 2015). What makes the identification task difficult is that some components are not clearly interpretable and may contain both, artifactual and neuronal activity.

   Prototypical examples of a neuronal, a muscle, and an eye blink component are displayed in Figure 2.5. Neuronal components (Figure 2.5a) typically display

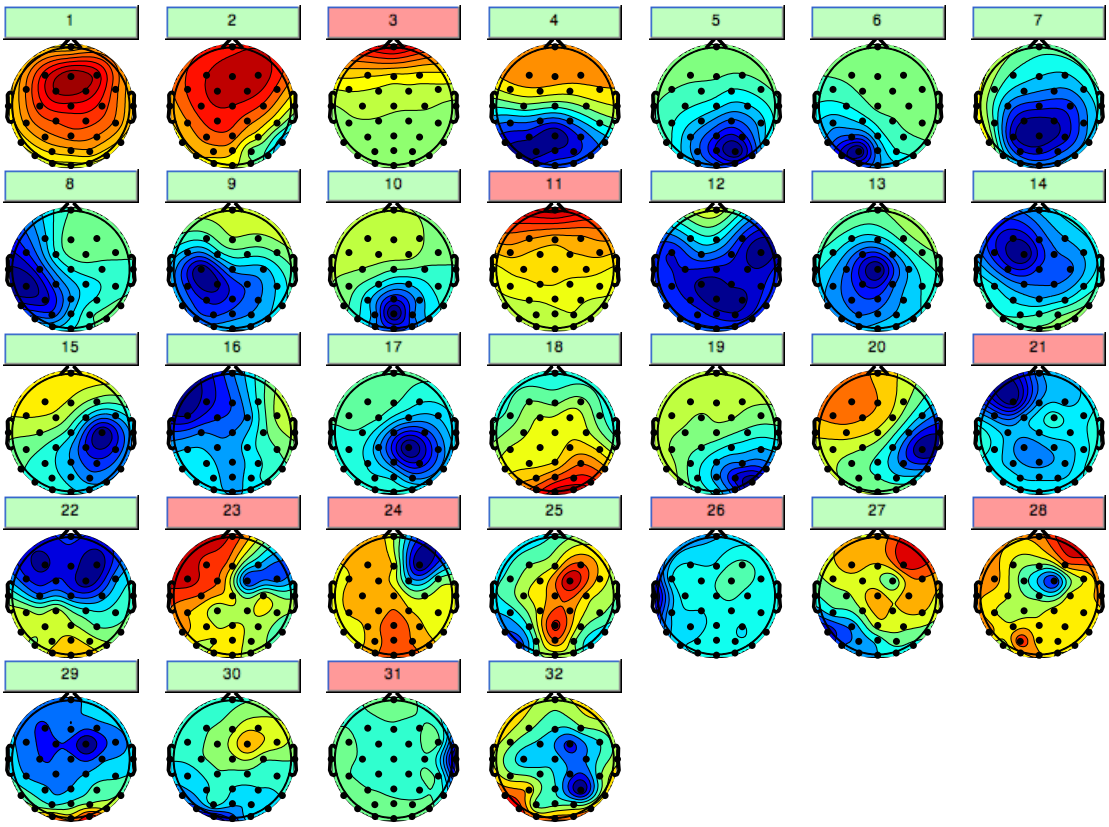


Figure 2.6: Example ICA decomposition of EEGLAB sample data. The plot shows the spatial pattern of each ICA component and the labels computed by MARA (red: artifact/reject, green: neuronal/accept).

a peak in the spectrum around 10 Hz (the so-called alpha peak) and a dipolar scalp map. Muscular components (Figure 2.5b) are often characterized by spatially localized activity and comparatively high power above 20 Hz. Eye blink artifacts (Figure 2.5c) are characterized by a strong frontal activation in the scalp map and a steep power spectrum without an alpha peak.

5. After artifactual ICs have been identified, the EEG can be reconstructed without them.

Contributions of this thesis. In Chapter 3 of this thesis, we address point 4 and point 2 of the ICA-based artifact processing pipeline outlined above.

Most of our work on artifact rejection addresses point 4. We develop and test an open-source EEGLAB-plugin MARA (Multiple Artifact Rejection Algorithm)

which avoids the time-consuming hand-rating process of independent components by automatically classifying them into artifactual and non-artifactual components (Winkler et al., 2011; Winkler et al., 2014). It is based on a linear classifier which is easy to understand (cf. (Müller et al., 2003)), and has been used in our group (Höhne and Tangermann, 2014; Hwang et al., 2015) and by others (Gomez Rojas, 2012; Nagya et al., 2014; Ho et al., 2015; Alday, 2015; Tóth, 2015; Wang et al., 2015). While other available EEGLAB-plugins mostly focus on eye artifacts (Viola et al., 2009; Mognon et al., 2011; Bigdely-Shamlo et al., 2013), MARA solves the binary classification problem 'artifact vs. not an artifact'. It is therefore also able to handle muscle artifacts (see Figure 2.6). A multi-class method which can also classify muscle artifacts has only recently been made available as an EEGLAB-plugin (Frølich et al., 2015a).

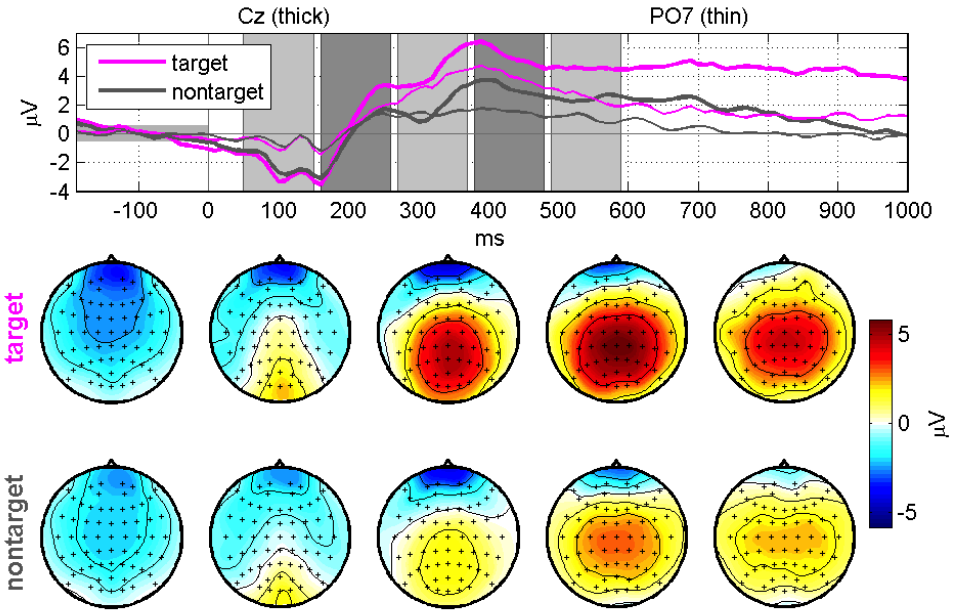
We further used MARA to address point 2 of the artifact pipeline. Without prior high-pass filtering, ICA often produces visibly poor separation of artifactual and non-artifactual activity. Recent research also indicates that high-pass filtering improves reliability (Groppe et al., 2009) and measures of independence and dipolarity (Zakeri et al., 2014) of the estimated independent components. The question then arises which cut-off frequency to use in practice. This is an empirical question, that probably depends on the type of data one wants to analyze. We address this question in Section 3.3 (Winkler et al., 2015a) in one Event-Related Potential (ERP) study.

2.1.4 Event Related Potentials (ERPs)

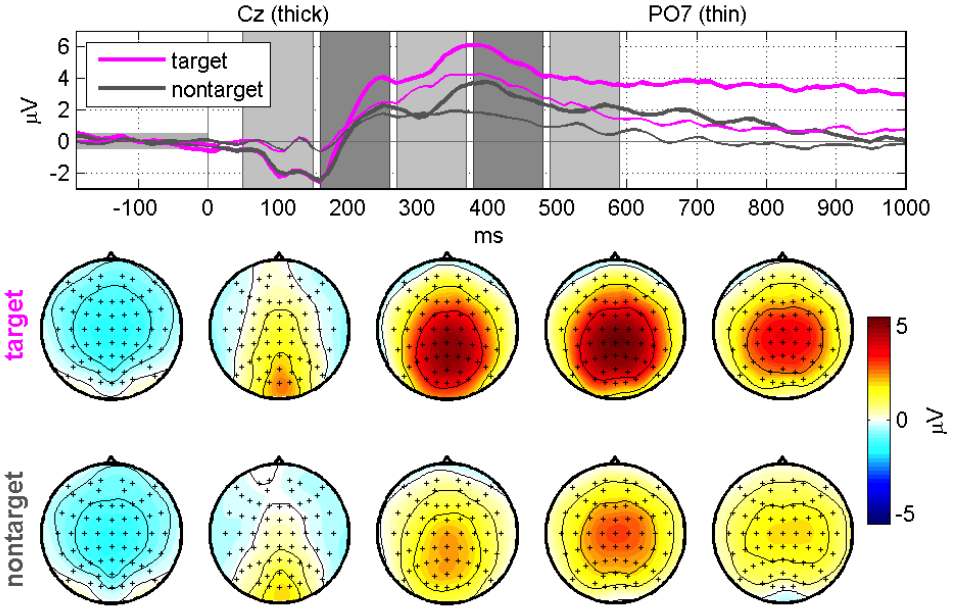
An Event-Related Potential (ERP) is computed by averaging EEG activity after the presentation of visual, somatosensory or auditory stimuli. Currently, ERP is one of the most widely used methods in cognitive neuroscience research (Fabiani et al., 2000) and is extensively used for brain-computer-interfacing (Farwell and Donchin, 1988; Treder and Blankertz, 2010; Höhne et al., 2011; Blankertz et al., 2011) and mental state monitoring (Blankertz et al., 2010b; Müller et al., 2008).

As an example, Figure 2.7 depicts event-related potentials occurring during visual stimulation administrated to closed eye lids (Hwang et al., 2015). Twelve healthy participants were asked to attend one of three different visual stimuli which were flashed sequentially. The stimuli that participants attend to are called the target stimulus, while the unattended stimuli are called the non-target stimuli. We can see a characteristic P300 ERP component, which is a peak that occurs approximately 250 to 500 ms post stimulus. It is thought to reflect an attention-dependent cognitive component.

When no artifact removal is performed, the ERP is contaminated by stimulus-specific eye movements as evidenced by activity in the frontal electrodes (Fig-



(a) Without artifact removal



(b) After ICA-based artifact removal with MARA

Figure 2.7: Grand-average ERPs for target and non-target stimuli of 12 subjects of a study in which visual stimuli were administered on closed eyelids, with and without prior ICA-based artifact removal with MARA. Plots show the time courses at two electrodes and average activation scalp maps in the five marked intervals. Figures are taken from (Hwang et al., 2015), with permission.

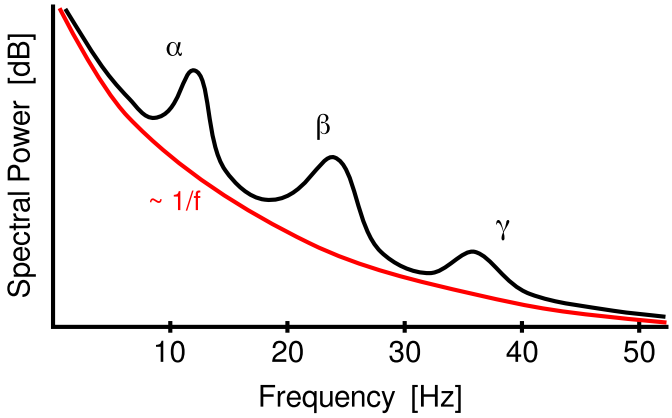


Figure 2.8: Idealized spectrum of EEG activity. Taken from the lecture 'Brain Computer Interfacing' of Benjamin Blankertz, with permission.

ure 2.7a). Eye activity is removed by ICA-based artifact removal using MARA (Figure 2.7b). In this thesis, we contribute further evidence that ICA-based artifact removal using MARA may be a beneficial for ERP computation in Section 3.3.

2.1.5 Oscillatory Activity

EEG signals demonstrate oscillations in characteristic frequency ranges. This rhythmic neuronal activity is a fundamental property of neuronal networks. Brain rhythms correlate with particular states of mind (such as level of attentiveness or sleeping), vary considerably over time, and occur in certain characteristic forms during seizures and coma. Reviews on EEG rhythms can be found in (Buzsáki and Draguhn, 2004; Klimesch et al., 2007; Uhlhaas et al., 2008; Wang, 2010).

Oscillatory EEG activity is commonly divided in 5 frequency bands, the delta (0-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma-band (30-80 Hz). The EEG power spectrum is inversely proportional to frequency, and may exhibit spectral peaks in the characteristic frequency bands (Figure 2.8). Most prominent is the alpha-peak, which is the dominant frequency in the human EEG of adults.

EEG reflects the activity of numerous neurons that are spatially aligned and discharge synchronously. The amplitude of EEG signals thus strongly depends on how synchronous the underlying neuronal activity is. Synchrony may change in response to certain tasks. This phenomenon is called Event-Related Desynchronization (ERD). ERD is computed as the relative difference in band power during the performance of a task compared to a reference period (Pfurtscheller and Aranibar, 1979). Neurophysiologically, ERD indicates that neurons no longer discharge synchronously.

The functional role of oscillations is far from being understood and an active area

of neuroscientific research. The following extract of the recent textbook by Bear, Paradiso & Conners (Bear et al., 2015) summarizes:

> Cortical rhythms are fascinating to watch in an EEG, and they parallel so many interesting human behaviors that we are compelled to ask: Why so many rhythms? More importantly, do they serve a purpose? There are no satisfactory answers yet. Ideas abound, but definitive evidence is scarce. One hypothesis for sleep-related rhythms is that they are the brain's way of disconnecting the cortex from sensory input. [...] A function for fast rhythms in the awake cortex has also been proposed. One scheme for understanding visual perception takes advantage of the fact that cortical neurons responding to the same object are synchronously active. Walter Freeman, a neurobiologist at the University of California, Berkeley, pioneered the idea that neuronal rhythms are used to coordinate activity between regions of the nervous system. [...] The evidence for this idea is indirect, far from proven, and understandably controversial. For now, the functions of rhythms in the cerebral cortex are largely a mystery. One plausible hypothesis is that most rhythms have no direct function. Instead, they may be intriguing but unimportant by-products of the tendency for brain circuits to be strongly interconnected, with various forms of excitatory feedback.

Contributions of this thesis. An intriguing way to investigate the functional role of oscillations is to induce them with brain stimulation techniques such as repetitive Transcranial Magnetic Stimulation (rTMS) and Transcranial Alternating Current Stimulation (TACS) (Thut et al., 2012; Herrmann et al., 2013).

In Chapter 4 of this thesis, we explore a fundamentally different approach that does not require direct intervention in the nervous system, by using statistical methods for the inference of cause-effect relations. We use the concept of 'Granger causality' (Granger, 1969), which operationalizes causality in time series based on the idea that the cause should precede its effect (cf. Section 2.3.3 for definition and limits). More specifically, we consider the case in which we simultaneously measure EEG band power ϕ and a behavioral target variable of interest z over time. For example, an actively researched question in the field of Brain-Computer Interfaces (BCIs) is whether (and how) oscillatory sources influence the control performance of a user during a BCI experiment (Grosse-Wentrup et al., 2011; Maeder et al., 2012). In the Granger causal setting, the target variable z would be the BCI's control performance per trial, and the goal is to identify a neuronal source whose power time course Granger causes z.

The simplest approach to testing for Granger causality is to compute the bandpower and test for Granger causality separately for each electrode. However, we

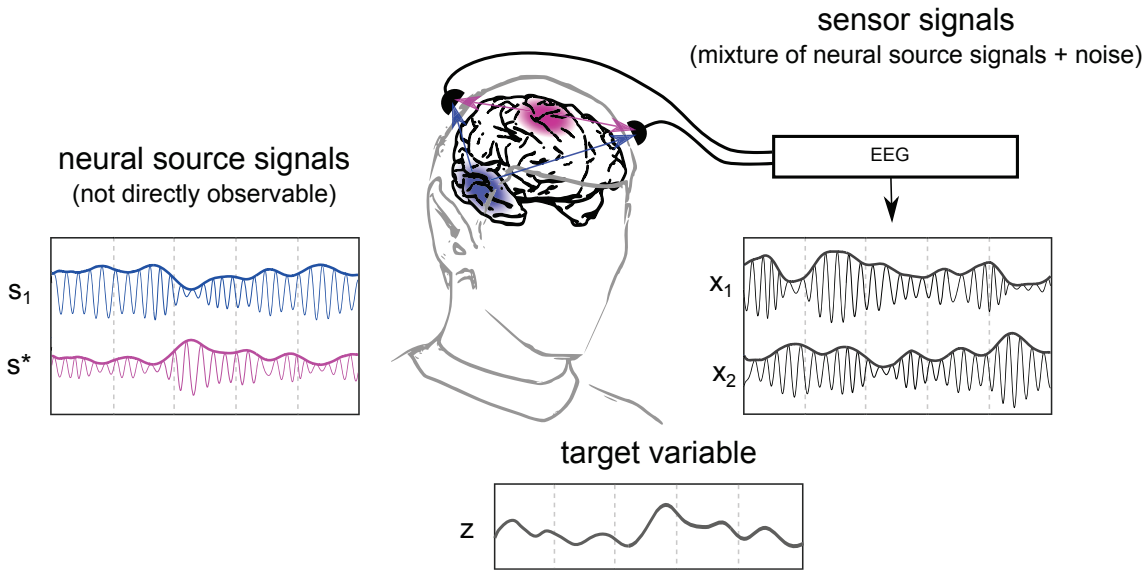


Figure 2.9: Problem setting in Chapter 4. In this simple cartoon example, there are two active brain sources (s_1 and s^*). The power dynamics of source s^* are causally related to an externally observed target variable z, while the power dynamics of s_1 are unrelated to z. EEG sensors record a linear mixture of the source activity, resulting in channel signals x_1 and x_2. The appropriate course of action is to recover the time course of s^* before the computation of band power. Information contained in the target variable z can be used to recover s^*. Taken from (Winkler et al., 2015b)

achieve a higher signal-to-noise ratio by recovering the underlying signals from scalp recordings prior to the computation of band-power (Figure 2.9). In Chapter 4, we therefore investigate which signal processing methods are best suited to extract rhythmic activity of interest from the sensors measurements according to the linear generative model of EEG given in (2.1). We develop a method, referred to as GrangerCPA (Granger Causality Power Analysis), which makes use of the assumed dependency to the target variable z, and we compare it to state-of-the art methods such as ICA in simulated and real EEG recordings.

2.1.6 Brain-Computer-Interfacing (BCI)

An interesting use-case of EEG is Brain-Computer-Interfacing. Brain-computer interfaces (BCI) are systems that allow a direct connection between brain and computer without the use of the extremities (Dornhege et al., 2007; Wolpaw and Wolpaw, 2012). BCIs are based on the single-trial classification of the ongoing EEG signal, and are developed with the goal to improve the life quality of disabled individuals

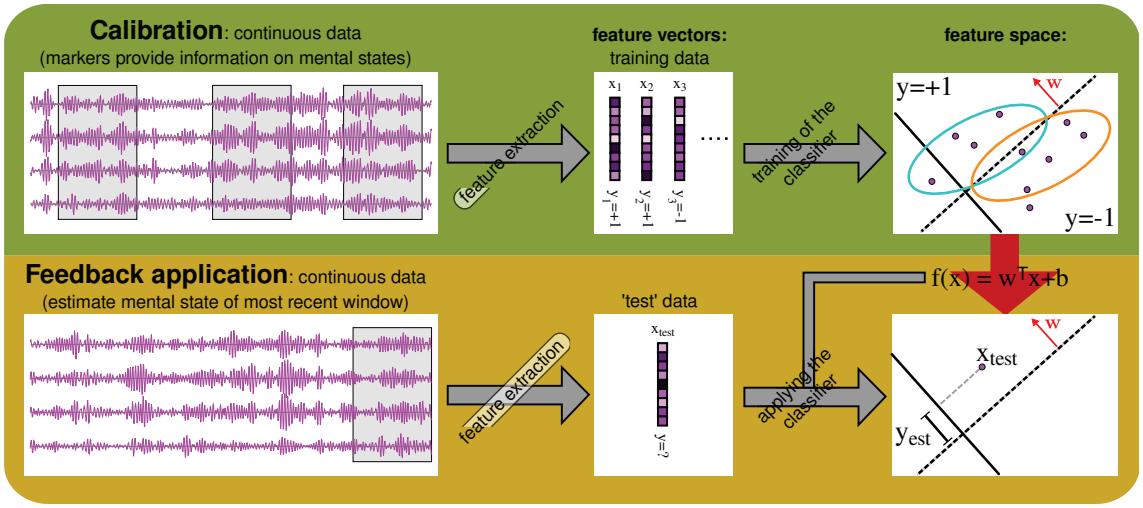


Figure 2.10: The two phases of a BCI system based on machine learning techniques. During the calibration phase the user is required to perform certain pre-defined mental tasks. Meaningful features are extracted from the EEG data and are used to train a classifier. This classifier is then used in the second phase, allowing the user to control a feedback application. The figure is taken from the lecture 'Brain Computer Interfacing' of Benjamin Blankertz, with permission.

(del R. Millán et al., 2010; Acqualagna and Blankertz, 2013; Höhne et al., 2014).

Research on non-medical application scenarios investigates whether it is possible to continuously monitor fatigue, attention, emotional engagement and mental workload in operational environments or during driving (Berka et al., 2008; Müller et al., 2008; Winkler et al., 2010; Blankertz et al., 2010b; Venthur et al., 2010; Haufe et al., 2011). Furthermore, machine learning methods which were successful in the field of Brain-Computer-Interfacing have been used in other applications domains, for example to investigate the subconscious processing of speech quality (Porbadnigk et al., 2013; Porbadnigk et al., 2015; Görnitz et al., 2014) and in prosthetic device control (Hahne et al., 2012; Vidovic et al., 2014; Hwang et al., 2014; Kauppi et al., 2015).

A BCI system based on machine learning techniques typically requires two phases, as illustrated in Figure 2.10. During the calibration or training phase the user is required to perform certain pre-defined mental tasks. The collected EEG data is then used to train a classifier, which enables the user to communicate via the system in the feedback phase (Blankertz et al., 2002).

BCI systems can be based on different neurophysiological phenomenon. ERP-based systems can be used for attention-based typewriting. In a typical application,

several symbols or letters are displayed on the screen. During each trial, the user is instructed to focus on one symbol or letter, while the symbols are intensified ('flashed') one after another in random order. The attended letter induces a different ERP than the non-attended letters. ERPs can thus be distinguished by machine learning methods, which are well described in the tutorial by (Blankertz et al., 2011).

BCI systems can also be based on Motor Imagery, in which imagined movements (typically left hand, right hand, foot) are detected by the system. These systems exploit spatially localized event-related desynchronization patterns that occur during Motor Imagery. To extract these patterns with an acceptable signal-to-noise ratio, the Common Spatial Patterns (CSP) method is typically employed. CSP finds a linear combination of signals from several electrodes in order to maximize the spectral power differences of the classes. The mathematical details are well described in (Blankertz et al., 2008b).

Contributions of this thesis. If a BCI system developed and tested on healthy subjects is (unconsciously) controlled by artifacts, it will be of little use in patients who may not be physically capable to produce these artifacts. Even in healthy subjects, artifacts may reduce the signal-to-noise ratio and impact negatively on the BCI's performance, especially if the stimulus during training induces other artifacts than those from online use (Frølich et al., 2015b; Brandl et al., 2015). Strategies to overcome this problem include the use of regularization methods for CSP (Blankertz et al., 2008a; Kawanabe et al., 2009; Samek et al., 2012; Samek et al., 2014) and the explicit removal of outlier trials or electrodes.

ICA-based artifact cleaning might also help to improve BCI performance. As a first proof-of-concept, Halder et al. (Halder et al., 2007) applied ICA-based artifact cleaning to data from three participants who performed motor imagery. Depending on whether artifacts were systematically co-activated with the task or not, opposite effects of artifact cleaning on BCI classification performance were demonstrated. To the best of our knowledge, only small data sets of one or two participants had been analyzed since then. In Section 3.2 we therefore analyze the effect of ICA-based artifact removal with MARA both in an ERP-based (21 participants) and a Motor Imagery based BCI-study (80 participants). We find that, while we might be more sure that artifacts are not used for BCI control, the overall performance of the analyzed BCI paradigms did not benefit from ICA-based cleaning.

2.2 Blind Source Separation (BSS)

The blind source separation (BSS) problem consists of recovering a set of source signals $S \in \mathbb{R}^{K \times T}$ from a set of mixed signals $X \in \mathbb{R}^{M \times T}$. The goal is to invert the

linear model

$$X = A \cdot S \tag{2.2}$$

with little prior knowledge on the source signals S or the mixing matrix $A \in \mathbb{R}^{M \times K}$. Here K denotes the number of source signals, M denotes the number of sensors and T denotes the number of available time points. Without loss of generality, it is assumed that the observed variables X and the hidden sources signals S have zero mean.

Note that for most BSS applications, it would be more realistic to assume an additive measurement noise term as in the generative model of EEG in (2.1). However, the BSS problem is typically formulated without the noise term, because the noise-free model is difficult enough to estimate and it performs reasonably well in many applications (Hyvärinen and Oja, 2000).

The BSS-problem is in general highly underdetermined, but useful solutions can be derived under a variety of assumptions. A demixing matrix $\hat{W} \in \mathbb{R}^{K \times M}$ is estimated such that the estimated sources

$$\hat{S} = \hat{W} \cdot X \tag{2.3}$$

best fulfill pre-defined assumptions. The rows of $\hat{W}$ are called spatial filters.

Numerous algorithms exist to estimate $\hat{W}$. We will describe Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Spatio-Spectral Decomposition (SSD) in the next sections.

2.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a wide-spread pre-processing step used for the dimensionality reduction of multivariate data. The data is projected onto a lower-dimensional subspace in a way that minimizes the information lost in terms of mean squared error. This is done by finding directions in the data which explain most of its variance. PCA is computed by solving an eigendecomposition of the covariance matrix of the data (see e.g. (Bishop, 2006)). The resulting spatial filters are orthogonal to each other.

2.2.2 Independent Component Analysis (ICA)

In the field of neuroscience, one of the most popular BSS algorithms is Independent Component Analysis (ICA; (Jutten and Herault, 1991)), which solves the Blind Source Separation problem under the assumption that the sources S are mutually independent. In contrast to PCA, ICA methods generally yield non-orthogonal filters, as illustrated in Figure 2.11.

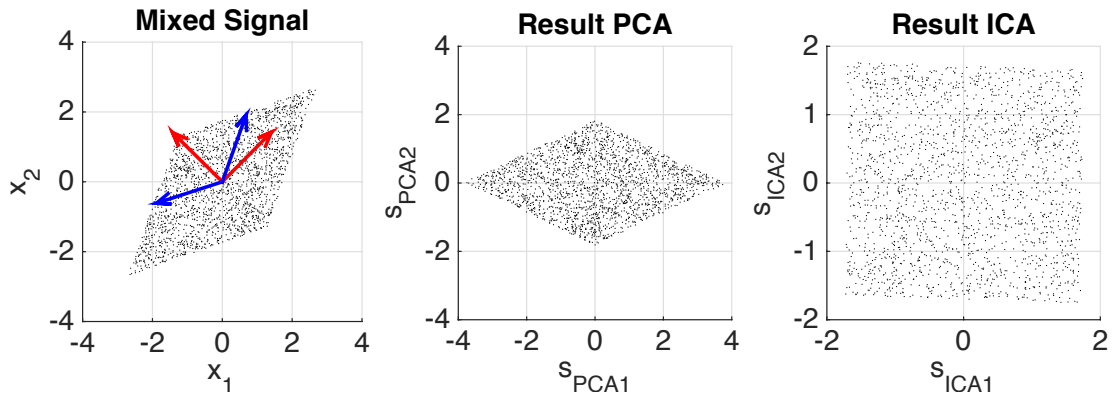


Figure 2.11: Illustrative comparison between ICA and PCA. The mixed signals $X \in \mathbb{R}^{2\times T}$ (left) should be linearly transformed into sources $\hat{S} = \hat{W}X$. The orthogonal filters of PCA are indicated with red arrows, the typically non-orthogonal filters of ICA are blue arrows. (middle) PCA finds uncorrelated sources s_{PCA1}, s_{PCA2}, the first of which explains maximal variance. Resulting sources are uncorrelated, but not necessarily independent. (right) ICA finds independent sources s_{ICA1}, s_{ICA2}.

Multiple ICA algorithms exist, because a number of measures can be used to assess whether two time series are statistically independent. Two broad classes of algorithms can be distinguished: (1) methods which rely on higher-order statistics to define independence, and (2) methods which rely on second-order statistics only by taking temporal dependencies in the time series into account. Note that the term ICA is sometimes used to refer to the first model class only.

Higher-order methods. Following (Hyvärinen and Oja, 2000; Hyvärinen et al., 2001), the sources S can be recovered by relying on higher-order statistics under the following assumptions:

1. The sources $S = (s_1, \ldots, s_K)^\top$ are mutually independent.

   This assumption is stronger than requiring the sources not to be correlated. The sources $s_1, \ldots, s_K$ are uncorrelated when their covariances are 0, that is, when their second order moments are 0. They are statistically independent when a component s_i does not contain any information about another component s_j $(i \neq j)$, that is, when moments of arbitrary order are 0.

2. The sources $s_1, \ldots s_K$ are not normally distributed.

   Intuitively, normally distributed components are 'too simple'. They can be described solely by their first and second order moments, while all higher order moments are 0.

3. It is typically assumed that the mixing matrix A is square.

Note that it is not possible to determine the variance of the independent source components. This is because S and A are unknown, and every scalar factor in one source can be compensated by dividing the corresponding column in the mixing matrix (the pattern) by this factor. Also, the order of the components is arbitrary.

The demixing $\hat{W}$ is estimated by maximizing some criterion of independence (Cardoso and Souloumiac, 1993; Comon, 1994; Bell and Sejnowski, 1995; Amari et al., 1996; Pham and Garat, 1997; Hyvärinen and Oja, 1997; Hyvärinen, 1999; Müller et al., 1999). A good overview can be found in (Hyvärinen and Oja, 2000; Hyvärinen et al., 2001; Comon and Jutten, 2010).

Second-order methods. Second-order methods take the temporal structure of the time series into account and enforce decorrelation over time (Molgedey and Schuster, 1994; Belouchrani et al., 1997; Ziehe and Müller, 1998). In this thesis, we use TDSEP (Temporal Decorrelation source SEParation) (Ziehe and Müller, 1998; Ziehe et al., 2004), which is equivalent to SOBI (Second Order Blind Identification) (Belouchrani et al., 1997). TDSEP/SOBI amount to finding a demixing W which leads to minimal cross-covariances over several time-lags between all pairs of components of S . The assumption is therefore that the sources $s_1, \ldots, s_K$ have non-zero autocovariances. Non-Gaussianity is not required.

TDSEP/SOBI is based on the following steps:

1. Whitening of the data. Whitening is a linear transformation U which decorrelates and scales the data X such that the covariance matrix of the whitened data $X' := UX$ equals the identity. A solution is given by

$$U := \Lambda^{-1/2}\Phi^T. \tag{2.4}$$

   where the column of Φ are the eigenvectors of the covariance matrix $\frac{1}{T}XX^\top$ and Λ is a diagonal matrix which contains the corresponding eigenvalues. (It is easy to see that indeed $\frac{1}{T}X'X'^\top = U\frac{1}{T}XX^\top U^\top = \Lambda^{-1/2}\Phi^T\Phi\Lambda\Phi^T\Phi\Lambda^{-1/2} = I$ holds.)

   Whitening is a useful pre-processing step for ICA algorithms because it reduces the BSS problem to finding an orthogonal demixing matrix. We are now searching for a new demixing matrix W' such that $W'X' = S$, which is orthogonal: $W'W'^\top = W'\frac{1}{T}X'X'^\top W'^\top = \frac{1}{T}SS^\top = I$. For the last equality, we assumed without loss of generality that the independent source components are each scaled to 1.

2. Time-lagged cross-covariance matrices are simultaneously diagonalized. Denote with

$$\hat{C}_{X'}(\tau) := \frac{1}{T}\sum_{t=1}^{T} X'(t)X'^{\top}(t-\tau) \quad \text{and} \quad \hat{C}_S(\tau) := \frac{1}{T}\sum_{t=1}^{T} S(t)S^{\top}(t-\tau) \tag{2.5}$$

the empirical cross-covariance matrices at time lag $\tau \in \mathbb{Z}$, where $X'(t)$ and $S(t)$ are the t-th column of X' and S. Now, note that (1) $W'C_{X'}(\tau)W'$ equals the cross-covariance matrix $\hat{C}_S(\tau)$ of the source components S at time-lag τ; and (2) The independence assumptions yields that this cross-covariance matrix of the source components S at time-lag τ is a diagonal matrix. W' can thus be computed as the matrix that jointly diagonalizes a set of whitened cross-covariances $C_{X'}(\tau)$. Throughout this thesis, we use $\tau = 1, \ldots, 99$.

3. Finally we obtain $W = W' \cdot U$.

2.2.3 Spatio-spectral decomposition (SSD)

The purpose of spatio-spectral decomposition (SSD) (Nikulin et al., 2011) is to extract brain oscillations in a frequency band of interest. It is based on the assumption that noise sources produce signals in a relatively broad frequency range, while the brain oscillations of interest are contained in a narrow spectral band. Algorithmically, it maximizes the signal power in a frequency band of interest while simultaneously minimizing it at neighboring frequencies. That is, SSD seeks spatial filters $\mathbf{w} \in \mathbb{R}^M$ which maximize

$$\mathrm{SNR}(\mathbf{w}) = \frac{\mathbf{w}^{\top}\Sigma_{\mathrm{sig}}\mathbf{w}}{\mathbf{w}^{\top}\Sigma_{\mathrm{noise}}\mathbf{w}} \tag{2.6}$$

where Σ_{sig} is the covariance of the data filtered in the frequency band of interest and Σ_{noise} is the covariance of the data filtered in the sidebands.

The entire SSD demixing matrix can be computed by solving a generalized eigenvalue problem in a matter of seconds (Nikulin et al., 2011). SSD has been shown to be an effective tool for preprocessing, because it extracts a low dimensional subspace which captures oscillatory activity in the frequency range of interest (Haufe et al., 2014a).

2.3 Causal inference

The learning of cause-effect relationships is a challenging task, for which controlled randomized experiments are necessary. However, such experiments are often unethical, expensive or just impossible. For example, research on brain connectivity

tries to identify which brain region communicates with another brain region in the brain's normal activity.

The scientific methodology for the inference of cause-effect relationships from uncontrolled observational data is subject to intense research. While some methods incorporate specific domain knowledge into the model (e.g. (Friston et al., 2003; Patel et al., 2006)), most statistical methods fall into one of the following three classes. First, Causal Bayesian Networks use conditional independence relationships between variables in order to infer a set of possible causal models. For a second class of models, so-called additive noise models, it is possible to show that the causal model can be fully identified under mild conditions (non-linearity or non-gaussianity). For the analysis of time series a third class of models, which is based on temporal precedence, is very popular.

2.3.1 Causal Bayesian Networks

In the framework of Causal Bayesian Networks, cause-effect relations are represented as a graph in which the nodes represent random variables and directed edges depict causal influences (Spirtes et al., 2000; Pearl, 2009; Mumford and Ramsey, 2014). The variables are assumed to have a causal ordering, that is, they can be represented as a directed acyclic graph (DAG).

Knowledge about causal relations of the variables can be gained by identifying conditional independence relationships in the data. Such causal inference is based on theoretical insights, which state that some conditional independency relations characterize certain causal directionalities. Consider for example the following dependencies of three random variables x, y and e: x and e are dependent, y and e are dependent, but x and y are independent. The only natural explanation is that x and y are a common cause of e, that is $x \rightarrow e \leftarrow y$.

Various methods exist for the identification of Causal Bayesian Networks, most of which assume independent and identically distributed (iid) data, an assumption which is violated for EEG data. Note also that several possible models may display the same conditional independence relations. For example, the three models $x \rightarrow e \rightarrow y$, $x \leftarrow e \leftarrow y$ and $x \leftarrow e \rightarrow y$ all imply that x is independent of y conditional on e. Causal Bayesian Network methods may therefore only result in a set of equally likely models.

2.3.2 Acyclic causal models with additive noise

It is possible to uniquely identify a causal model without the use of temporal information by placing additional assumptions on the data generation process. A popular model is the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al., 2006), which assumes that

1. the observed variables can be represented as a directed acyclic graph with additive non-Gaussian noise,
2. the data generating process is linear, and
3. there are no unobserved hidden causes.

Under these assumptions, the causal model can be completely identified by relying on higher order distributional statistics (Shimizu et al., 2006).

The mathematical model can be written as

$$X = B \cdot X + \upsilon \tag{2.7}$$

where $X \in \mathbb{R}^{M \times T}$ contains T data points of M observed variables, $\upsilon \in \mathbb{R}^{M \times T}$ contains mutually independent, non-gaussian innovations (noise), and $B \in \mathbb{R}^{M \times M}$ is a square matrix that could be permuted to strict lower triangularity if one knew a causal ordering. For example, for two variables $x \to y$, the model reads

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \upsilon_x \\ \upsilon_y \end{bmatrix} \tag{2.8}$$

with $b \in \mathbb{R}$.

Now, the key insight is that the observed variables X are a linear combination of the innovations. That is, Equation (2.7) can be rewritten as $X = A \cdot \upsilon$ where $A := (I - B)^{-1}$. We can thus compute A using ICA, and then infer the causal ordering in B from it. Note that it may be more efficient to estimate LiNGAM with procedures other than ICA (Shimizu et al., 2011; Hyvärinen and Smith, 2013). Extensions of the framework consider for example temporal structure

(Hoyer et al., 2009; Zhang and Hyvärinen, 2009; Peters et al., 2014) provided a generalization of additive noise models to nonlinear models. They show that nonlinearities can in fact help to identify the causal model. To illustrate the procedure, let us again consider two variables, x and y, which now stem from the causal model $y = f(x) + \upsilon_y$, where f is an arbitrary function and x is independent from υ_y. (Hoyer et al., 2009) show that, even if both x and υ_y are Gaussian, the joint probability distribution of x and y could only arise from a backward model $x = g(y) + \upsilon_x$ (with y independent of υ_x), if f is linear.

In brief, the causal model $x \to y$ can then be tested as follows:

1. Perform a non-linear regression of y on x to estimate a function f such that $y \approx f(x)$,
2. Perform a significance check whether the residuals $\upsilon_y = y - f(x)$ are independent of x. Accept the causal model $x \to y$ only if the null hypothesis of independence cannot be rejected.

This methods is not directly applicable to time series, because it assumes iid data. Time series data typically exhibit strong autocorrelation structure, which have to be taken into account in most analysis (cf. (Bartz and Müller, 2014)). An interesting extension of additive noise models which allow nonlinear and instantaneous effects to time series has recently been proposed (Peters et al., 2013).

2.3.3 Granger causality and other time-lagged measures

In time-series analysis, inference about cause-effect relationships is commonly based on the principle that the cause should precede its effect. Particularly, the concept of Granger causality (GC) has gained popularity as a simple testable definition of causality based on temporal precedence. Since its introduction in 1969 many extensions have proposed, but we will focus here on the simple bivariate linear case. We will also mention the shortcoming of Granger causality and related measure, and outline two proposed remedies, the Phase Slope Index (PSI) (Nolte et al., 2008) and time reversal testing (Haufe et al., 2013).

Linear Granger causality (GC). In its original formulation, a time series x_t is said to Granger-cause a time series y_t, if the past of x_t helps to predict y_t above what can be predicted by using 'all other information in the universe' besides x_t (Granger, 1969). In practical situations in which only two time series are available, it is common to consider only the information contained in the past of x_t and y_t (Hamilton, 1994).

The standard Granger causality test is given by the comparison of the goodness of fit of two autoregressive (AR) models. First, y_t is modeled as a function of a predefined number p of its most recent past values. Second, y_t is modeled as a function of both its own past values and the past values of x_t. Finally, Granger causality tests whether the second regression model explains significantly more variance of y_t than the first regression model.

Granger causality is grounded in the theory of autoregressive modeling. We assume the data has been generated by a stable bivariate vector autoregressive process of lag order p (VAR(p) process), $\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \mathbb{R}^2$,

$$\mathbf{z}_t = A_1 \mathbf{z}_{t-1} + \ldots + A_p \mathbf{z}_{t-p} + \epsilon_t \ , \tag{2.9}$$

where $\epsilon_t \in \mathbb{R}^2$ is a 2-dimensional white noise process (that is, $\langle \epsilon_t \rangle = 0$, $\langle \epsilon_t \epsilon_{t-h}^\top \rangle = 0$ for $h \neq 0$, and $\langle \epsilon_t \mathbf{z}_{t-h}^\top \rangle = 0$ for $h \in \mathbb{N} \setminus \{0\}$, $\langle \cdot \rangle$ denotes expectation) with residual covariance matrix

$$\Sigma = \langle \epsilon_t \epsilon_t^\top \rangle = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{bmatrix} \ . \tag{2.10}$$

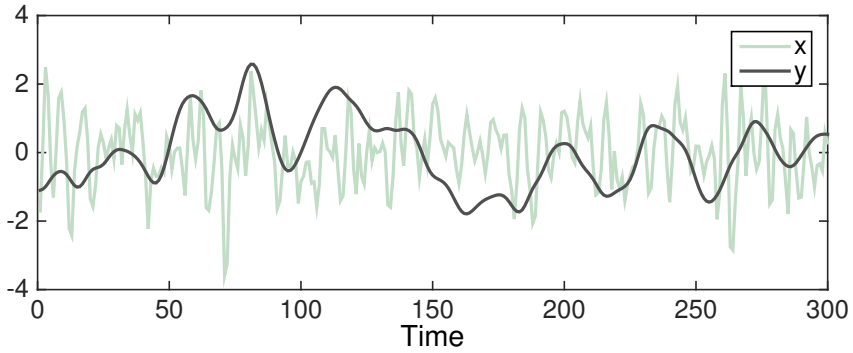


Figure 2.12: Example of a simulated VAR(5) process where x_t Granger causes y_t.

The noise variables ϵ_t are also called innovations or residuals. Stability requires that $\det(I - A_1\lambda - \ldots - A_p\lambda^p) \neq 0$ for all $\lambda \in \mathbb{C}$ with $|\lambda| \leq 1$, and ensures that the process is stationary.

Note that VAR models are able to capture a surprisingly large range of time series dynamics (see Figure 2.12 for an example). In fact, a fundamental result from time series analysis, Wold's decomposition theorem (Wold, 1938), implies that 'under quite general conditions, every stationary, purely nondeterministic, process (without a deterministic component) can be approximated well by a finite order VAR process' (Lütkepohl, 2007).

Following (Geweke, 1982), x_t and y_t possess themselves autoregressive (AR) representations, which we denote by

$$x_t = \sum_{k=1}^{\infty} a_k x_{t-k} + \xi_t^x, \qquad \mathrm{Var}(\xi_t^x) =: \Sigma_x \quad \text{and} \tag{2.11}$$

$$y_t = \sum_{k=1}^{\infty} b_k y_{t-k} + \xi_t^y, \qquad \mathrm{Var}(\xi_t^y) =: \Sigma_y \; . \tag{2.12}$$

where the residuals ξ_t^x and ξ_t^y of these two univariate processes are each serially uncorrelated. Directed Granger-causal information flow is then defined as (Geweke, 1982)

$$F_{y\to x} := \log\left(\frac{\Sigma_x}{\Sigma_{xx}}\right) \quad \text{and} \quad F_{x\to y} := \log\left(\frac{\Sigma_y}{\Sigma_{yy}}\right) \; . \tag{2.13}$$

For an information flow from x to y it holds that $F_{x\to y} > 0$. Under the assumption of Gaussian-distributed residuals, $F_{y\to x}$ and $F_{x\to y}$ are asymptotically χ^2 distributed, giving rise to an analytical test of their significance (Geweke, 1982). An asymptotically equivalent test is given by an F-test of the goodness-of-fit of the two models, cf. (Hamilton, 1994; Bressler and Seth, 2011). If the distribution of the residuals

is unknown, non-parametric methods such as permutation testing can be used for significance testing (Anderson and Robinson, 2001).

As variables in physical systems often mutually influence each other, it is also of interest to determine the *net* driver of the interaction by assessing whether more information is flowing from x_t to y_t then from y_t to x_t or vice versa. Following (Nolte et al., 2008; Nolte et al., 2010), net Granger causality is defined as the difference Granger causality scores, that is

$$F_{x \to y}^{(net)} := F_{x \to y} - F_{y \to x} \tag{2.14}$$

Because the analytical distributions of these differences are unknown, statistical significance of net Granger causality scores needs to be assessed using resampling methods.

Related measures. As the concept of Granger causality has received a great deal of attention, several extensions to non-linear and multivariate models have been proposed (Marinazzo et al., 2008; Barrett et al., 2010; Vicente et al., 2011). Important is also the extension to the spectral domain (Geweke, 1982), in which we can obtain a representation of Granger causality indices as a function of frequency. This feature is helpful in some neuroscientific applications, because neurophysiological interaction often depends on synchronous oscillatory activity in well-known frequency bands. Several other measures related to spectral Granger causality, such as the Directed Transfer Function (DTF) (Kaminski and Blinowska, 1991) and the Partial Directed Coherence (PDC) (Baccalá and Sameshima, 2001) have also gained popularity.

Limitations of Granger causality. While being a widespread tool, Granger causality and related measures suffer from several limitations:

- Temporal precedence does not necessarily imply causality. Consider the following example from economics: interest rates or consumer confidence ratings predict economic development, but they reflect human forward-looking behavior and a causal relationship cannot necessarily be inferred (Hamilton, 1994). Similarly, a rooster who crows before sunrise does not cause the sunrise.

- Hidden common drivers cannot be detected. Already Granger pointed out that standard Granger causality can lead to spurious results if not all all relevant variables are incorporated in the model (Granger, 1969).

- Granger causality is susceptible to measurement noise. If two sensors measuring the same signal are superposed with noise, they mutually help predicting each other's future (Nalatore et al., 2007; Nolte et al., 2008). This is a problem

for example in the study of brain connectivity with EEG, because the activity at a given sensor is a mixture of contributions from several neuronal sources (Gómez-Herrero et al., 2008; Schoffelen and Gross, 2009; Haufe et al., 2010; Ewald et al., 2012; Haufe et al., 2013).

- Spurious Granger causality has also been reported to arise due to downsampling and temporal aggregation (Tiao and Wei, 1976; McCrorie and Chambers, 2006; Zhou et al., 2014). This may pose serious problems for example in functional magnetic resonance imaging (fMRI) (Seth et al., 2013; Smith et al., 2011).

- Like most other methods, Granger causality requires the signals to be wide-sense stationary, meaning that the mean and cross-covariance are not allowed to vary with respect to time. To deal with non-stationarity in the mean, a widely-used remedy is to test for so-called 'cointegration' (Engle and Granger, 1987), a technique which received the Nobel prize of economics in 2003.

Due to these above mentioned limitations, Granger referred to a positive outcome of his test as the identification of a 'prima facie' cause, meaning 'at first sight' or 'until revoked'.

In the last years, the problem of spurious Granger causality in the presence of measurement noise received more attention. Novel causality metrics or remedies which are more robust with respect to measurement noise and volume conduction have been proposed (Nalatore et al., 2007; Nolte et al., 2008; Vicente et al., 2011; Haufe et al., 2013; Vinck et al., 2015). In the following, we will review the Phase Slope Index (PSI) (Nolte et al., 2008), and time-reversed causality testing (Haufe et al., 2013).

Phase-Slope-Index (PSI). One measure which is also based on the idea of temporal precedence, but in contrast to Granger causality not on predictability, is the Phase Slope Index (PSI) (Nolte et al., 2008). PSI has been proposed as a more noise-robust metric of information flow and is based on the slope of the phase of cross-spectra between two time series.

To define PSI, let us first introduce the following quantities. The empirical cross-spectrum $\hat{G}_{xy}$ of two time series x_t and y_t is a complex number computed as the mean over L data segments

$$\hat{G}_{xy}(f) := \frac{1}{L} \sum_{l=1}^{L} \bar{x}(l, f) \bar{y}^*(l, f) \tag{2.15}$$

where $\bar{x}(f,l)$ is the Fourier transform of x in data segment l, and * denotes complex conjugation. The auto-spectra of the two signals

$$\hat{G}_{xx}(f) := \frac{1}{L}\sum_{k=1}^{K} \bar{x}(l,f)\bar{x}^*(l,f) \quad \text{and} \quad \hat{G}_{yy}(f) := \frac{1}{L}\sum_{l=1}^{L} \bar{y}(l,f)\bar{y}^*(l,f) \tag{2.16}$$

are real numbers which corresponds to the mean power of x and y at frequency f. The complex coherency γ_{xy} is then defined as the normalized cross-spectrum

$$\gamma_{xy}(f) := \frac{\hat{G}_{xy}(f)}{\sqrt{\hat{G}_{xx}(f)\hat{G}_{yy}(f)}} \tag{2.17}$$

and describes the linear relationship of two time series at a specific frequency f. The *Phase Slope Index (PSI)* $\Psi_{x\to y}$ is now defined as

$$\Psi_{x\to y} := Im\left(\sum_{f\in\mathcal{F}} \gamma_{xy}^*(f)\gamma_{xy}(f+\delta f)\right) \tag{2.18}$$

where $\mathcal{F}$ is the set of frequencies over which the slope is summed, δf is the frequency resolution and $Im(\cdot)$ denotes taking the imaginary part. Typically, $\mathcal{F}$ contains all frequencies, but it can be restricted to a specific band if desired. Statistical significance can be assessed using resampling techniques.

PSI corresponds to an average of the slope of the phase spectrum. As with Granger causality, PSI is based on the idea that interactions require time. If two waves travel at similar speed, then the phase difference between driver and recipient increases with frequency and we expect a positive slope of the phase spectrum. One advantage of PSI over Granger causality is that it robustly rejects causal interpretations for mixtures of non-interacting signals such as correlated noise sources, because mixtures of independent sources do not induce an imagery part of coherency (Nolte et al., 2004).

Time reversal and contributions of this thesis. Another remedy to avoid false detections of causal interactions was recently proposed by (Haufe et al., 2012; Haufe et al., 2013). They proposed to contrast causality measures applied to the original time series $\mathbf{z}_t$ with the same measures obtained from time-reversed signals $\tilde{\mathbf{z}}_t := \mathbf{z}_{-t}$.

Let us formalize time reversal for Granger causality. Theoretical results of (Andel, 1972) state that the time-reversed signal of any VAR(p) process has again a VAR(p) representation that can be expressed analytically in terms of the original process. That is, any stable VAR(p) process (2.9) $\mathbf{z}_t$ has a time-reversed representation

$$\mathbf{z}_t = \tilde{A}_1\mathbf{z}_{t+1} + \tilde{A}_2\mathbf{z}_{t+2} + \ldots + \tilde{A}_p\mathbf{z}_{t+p} + \tilde{\epsilon}_t \tag{2.19}$$

that is again of order p with uniquely defined autoregressive coefficients $\tilde{A}_1, \ldots \tilde{A}_p$ and residual covariance matrix

$$\tilde{\Sigma} = \begin{bmatrix} \tilde{\Sigma}_{xx} & \tilde{\Sigma}_{xy} \\ \tilde{\Sigma}_{xy} & \tilde{\Sigma}_{yy} \end{bmatrix} . \tag{2.20}$$

In analogy to the original time series, let us define the time-reversed Granger scores as

$$\tilde{F}_{\tilde{y}\to\tilde{x}} := \log\left(\frac{\tilde{\Sigma}_x}{\tilde{\Sigma}_{xx}}\right) \quad \text{and} \quad \tilde{F}_{\tilde{x}\to\tilde{y}} := \log\left(\frac{\tilde{\Sigma}_y}{\tilde{\Sigma}_{yy}}\right) , \tag{2.21}$$

where $\tilde{\Sigma}_x$ and $\tilde{\Sigma}_y$ denote the residual variances from the time-reversed restricted models

$$x_t = \sum_{k=1}^{\infty} \tilde{a}_k x_{t+k} + \tilde{\xi}_t^x, \quad \mathrm{Var}(\tilde{\xi}_t^x) =: \tilde{\Sigma}_x \quad \text{and} \tag{2.22}$$

$$y_t = \sum_{k=1}^{\infty} \tilde{b}_k y_{t+k} + \tilde{\xi}_t^y, \quad \mathrm{Var}(\tilde{\xi}_t^y) =: \tilde{\Sigma}_y . \tag{2.23}$$

The net Granger causality scores are defined as

$$\tilde{F}^{(net)}_{\tilde{x}\to\tilde{y}} := \tilde{F}_{\tilde{x}\to\tilde{y}} - \tilde{F}_{\tilde{y}\to\tilde{x}} . \tag{2.24}$$

Finally, we define the differences of the Granger scores obtained on original and time-reversed signals as

$$\tilde{D}^{(net)}_{x\to y} := F^{(net)}_{x\to y} - \tilde{F}^{(net)}_{\tilde{x}\to\tilde{y}} . \tag{2.25}$$

In Difference-based time-reversed Granger causality (TRGC), we then infer a net information flow from x_t to y_t if

$$\tilde{D}^{(net)}_{x\to y} > 0 , \tag{2.26}$$

that is, we require that net Granger causality from x_t to y_t is reduced on the time-reversed signals.

(Haufe et al., 2013) showed that TRGC robustly rejects causal interpretations for mixtures of independent noise sources. The mathematical basis for the noise robustness property of time-reversal are the following simple observations:

1. The transpose of the cross-covariance matrices $C_{\mathbf{z}}(\cdot)$ of $\mathbf{z}_t$ is equal to the cross-covariance matrices $\tilde{C}_{\tilde{\mathbf{z}}}(\cdot)$ of the time-reversed series $\tilde{\mathbf{z}}_t$, that is

$$\tilde{C}_{\tilde{\mathbf{z}}}(h) = \langle \tilde{\mathbf{z}}_t \tilde{\mathbf{z}}_{t-h}^\top \rangle = \langle \mathbf{z}_t \mathbf{z}_{t+h}^\top \rangle = C_{\mathbf{z}}(-h) = \big(C_{\mathbf{z}}(h)\big)^\top \tag{2.27}$$

where $h \in \mathbb{Z}$, $\langle\cdot\rangle$ denotes expectation, and $\mathbf{z}$ resp. $\tilde{\mathbf{z}}$ are assumed to have been transformed to have zero mean.

2. If a series η_t only contains a mixture of independent noise sources, all its cross-covariance matrices will be symmetric, that is $\big(C_\eta(h)\big)^\top = C_\eta(h)\ \forall h \in \mathbb{Z}$ (Nolte et al., 2006).

For mixtures of independent noise sources, any causality measure that is solely based on a series' cross-covariance matrices (such as standard linear Granger causality) therefore yields the same result on the original and the time-reversed signals. Their difference equals zero.

Time-reversed Granger causality (TRGC) thus displays an intriguing noise robustness property. Furthermore, it also yielded very encouraging results in simulations (Haufe et al., 2013; Vinck et al., 2015) and is fast to compute. However, its behavior in the presence of causal interactions was still poorly understood. In particular, it was unclear how Granger causality scores computed on time-reversed signals link to the causal interactions on the original time-series, and therefore whether TRGC correctly indicates the direction of causality.

The intuitive idea behind time reversed testing is that, if temporal order is crucial to tell a driver from a recipient, directed information flow should be reduced if the temporal order is reversed. However, a mathematical proof for this intuition is required. Theoretical guarantees have only been derived for special cases in which either the signal's auto- and cross-covariances are very small in magnitude, or in which both signals have very similar autocorrelations (Vinck et al., 2015).

In Chapter 5, we study the time-reversal of (linear) finite-order VAR processes to prove that, in the case of unambiguous unidirectional information flow from x to y, difference-based TRGC indeed yields the correct result $\tilde{D}^{(net)}_{x\to y} \geq 0$ (Winkler et al., 2015c).

2.4 List of abbreviations

AR	autoregressive
BCI	Brain-Computer-Interface
BSS	Blind Source Separation
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
ERP	Event-Related Potential
ERD	Event-Related Desynchronization
fMRI	functional Magnetic Resonance Imaging
GrangerCPA	Granger Causal Power Analysis
ICA	Independent Component Analysis
IC	Independent Component
LDA	Linear Discriminant Analysis
MARA	Multiple Artifact Rejection Algorithm
PCA	Principal Component Analysis
PSI	Phase Slope Index
SNR	Signal-to-noise ratio
SPoC	Source Power Correlation
SSD	Spatio-Spectral Decomposition
CSP	Common Spatial Patterns
SVM	Support Vector Machine
TRGC	Time-reversed Granger causality
VAR(p) process	Vector autoregressive process of order p

3 Automatic artifact removal for EEG signals

3.1 Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG signals

Irene Winkler, Stefan Haufe and Michael Tangermann. Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals. *Behavioral and Brain Functions*, 7:30, 2011
`http://www.behavioralandbrainfunctions.com/content/7/1/30/`

Short summary. We constructed a linear component classification method that automates the process of hand-selection of artifactual independent components. We later call this algorithm MARA (Multiple Artifact Rejection Algorithm). The core of MARA is a linear classifier based on six features from the spatial, the spectral and the temporal domain. Features were optimized to solve the binary classification problem 'reject vs. accept'. Thus, the classifier is not limited to a specific type of artifact, and should be able to handle eye artifacts, muscular artifacts and loose electrodes equally well.

The paper describes the feature selection procedure and the validation of the classifier in detail. After the construction of the classifier on labeled ICA components from a reaction time experiment, its performance was validated on new components of the same study and two other EEG studies.

Contributions. This was my first scientific paper, and the adopted research methodology was closely supervised by Michael Tangermann. I wrote the majority of the paper and carried out all implementations, except for the current density norm feature provided by Stefan Haufe.

Winkler et al. Behavioral and Brain Functions 2011, 7:30
http://www.behavioralandbrainfunctions.com/content/7/1/30

BEHAVIORAL AND BRAIN FUNCTIONS

METHODOLOGY Open Access

Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals

Irene Winkler*, Stefan Haufe and Michael Tangermann

Abstract

Background: Artifacts contained in EEG recordings hamper both, the visual interpretation by experts as well as the algorithmic processing and analysis (e.g. for Brain-Computer Interfaces (BCI) or for Mental State Monitoring). While hand-optimized selection of source components derived from Independent Component Analysis (ICA) to clean EEG data is widespread, the field could greatly profit from automated solutions based on Machine Learning methods. Existing ICA-based removal strategies depend on explicit recordings of an individual's artifacts or have not been shown to reliably identify muscle artifacts.

Methods: We propose an automatic method for the classification of general artifactual source components. They are estimated by TDSEP, an ICA method that takes temporal correlations into account. The linear classifier is based on an optimized feature subset determined by a Linear Programming Machine (LPM). The subset is composed of features from the frequency-, the spatial- and temporal domain. A subject independent classifier was trained on 640 TDSEP components (reaction time (RT) study, n = 12) that were hand labeled by experts as artifactual or brain sources and tested on 1080 new components of RT data of the same study. Generalization was tested on new data from two studies (auditory Event Related Potential (ERP) paradigm, n = 18; motor imagery BCI paradigm, n = 80) that used data with different channel setups and from new subjects.

Results: Based on six features only, the optimized linear classifier performed on level with the inter-expert disagreement (<10% Mean Squared Error (MSE)) on the RT data. On data of the auditory ERP study, the same pre-calculated classifier generalized well and achieved 15% MSE. On data of the motor imagery paradigm, we demonstrate that the discriminant information used for BCI is preserved when removing up to 60% of the most artifactual source components.

Conclusions: We propose a universal and efficient classifier of ICA components for the subject independent removal of artifacts from EEG data. Based on linear methods, it is applicable for different electrode placements and supports the introspection of results. Trained on expert ratings of large data sets, it is not restricted to the detection of eye- and muscle artifacts. Its performance and generalization ability is demonstrated on data of different EEG studies.

Background

Signals of the electroencephalogram (EEG) can reflect the electrical background activity of the brain as well as the activity which is specific for a cognitive task during an experiment. As the electrical field generated by neural activity is very small, it can only be recognized by EEG if large assemblies of neurons show a similar behavior. Resulting neural EEG signals are in the range of micro volts only and can easily be masked by artifactual sources. Typical artifacts of the EEG are caused either by the non-neural physiological activities of the subject or by external technical sources. Eye blinks, eye movements, muscle activity in the vicinity of the head (e.g. face muscles, jaws, tongue, neck), heart beat, pulse and Mayer waves are examples for physiological artifact sources, while swaying cables in the magnetic field of the earth, line humming, power supplies or transformers can be the cause of technical artifacts.

Brain-Computer Interfaces (BCI) are based on the single trial classification of the ongoing EEG signal and can improve the life quality of disabled individuals especially

* Correspondence: irene.winkler@tu-berlin.de
Machine Learning Laboratory, Berlin Institute of Technology, Franklinstr. 28/29, 10587 Berlin, Germany

in combination with other assitive technology [1]. The exclusion of artifacts is of special interest for BCI applications, as the intended or unconscious use of artifacts for BCI control are usually not desirable when the BCI system is tested on healthy subjects. Furthermore, as averaging methods have to be avoided, these real-time systems BCIs rely on relatively clean EEG signals. The same holds true for other Mental State Monitoring applications, that monitor a subject's mental state continuously and on a fine granular time resolution to detect changes e.g. of wakefulness, responsiveness or mental workload as early as possible [2].

The two physiological artifacts most problematic for BCI applications are ocular (EOG) and muscle (EMG) artifacts. EOG activity is either caused by rolling of the eyes or by eye blinks which occur approx. 20 times per minute [3]. Both result in a low-frequency activity most prominent over the anterior head regions, with maximal frequencies below 4 Hz. In contrast, EMG activity (caused by chewing, swallowing, head or tongue movements) is usually a high-frequency activity (>20 Hz) which ranges from rather small to very large amplitudes [4].

For an extensive review of artifact reduction techniques in the context of BCI-systems, the reader can refer to Fatourechi et al. [5]. Since the rejection of artifactual trials amounts to a considerable loss of data, a method that removes the artifacts while preserving the underlying neural activity is needed. For example, linear filtering is a simple and effective method if artifactual and neural activity are located in non-overlapping frequency bands. Unfortunately, artifacts and the brain signal of interest do usually overlap. Nevertheless, ocular activity can be partially removed by regression-based methods, which subtract a part of the activity measured at additional electrooculogram (EOG) channels from the EEG (see [6] for a review). Regression-based methods require the reliable recording of additional EOG channels and are limited by the fact that the EOG is contaminated by brain activity which is removed as well. Furthermore, they cannot eliminate non-eye activity.

If artifactual signal components and neural activity of interest are not systematically co-activated due to a disadvantageous experimental design, methods of Blind Source Separation (BSS) like Independent Component Analysis (ICA) are promising approaches for their separation [7,8]. A common approach is the transformation of the EEG signals into a space of independent source components, the hand-selection of non-artifactual neural sources and the reconstruction of the EEG without the artifactual components (for an example of independent source components, see Figure 1). While assumptions for the application of ICA methods are only approximately met in practice (linear mixture of independent components, stationarity of the sources and the mixture, and prior knowledge about the number of components), their application usually leads to a good separation, with only a small number of hybrid components that contain both, artifacts and neural signals [9-12].

Existing methods for artifact rejection can be separated into hand-optimized, semi-automatic and fully automatic approaches. Semi-automatic approaches require user interaction for ambiguous or outlier components [13,14]. While fully automated methods were proposed for the classification of eye artifacts [15,16], these methods do not easily generalize to non-eye artifacts or even require the additional recording of the EOG [17,18]. Viola et al. and Mognon et al. [19,20] both developed an EEGLAB plug-in which finds artifactual independent components. Both plug-ins have a fully automatic mode that has been shown to recognize and reject major artifacts like eye blinks, eye movements and heart beats, while the detection of muscular or more subtle artifacts has not been reported. The plug-in developed by Viola et al. relies on a user-defined template, while Mognon's approach does not require user interaction.

Existing more flexible approaches for the general classification of different artifact types were reported for EEG data of epileptic patients [21], where the authors report a Mean Squared Error (MSE) of approx. 20% for their system based on a Bayesian classifier. Halder et al. [22] report a classification error below 10% for their Support Vector Machine (SVM) based system for a fixed number of electrodes if dedicated artifact recordings are available for the classifier training. But even if such optimized conditions are present, difficulties of separating muscle artifact components from neural components are common [22].

The review of the existing literature did not reveal a systematic screening of potentially discriminant features for the general task of artifact detection/removal. Moreover, most approaches restrict themselves to part of the available information, e.g. rely on spatial patterns only [19], or spatial patterns and spectral features [22], or spatial pattern and temporal features [20].

Our proposed solution for a general artifact detection method is motivated by the needs of EEG practitioners. First, it is desirable that a method efficiently and reliably detects all classes of artifacts, e.g. is not restricted to eye-, heart beat-, or muscle artifacts. Second, a practical method must be applicable post-hoc, i.e. without the need of dedicated artifact recordings at the time of the experiment. Third, it is difficult to convince EEG practitioners to use a method of artifact rejection if it is a black box and refuses introspection. As the goal must be to develop a method, that delivers interpretable and

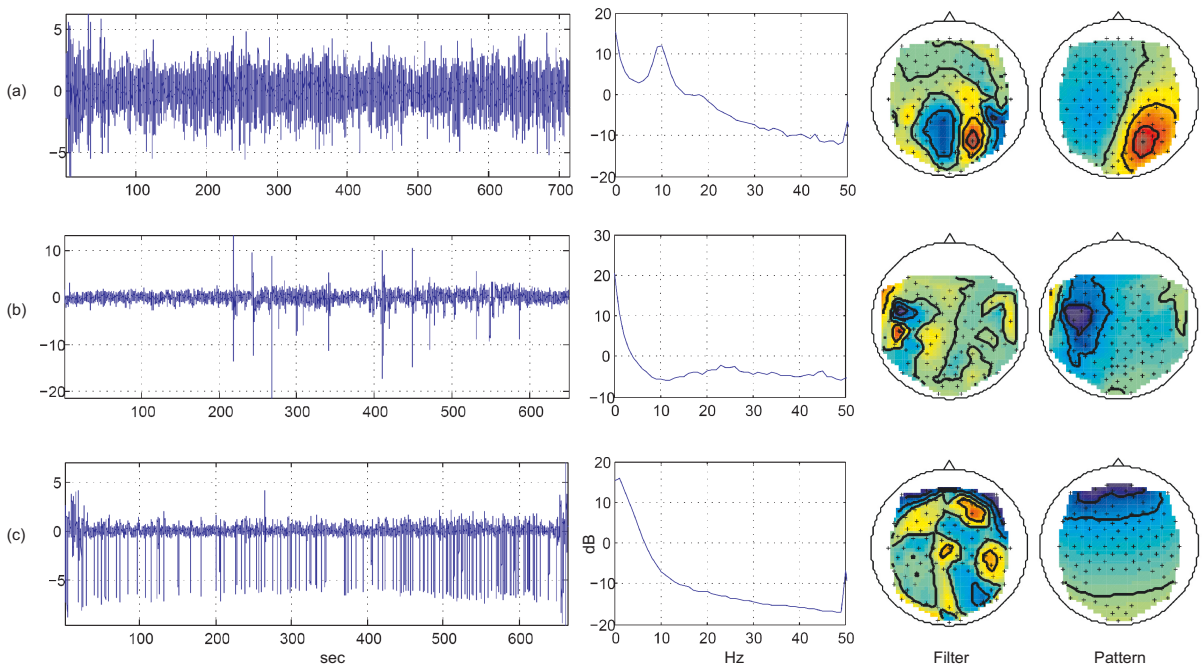


Figure 1 Three example independent source components. Time series (first column), spectrum (second column), filter (third column) and pattern (fourth column) of three components. The first row (a) shows an alpha generator in the occipital lobe. The second row (b) shows a rare muscle artifact component with an increased spectrum in higher frequencies. The third line (c) shows an eye artifact component that appears regularly, has an increases spectrum in lower frequencies and a typical front-back distribution in the pattern.

easy to understand results, we decided for a linear classification method. Luckily, linear methods have proven a high performance for a number of classification tasks in the field of EEG-based BCI systems. However, to be able to estimate the performance loss compared to a potentially better, but difficult to interpret, non-linear classification method, the results of a Gaussian SVM are reported in parallel.

We decided to use a sparse approach (sparsity in the features) although it is a mixed blessing. It leads to a trade-off between efficiency and interpretability, as redundant but slightly less discriminative features are removed with high probability from the overall set of features. This has to be kept in mind during the analysis of results. To reach the goal of a sparse method that delivers physiologically interpretable results, we decided to incorporate a thorough feature selection procedure in combination with a linear classification method that is based on features of all three available information domains of EEG data: the spatial domain (e.g. patterns of independent components), the frequency domain and the temporal domain.

The paper is organized as follows: In the methods section, a reaction time (RT) paradigm is introduced, as data from this study forms the basis for the construction of the proposed artifact detection method. After the signal pre-processing methods (including a temporal variant of ICA) are introduced, we describe 38 features that are candidates for the artifact discrimination task. Based on labels provided by EEG experts, a thorough feature selection procedure is described, that is used to condense the 38 features to a small subset. Furthermore, classification methods are introduced. The methods section ends with a description of two other EEG paradigms (auditory Event Related Potential (ERP) and motor imagery for BCI), that will be used to validate the generalization approach of the proposed artifact classifier. In the results section, the outcome of the feature selection procedure is given, together with the artifact classification performance on unseen data of the RT paradigm, data of a unseen auditory ERP paradigm. Finally the method is applied in the context of a motor imagery BCI setup, before the paper closes with a discussion.

Methods

In the following subsections, we will describe how the proposed new artifact classification method is set up. Then we will introduce two further studies that are utilized to test the classifier's generalizability.

Participants of the studies described below provided verbal and written informed consent and were free to

stop their participation at any time. All collected data was anonymized before any subsequent analysis or presentation took place.

Classifier Construction using a RT study

The artifact classifier is set up based on labeled independent components gained from a reaction time (RT) study.

Experimental Setup

Data from 12 healthy right-handed male subjects were used to train and to test the proposed automated component classification method. Every subject participated in one EEG recording session of approx. 5 hours duration. EEG was recorded from 121 approx. equidistant sensors and high pass filtered at 2 Hz. During this session, 4 repeated blocks of 3 different conditions (C0, C1, C2) were performed. Each block lasted approx. 45 minutes. During all three conditions, subjects performed a forced-choice left or right key press reaction time task upon two auditory stimuli in an oddball paradigm. The key press actions were performed with micro switches attached to the index fingers. During condition C0 subjects had to gaze at a fixation cross without any further visual task. Condition C1 introduced an additional distraction, as a video of a driving scene had to be watched passively on a screen. Condition C2 introduced an additional second task: subjects infrequently had to follow simple lane change instructions and control a steering wheel. By design, EEG recordings under condition C2 were inevitably more prone to muscle and eye artifacts, while C1 possibly stimulated eye movement artifacts, but not muscle artifacts. However, all subjects had been instructed during all conditions to avoid producing artifacts.

Unmixing and data split

To avoid the artificial split of signal components due to the high dimensionality of the data, the separation of the EEG signals by an ICA method was preceded by a dimensionality reduction by Principal Component Analysis (PCA) from 121 EEG channels in the sensor space into $k = 30$ PCA components. This choice of k was based on previous experience, but was probably not the optimal choice. The TDSEP algorithm (Temporal Decorrelation source SEParation) [23] was used to transform the 30 PCA components into 30 independent source components. PCA and TDSEP were applied in a subject specific way, i.e. PCA and TDSEP matrices were calculated seperately for each subject.

TDSEP is a BSS algorithm to estimate a linear demixing

$$WX = S \tag{1}$$

of a given multivariate time series $X = (x_1,..., x_k)^T$ into unknown, assumed mutually independent source components $S = (s_1,..., s_k)^T$. Note that both the demixing W and the source components S are unknown, and that BSS algorithms differ in the definition of independence between components. While ICA algorithms exploit higher order statistics, TDSEP relies on second-order statistics by taking the temporal structure of the time series into account. TDSEP amounts to finding a demixing W which leads to minimal cross-covariances over several time-lags between all pairs of components of S.

For a mathematical discussion, let $\Sigma(\tau) := E(X_w(t)X_w^T(t-\tau))$ be the cross-covariance matrix of the whitened data X_w at time-lag τ, where the whitening transformation linearly decorrellates and scales the data such that $\Sigma(0) = I$. Consider now that (1) Whitening reduces the BSS problem to finding an orthogonal demixing matrix $\tilde{W}$; (2) $\tilde{W}\Sigma(\tau)\tilde{W}^T$ equals the cross-covariance matrix of the source components S at time-lag τ; and (3) The independence assumption yields that the cross-covariance matrix of the source components S at time-lag τ is a diagonal matrix. TDSEP thus computes $\tilde{W}$ as the matrix that jointly diagonalizes a set of whitened cross-covariances $\Sigma(\tau)$. Here we use $\tau = 1,..., 99$.

In the context of EEG signals, TDSEP finds k independent components contributing to the scalp EEG. They are now characterized by their time course, a spatial pattern given by the respective column of the mixing matrix $A := W^{-1}$, and a spatial filter given by the respective row of the demixing matrix W. The pattern contains the projection strengths of the respective component onto the scalp electrodes, whereas the filter gives the projection strength of the scalp sensors onto the source component (see, e.g [24]). All resulting source components were hand labeled into artifactual and non-artifactual components by two experts who each labeled one half of the ICA components based on four plots per component, namely the time series, the frequency spectrum and one scalp plot of the component's filter and one of its pattern. Not all components were unambigious but instead contained a mixture of neural and artifactual activity. Discarding all those components which contain traces of artifacts would remove too much of the relevant neural activity. Therefore, only those mixed components were labeled as artifacts, that revealed a relatively small amount of neural activity compared to the strength of the artifact contained.

For the training of the proposed automated classification method, 23 EEG recordings of 10 minutes duration were taken from the first experimental block only, leading to 690 labeled source components. Neural

components and artifact components were approx. equally distributed (46% vs. 54%). Figure 1 shows typical examples of two artifacts and one neural component.

The trained classifier was tested on 36 unseen EEG recordings from the third experimental blocks. Among these 1080 source components were 47% neuronal components and 53% artifact components.

Feature Extraction

In order to provide substantial information to an automated classification method, we construct an initial feature set that contains 13 features from a component's time series, 9 features from its spectrum and 16 from its pattern. Based on this collection of 38 features a subset of the most discriminative features is determined in a feature selection procedure.

Features derived from a component's time series

1. ***Variance*** of a component's time series. It is not possible to determine the variances of the independent components, as both S and $A := W^{-1}$ are unknown, and the solution is thus undetermined up to scaling. We estimate the impact one independent component s_i has on the original EEG by calculating $\mathrm{Var}(\mathrm{std}(A_i) \cdot s_i)$ where A_i denotes the respective pattern. The idea here is to calculate the standard deviation of one independent component when its corresponding pattern has unit variance.
2. ***Maximum Amplitude***
3. ***Range*** of the signal amplitude
4. ***Max First Derivative***, approximated for the discrete signal $s(t)$ in t_i by $s'(t_i) \approx \frac{s(t_{i+1})-s(t_i)}{1}$
5. ***Kurtosis***
6. ***Shannon Entropy***
7. ***Deterministic Entropy***, a computationally tractable measure related to the Kolmogorov complexity of a signal [25]
8. ***Variance of Local Variance*** of time intervals of 1 s and of 15 s duration (2 separate features)
9. ***Mean Local Variance*** of time intervals of 1 s duration, and of 15 s duration (2 separate features)
10. ***Mean Local Skewness***, the mean absolute local skewness of time intervals of 1 s and 15 s duration (2 separate features)

The above 13 features were all logarithmized in a last step. With exception of the *Variance* feature all were calculated after standardization of the time series to variance 1. These features describe outliers in terms of unusual high amplitude values, as they are typically present in blinks and muscle artifacts. Furthermore, they are sensitive to non-stationarities and non-normal higher order moments in the time series signal, as they can be expected by muscle activity which typically is not present equally strong over the full duration of 10 min.

Features derived from a component's spectrum

1. ***k_1, λ, k_2*** and ***Fit Error*** describe the deviation of a component's spectrum from a prototypical 1/frequency curve and its shape. The parameters k_1, λ, k_2 >0 of the curve

$$f \mapsto \frac{k_1}{f^{\lambda}} - k_2 \tag{2}$$

are determined by three points of the log spectrum: (1) value at 2 Hz, (2) local minimum in the band 5-13 Hz, (3) local minimum in the band 33-39 Hz. The logarithm of k_1, λ, k_2 and of the mean squared error of the approximation to the real spectrum are used as features.

The spectrum of muscle artifacts, characterized by unusual high values in the 20-50 Hz range, are thus approximated by a comparatively steep curve with high λ and low k_1.

2. ***0-3 Hz, 4-7 Hz, 8-13 Hz, 14-30 Hz, 31-45 Hz***, the average log band power of the δ (0-3 Hz), θ (4-7 Hz), α (8-13 Hz), β (14-30 Hz) and γ (31-45 Hz) band.

Features derived from a component's pattern

1. ***Range Within Pattern***, logarithm of the difference between the minimal and the maximal activation in a pattern
2. ***Spatial Distance of Extrema***, logarithm of the Euclidean norm of the 2D-coordinates of the minimal and maximal activation in a pattern
3. ***Spatial Mean Activation Left, Left Frontal, Frontal, Right Frontal, Right, Occipital, Central***, logarithm of the average activation in 7 groups of electrodes as depicted in Figure 2
4. ***2DDFT***. Pattern without a "smooth" activity distribution do not originate from an easily traceable psychological source and are thus artifacts or mixed components. The spatial frequency of a pattern can be described by means of a two-dimensional discrete Fourier transformation. As a first step, the pattern is linearly interpolated to a quadratic 64x64 pattern matrix. The feature 2DDFT is the average logarithmic band power of higher frequencies of the 1st and 4th quadrant (see Figure 3) of the 2D-Fourier spectrum of the pattern matrix.
5. ***Laplace-Filter***. Laplace-filtering leads a second way of finding spatially high frequent patterns, as these have more defined edges. Similar to the 2DDFT-Feature, the pattern is linearly interpolated to a quadratic 64 × 64 pattern matrix. Then, a 3 × 3 Laplace filter is applied. The feature is defined as the

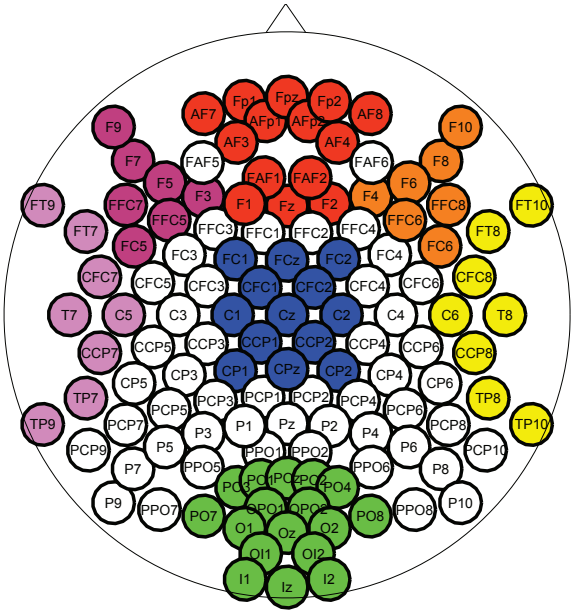


Figure 2 Scalp electrode sets. Mean activation in the 7 colored electrode groups are used as features.

logarithm of the Frobenius norm of the resulting matrix.

6. ***Border Activation***. This binary feature captures the spatial distribution at the borders of a pattern. It is defined as 1 if either the global maximum of the pattern is located at one of the outmost electrodes of the setup in Figure 4 (right), or if the local maximum of an electrode group in Figure 4 (left) is located at the outmost electrode of the group and if that local maximum deviates at least 2 standard deviations from the group average. Otherwise the feature is defined as -1. The idea behind this feature is that a pattern with maximal activation at its border is unlikely to be generated by a source inside the brain - it thus indicates an artifact.

7. ***Current Density Norm*** of estimated source distribution and strongest source's position ***x, y, z***. ICA itself does not provide information about the

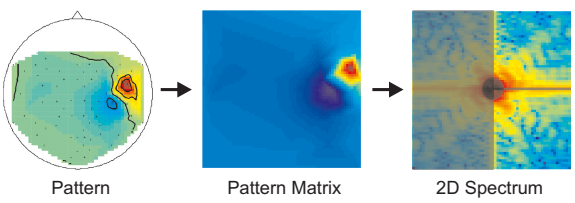


Figure 3 The feature 2DDFT. The feature 2DDFT is the average logarithmic band power of higher frequencies of the 1st and 4th quadrant of the 2D-Fourier spectrum of the pattern matrix.

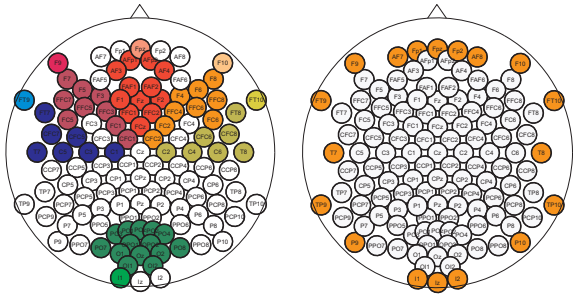


Figure 4 The feature Border Activation. Electrode groups (left) and electrodes (right) used to determine the feature Border Activation.

locations of the sources S. However, ICA patterns can be interpreted as EEG potentials for which a physical model is given by $a = Fz$. Here, $z \in \mathbb{R}^{3m}$ are current moment vectors of unknown sources at m locations in the brain and $F \in \mathbb{R}^{k\times\ 3\ m}$ describes the mapping from sources to k sensors, which is determined by the shape of the head and the conductivities of brain, skull and skin tissues. We consider $m = 2142$ sources which are arranged in a 1 cm grid. Source estimation can only be done under additional constraints since $k \ll m$. Commonly, the source distribution with minimal l_2-norm (i.e., the "simplest" solution) is sought [26]. This leads to estimates

$$\min_z ||Fz - a||^2 + \lambda||\Gamma z||^2 = (F^TF + \lambda\Gamma^T\Gamma)^{-1}F^Ta := J_\lambda a \quad (3)$$

where Γ approximately equalizes the cost of dipoles at different depths [27] and λ defines a trade-off between the simplicity of the sources and the fidelity of the model.

Since Eq. 3 models only cerebral sources, it is natural that noisy patterns and patterns originating outside the brain can only be described by rather complicated sources, which are characterized by a large l_2-norm. For an example, see Figure 5. We propose to use $f := \log||\Gamma z|| = \log||\Gamma J_\lambda \tilde{a}||$ as a feature for discriminating physiological from noisy or artifactual patterns. Here $\tilde{a} := a/||a||$ are normalized ICA patterns and $\lambda = 100$ was chosen from {0,1,10,100,1000} by cross-validation. To allow for a meaningful comparison of different f values over settings of varying numbers of electrodes, we pre-calculated ΓJ_λ on 115 electrodes and used only those rows that corresponded to the recorded electrodes. Note that while this approach is simple, it may not be the optimal choice when the set of electrodes varies.

Assuming a pattern is generated by only one source, we can estimate its 3D-coordinates x, y, z as the

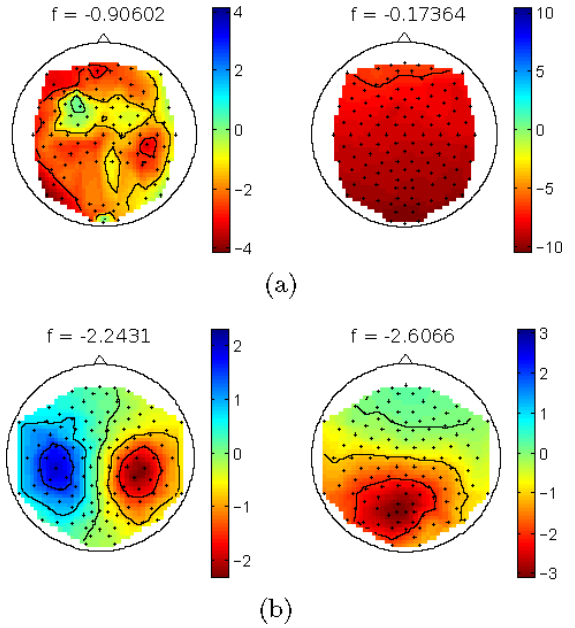


Figure 5 The feature Current Density Norm. (a) Two example patterns with high Current Density Norm $f = \log ||\Gamma z||$. (b) Two example patterns with low Current Density Norm.

location of maximal current density. Note that this is only a very simple source localization method.

Feature Selection and Classification

We conduct an embedded feature selection by using the weight vector of a Linear Programming Machine (LPM) [28]. Like all binary linear classifiers it finds a separating hyperplane $H : \mathbb{R}^d \ni \mathbf{x} \mapsto \text{sign}(\mathbf{w}^T \cdot \mathbf{x} + b) \in \{-1, 1\}$ characterized by a weight vector $\mathbf{w}$ and a bias term b. If the features are zero-mean and have same variance, their importance for the classification task can be ranked by their respective absolute weights $|w_i|$. The LPM is known to produce a sparse weight vector $\mathbf{w}$ by solving the following minimization problem:

$$\begin{aligned} &\min \ ||\mathbf{w}||_1 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } \ y_i(\mathbf{w}^T \cdot \mathbf{x} + b) \geq 1 - \xi_i \ \ (i = 1, \ldots, n) \\ &\qquad \xi_i \geq 0 \ \ (i = 1, \ldots, n) \end{aligned} \tag{4}$$

We thus apply a LPM to the training data in a 5 × 10 cross-validation procedure with the goal to obtain a ranking of the features according to $|E(w_i/||w_i||)|$. Beforehand, the LPM-hyperparameter C was set to C = 0.1 by a 5 × 10 cross-validation heuristic, such that LPM yielded good classification results while using a sparse feature vector, i.e. we selected C with the minimal number of features essential for the classification task (defined by $|E(w_i/||w_i||)|$ >0.1) while the cross-validation error deviates less than one standard error from the minimal cross-validation error.

Having obtained a ranking of the features, the additional information needed is how many of the best-ranked features are optimal for classification. With the goal in mind to find a good trade-off between feature size and error we proceed as follows: For every rank position, we compute the cross-validation error obtained by a classification based on the best-ranked features. Then the number of best ranked features is selected to be the minimum number of features yielding a cross-validation error which deviates less than one standard error from the minimal cross-validation error.

Obviously, the number of features depends on the classification method. We compare a LPM, a non-linear Support Vector Machine (SVM) with Gaussian kernel [29] and a regularized Linear Discriminant Analysis (RLDA) [24], where we use a recently developed method to analytically calculate the optimal shrinkage parameter for regularization of LDA [30,31]. Since a nested cross-validation is computational expensive, the hyperparameters of SVM and LPM are set by an outer cross-validation, i.e. they are estimated on the whole training set which leads to a slight overfitting on the training data.

As a last step, the final classifier was trained on the full training data (690 examples) on the selected features, and tested on unseen test data (1080 examples).

Validation in an auditory ERP study

To evaluate the artifact detection performance beyond the training domain, data from 18 healthy subjects were used to test the proposed automated component classification method in a completely different setup of an auditory ERP study.

Experimental Setup

A group of 18 subjects of 20 to 57 years of age (mean = 34.1, SD = 11.4) underwent an EEG recording of approx. 30 min duration using 64 Ag/AgCl electrodes of approx. equidistant sensors. EEG was band-pass filtered between 0.1-40 Hz. Note that this setup differs from the RT experiment, where EEG was recorded from 121 electrodes and high-pass filtered at 2 Hz.

The subjects were situated in the center of a ring of six speakers (at ear height). During several short trials they listened to a rapid sequence (Stimulus Onset Asynchrony = 175 ms) of six auditory stimuli of 40 ms duration. The six stimuli varied in pitch and noise. Each stimulus type was presented from one speaker only, and each speaker emitted one stimulus type only such that *direction* was a discriminant cue in addition to the pitch/noise characteristics. Subjects had to count the number of appearances of a rare target tone, that was

presented in a pseudo-random sequence together with 5 frequent non-target tones (ratio 1:5).

Unmixing and Classification

A PCA reduced the dimensionality of the EEG channels to 30 PCA components. Then, the TDSEP algorithm was used to transform the 30 PCA components into 30 independent source components. The resulting 540 source components were hand labeled by two experts into artifactual and non-artifactual source components. One of the experts had participated in the rating of the RT-study. Both experts rated all independent components. On average, the experts identified 28% neuronal components and 72% artifactual components (expert 1: 25% neuronal components, expert 2: 31% neuronal components). The labeled data was used to test how the artifact classifier generalizes to new data acquired in a different experimental setup by training the classifier solely on the training data from the RT experiment and applying it to this unseen data set.

Application to Motor Imagery BCI

To investigate the possibility of removing relevant neural activity, we incorporated our automatic ICA-classification step in a motor-imagery BCI system. In this offine analysis we investigate how an ICA-artifact reduction step affects the classification performance of a motor imagery BCI system based on the Common Spatial Patterns (CSP) method. For a detailed discussion of CSP the reader is referred to [32].

Experimental Setup

Eighty healthy BCI-novices performed first motor imagery with the left hand, right hand and both feet in a calibration (i.e. without feedback) measurement. Every 8 s one of three different visual cues (arrows pointing left, right, down) indicated to the subject which type of motor imagery to perform. Three runs with 25 trials of each motor condition were recorded. A classifier was trained using the pair of classes that provided best discrimination: CSP filters were calculated on the band-pass filtered signals and the log-variance of the spatially filtered signals were used to train a LDA. In a feedback measurement subjects could control a 1D cursor application in three runs of 100 trials [33].

Motor Imagery BCI preceded by ICA-based artifact reduction

The steps conducted to incorporate the artifact reduction are illustrated in Figure 6. The first step consists of a dimensionality reduction from about 90 EEG channels in the sensor space into k = 30 PCA components. As in the previous experiments, TDSEP was used to transform the 30 PCA components into 30 independent source components. Then, the component classifier trained on the RT experiment was applied. The components were ranked based on the classifiers output, which was used as a surrogate for the probability of being an artifact. Retaining a smaller or larger number of sources corresponds to an either very strict or soft policy for the removal of potential artifactual sources. We retained 6 to 30 source components of the most probable true neural sources, and removed the others. Further analysis was performed on the remaining sources, i.e. CSP filters were determined on the remaining independent source components and the log-variance of the spatially filtered signals were used to train an LDA.

Note that ICA artifact reduction methods usually reconstruct the EEG from the remaining neural sources. However, CSP solves an eigenvalue problem and requires the covariance matrix of the data to have full rank. Thus, CSP cannot be applied to the reconstructed EEG.

The application to the feedback measurement in a manner that allows for real-time BCI applications is straightforward: After un-mixing the original data according to the ICA filters determined on the calibration measurement, the previously determined 6 to 30 sources were selected for band-pass and CSP filtering and log-variance determination in order to form the test data features. To estimate the influence of the artifact reduction step on BCI performance, we compared the classification performance with artifact reduction (depending on the number of selected sources) with the standard CSP procedure using no artifact reduction.

Results

In the following subsections, the results of the classifier model selection and its additional validation on new data sets is presented.

Model Selection: RT study

The ranking of the features obtained by applying a LPM to the training data set of the RT study is shown in Table 1. Figure 7 shows the cross-validation errors for SVM, RLDA and LPM plotted against the size of the feature sub-set used for classification. The shape of the three curves reveals that at first, classification performance improves when adding features to the feature set. These features contain necessary but not redundant information. However, adding more than a certain number of features does not improve classification performance - these features only contain redundant information. Classification error slightly increases when more features are used for the classification task, which indicates that the classifier overfits on noisy and irrelevant features.

The fact that LPM performance is in the range of the RLDA classifier indicates that the feature ranking was suitable for our analysis (and not just for the LPM classifier). Given the ranking, the minimum number of

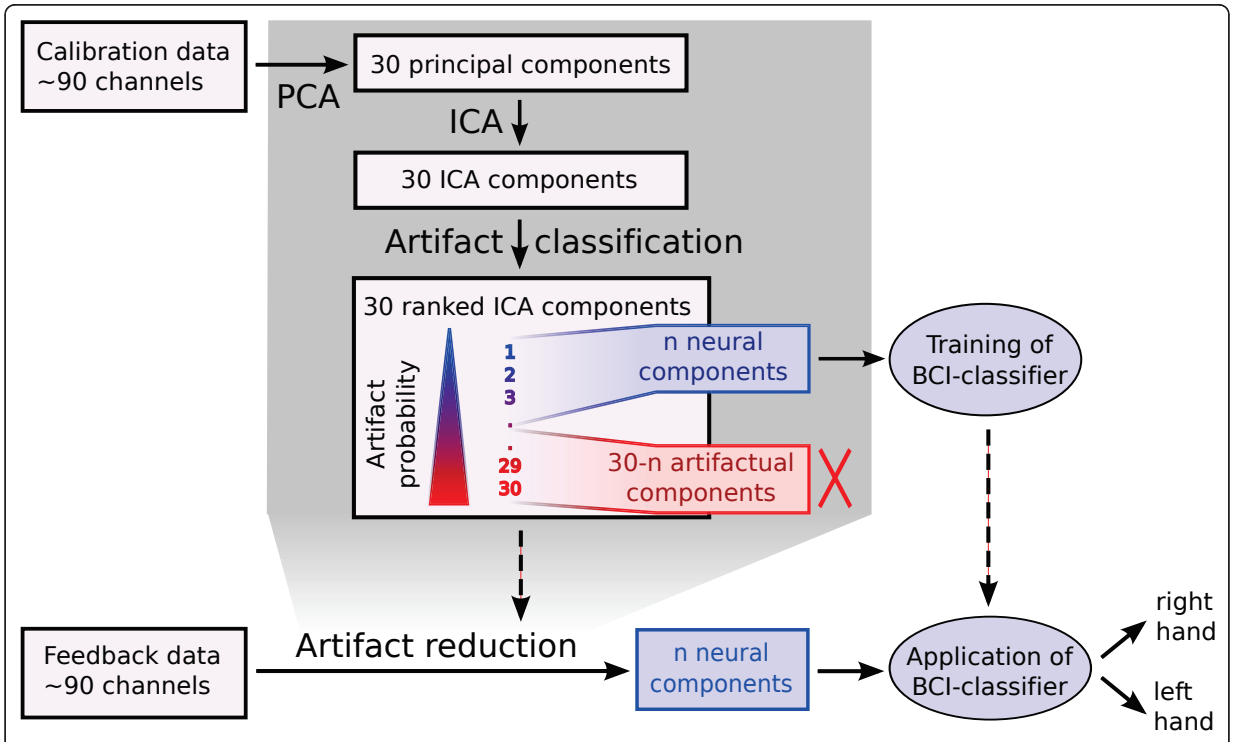


Figure 6 Artifact reduction step included in the standard CSP-procedure. The linear artifactreduction transformation of the original EEG into 6 - 30 signal components is calculated in the calibration phase. This transformation is applied to the feedback data.

features yielding a cross-validation error which deviates less than one standard error from the minimal cross-validation error is 9 for the SVM and only 6 for the RLDA. The SVM classifier slightly outperforms the RLDA classifier on the training data, but since our goal is to construct a simple linear classifier, we decided to use the RLDA classifier with the 6 best-ranked features. Notice that while SVM outperforms RLDA on the training data, this effect might be due to overfitting and disappears on the test data, as is shown in the next section.

The 6 best-ranked features are *Current Density Norm*, *Range Within Pattern*, *Mean Local Skewness 15 s*, λ, *8-13 Hz* and *FitError*. They incorporate information from the temporal, spatial and frequency domain.

Validation 1: RT study

Testing the trained classifier on unseen data from the RT study (1080 examples from experimental block 3) leads to an mean-squared error (MSE) of 8.9% only, which corresponds to a high agreement with the expert's labeling. Interestingly, testing a trained SVM classifier (based on 9 selected features) leads to an error of 9.5%. Thus, after feature selection, the RLDA classifier performs as good as a SVM classifier on unseen test data.

Let's take a moment to interpret the obtained classifier: The weight vector **w** is given in Table 2. It shows that a high current density norm of a component indicates an artifactual component. Recall from the definition of the Current Density Norm feature that these components are in fact difficult to explain by a prominent source within the brain. Furthermore, components with a high range within the pattern (i.e. outliers in the pattern), a high local skewness (i.e. outliers in the time series), high λ (i.e. a steep spectrum typical for muscle artifacts) and low spectral power in the 8-13 Hz range (i.e. no prominent alpha peak) are rated as artifacts by the classifier. Interestingly, a low FitError, i.e. a low error when approximating the spectrum by a 1/f curve, indicates an artifact for the classifier. This is due to the fact that components which have no alpha peak in the spectrum are most probably artifacts. Notice that the FitError feature in itself is not very informative, because a high FitError cannot distinguish between components with a large alpha peak (which contain most probably neural activity) and components with an unusual high spectrum in higher frequency (which indicates muscle activity). However, in combination with the other five features, the FitError feature carries additional information which improve classification performance.

It is interesting to take a closer look at the performance of single features, which is also given in Table 2. The best one, Current Density Norm, leads to a MSE of

Table 1 Ranking of features obtained by LPM.

Feature	Weight
Current Density Norm	●
Range Within Pattern	●
Mean Local Skewness 15 s	●
λ	●
8-13 Hz	●
FitError	●
Border Activation	•
2DDFT	•
Spatial Mean Activation Central	•
Max First Derivative	•
Variance	•
k_2	·
Spatial Mean Activation Left	·
Spatial Mean Activation Left Frontal	·
Laplace-Filter	·
Mean Local Variance 15 s	·
14-30 Hz	·
4-7 Hz	·
Mean Local Variance 1 s	·
Spatial Distance of Extrema	·
Spatial Mean Activation Occipital	·
k_1	·
Maximum Amplitude	·
y	·
Spatial Mean Activation Right	·
Kurtosis	·
x	·
0-3 Hz	·
Deterministic Entropy	·
Spatial Mean Activation Frontal	·
z	·
Variance of Local Variance 1 s	·
Range	·
Spatial Mean Activation Right Frontal	·
Variance of Local Variance 15 s	·
Mean Local Skewness 1 s	·
31-45 Hz	·
Shannon Entropy	·

The diameter of the black circles visualizes the absolute LPM weights $|E(w_i/\|w\|)|$ per feature after learning on the training data set. (The LPM-hyperparameter C had been set to $C = 0.1$ based on the cross-validation performance.)

14.1% on the test data of the RT study. The combination of the six features from all three domains improves the error substantially compared to even the best single feature. This shows that features which are far from optimal in single classification have a positive contribution in combination with other features.

Looking at the complete test set of 1080 components, 75 of them were misclassified as artifacts and 21 components were misclassified as neural sources. A detailed visual analysis of these cases reveals, that most of them were mixed components that contained both, artifacts and brain activity. Out of the 21 components which were misclassified as neural activity only two were eye movements and none were blinks. In some rare cases, examples which had been mislabeled by the expert could be identified. Figure 8 shows an example of a misclassified mixed component.

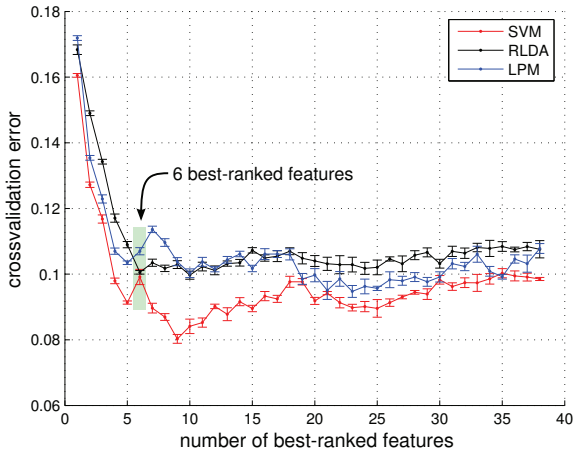


Figure 7 Cross-validation error for SVM, RLDA and LPM against the number of best-ranked features. A 10-fold cross-validation was repeated 5 times and standard errors are plotted. The SVM and LPM hyperparameters were selected by an outer cross validation. The number of 6 best-ranked features was determined for building the final classifier, as the estimated error of the RLDA starts to increase significantly for higher numbers of features.

To quantify the classification performance on muscle artifacts, we asked one expert to review the 574 artifactual components of the test set for muscle activity. The expert identified 388 components which contained muscle activity (which corresponds to 67.5% of the artifactual components and 17.2% of all the components). Out of the 21 artifactual components which were

Table 2 Feature weight vector and test errors.

Feature	Feature weight	Test Error RT	Test Error ERP
Current Density Norm	0.342	0.141	0.488
Range Within Pattern	0.574	0.151	0.186
Mean Local Skewness 15 s	0.317	0.309	0.442
λ	0.569	0.177	0.144
8-13 Hz	-0.219	0.166	0.138
FitError	-0.286	0.424	0.640
Combined		0.089	0.147

Feature weights w_i for each feature x_i of the classifier $H : \mathbb{R}^6 \ni x \alpha \text{ sign}(w^T \cdot x + b) \in \{-1, 1\}$ with $1 \triangleq$ Artifact and $-1 \triangleq$ Neuronal activity. Test error (MSE) for the 6 single features and for the combined classification for the RT experiment and the ERP experiment.

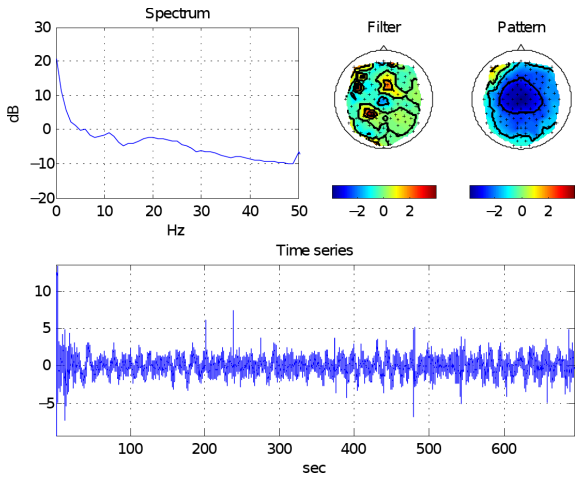


Figure 8 Example for a misclassified component. Mixed component that combines central alpha activity and slow Mayer Waves [35]. The human expert considered the mixed component a neural source, but the classifier labeld it as an artifact.

misclassified as neuronal components, only 12 contained muscle activity (57.1%). This indicates that muscle artifacts were handled equally well by the classifier as other types of artifacts.

The performance of a system on the classification task has to be judged in the light of the fact that inter-expert disagreements on EEG signals are often above 10% [34]. For our data, we asked one expert to re-label the 690 components of the training set, two years after the original labeling. The MSE between the new and the former rating was 13.2%. Thus, the prediction performance of our proposed classification method was comparable to the ranking of an human expert.

Validation 2: Auditory ERP study

The classifier trained on RT data and applied to 540 components of the auditory ERP study leads to an average MSE of 14.7% only for the classification of artifacts (expert 1: 15.7%, expert 2: 13.7%). On average over both experts, 18 of the 540 components were misclassified as artifacts and 61.5 components were misclassifed as neural sources (expert 1: 12 - 73, expert 2: 24 - 50).

Table 2 also shows the classification results for every single feature and for the combined classification for the auditory ERP data. The classification performance of the three features *Range Within Pattern*, λ and *8-13 Hz* is comparable to those in the RT experiment. They generalize very well over different experimental setups. However, the single feature classification performance for the remaining three features, *Current Density Norm*, *Mean Local Skewness 15 s*, and *FitError*, was close to chance level. This does not imply, however, that these features are unimportant for the classification tasks in the combined feature set. To asses the relevance of each feature in the combined feature set, we trained a RLDA on the ERP data using the labels of expert 1 and report the feature weights of the weight vector - *Current Density Norm* 0.139; *Range Within Pattern* 0.355; *Mean Local Skewness 15 s* 0.255; λ 0.531; *8-13 Hz* -0.710; *FitError* 0.059. We found that while the feature weight of *Current Density Norm* and *Mean Local Skewness 15 s* slightly decreased compared to the feature vector trained on the RT data, these were still far away from zero and thus carry information for the classification task.

Validation 3: Application to Motor Imagery BCI

Figure 9 (left) plots the BCI classification error (1-AUC) against the number of remaining independent

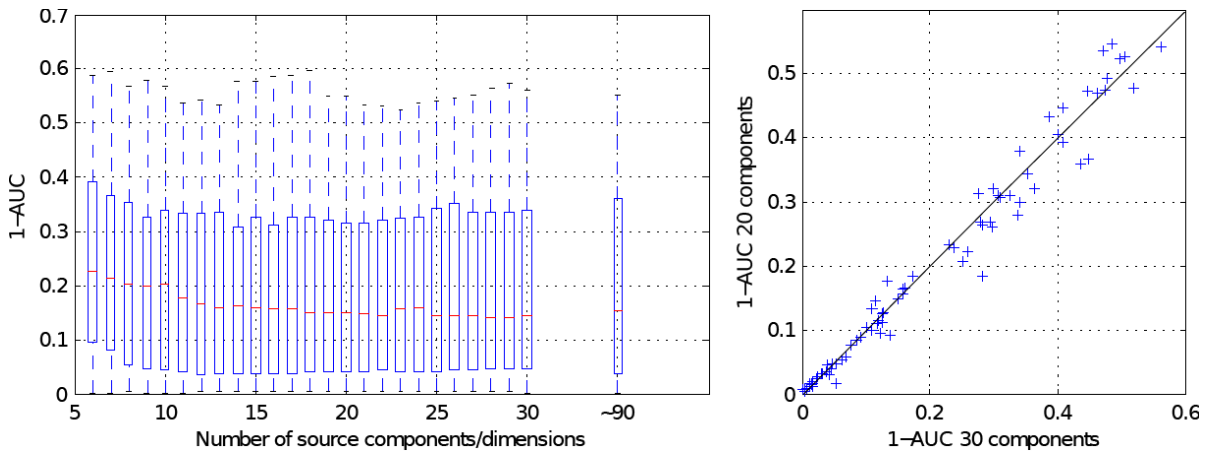


Figure 9 Influence of ICA-based artifact reduction in a motor imagery BCI tested with 80 subjects. Left: box plot of classifition errors (1-AUC) against the number of remaining independent sources compared with no artifact reduction. Right: scatter plot of classification errors (1-AUC) of each subject when removing 10 source components vs. using all 30 source components.

components, including one entry for the standard procedure without artifact reduction. Reducing the dimensionality of the data to 30 dimensions by PCA does not affect BCI performance. Moreover, consecutively removing components does not impair BCI performance at first, as these are artifactual components according to the classifier. Performance breaks down only when a strict removing policy is applied and less than about 12 sources (out of ~90 original channels) are retained, which have been ranked as neural sources by the classifier. The ranking of the classifier was confirmed by a visual analysis of the source components. Following the ranking of very probable artifacts to less probable artifacts, the inspection resulted in clear artifactual components to components that contained mixtures of neural and artifactual activity.

Figure 9 (right) shows a scatter plot of classification errors (1-AUC) for each subject when removing 10 source components vs. using all 30 source components. For this soft policy for removing artifactual components the variance between subjects is very small, especially for subjects with good classification rates.

Discussion

To summarize, we have constructed a subject-independent, fast, efficient linear component classification method that automates the process of tedious hand-selection of e.g. artifactual independent components. The proposed method is applicable online and generalizes to new subjects without re-calibration. It delivers physiologically interpretable results, generalizes well over different experimental setups and is not limited to a specific type of artifact. In particular, muscle artifacts and eye artifacts (besides other types) are recognized.

The proposed artifact classifier is based on six carefully constructed features that incorporate information from the spatial, the temporal and the spectral domain of the components and have been selected out of 38 features by a thorough feature selection procedure. After its construction on data from a reaction time experiment, the classifier's performance was validated on two different data sets: (1) on unseen data of a second condition of the original reaction time study - here the classifier achieved a classification error of 8.9%, while disagreement between two ratings of experts was 13.2%. (2) on unseen data of an auditory oddball ERP study - here the classifier showed a classification error of 14.7% in comparison to 10.6% of disagreement between experts. The classification error is remarkable low given that the second study has been recorded with half the number of electrodes, under a completely different paradigm, and contained a significantly higher proportion of artifactual components.

We could show that the generalization over different EEG studies is possible, which is in line with the findings of Mognon et al. [20] who demonstrated the generalization of an artifact classifier to a different laboratory and to a different paradigm. Although their method is simple and efficient, it so far does not recognize muscle artifacts.

Compared to the classification results of Halder et al. [22], who reported 8% of error for muscle artifacts and 1% error for eye artifacts, the classification error of our solution is slightly higher. A major difference between the two approaches is the way the training data was generated. Halder et al. reported, that subjects had specifically been instructed to produce a number of artifacts under controlled conditions for the classifier training. It can be speculated that such a training set contains stronger artifacts and less erroneous labels. Nevertheless, the results of Halder et al. were generated based on EEG recordings of only 16 electrodes. Without adjustments, it can only be applied to EEG recordings with 16 electrodes. In contrast, our method is applicable to different EEG setups. However, we only tested the generalization ability over different EEG studies on electrode sets that covered the whole scalp with approx. equidistant sensors. Whether the classifier is applicable to deal with EEG data recorded with further reduced electrode sets remains an open question that could be analyzed in the future.

To assess the danger of false positives introduced by our artifact detection method, we evaluated the influence of a strong artifact reduction on the classification performance of a standard motor imagery BCI task. An offline analysis of data acquired from 80 healthy subjects demonstrated that removing up to 60% of the sources (that were ranked according to their artifact classifier rating) did not impair the overall BCI classification performance. Note that we discarded the same number of components per subject in order to analyse the effect of false positives. In a practical BCI-system, it would probably be beneficial to apply a threshold on the propability of being an artifact per component instead.

While the suitability of our approach to remove large artifactual subspaces of the data is a welcome result, an open question remains that addresses the potential performance increase by careful artifact removal. Why didn't the removal of few artifactual sources improve the average motor imagery BCI performance? It is known that CSP is rather prone to outliers if the training data set is small [36]. Strategies to overcome this problem include the use of regularization methods for CSP (such as invariant CSP [37] or robust CSP [38]), the explicit removal of outlier trials or channels and regularization in the following classification step. As the standard evaluation procedures for motor imagery data

contained counter measures already (channel rejection and trial rejection based on variance), and the number of training data was considerably large, the overall influence of artifacts on the motor imagery data set probably was small. Furthermore, we observed, that for subjects with very good motor imagery classification rates, artifacts did not play any role at all. We conjecture that in the other subjects, artifacts either obstructed the relevant neural activity (cases where a slight improvement by artifact removal was obtained) or artifacts played some role in the control of the BCI system (cases where artifact removal slightly reduced the performance).

In addition to the construction of an efficient, sparse and interpretable classifier, our feature-selection methodology leads to valuable insights into the question of which features are best suited for the discrimination of artifactual and neuronal source components. However, it needs to be kept in mind that while the six identified features were arguably an exceptionally suitable feature set, these features were probably not the overall optimal choice. Furthermore, the question remains if the selected features generalize to other EEG data. Single feature classification performance drops on the ERP data for three of the six features (*Current Density Norm*, *Mean Local Skewness 15 s* and *Fit Error*). However, both *Current Density Norm* and *Mean Local Skewness 15 s* carry important information in the combined classification (when used together with the other four features). Still, the non-redundant information carried by the *Current Density Norm* feature drops substantially – a problem that may be caused by the use of a fixed matrix ΓJ_λ which had been determined on the RT setup of 115 electrodes. We found no obvious explanation for the importance change for the *Fit Error* feature, however.

In any case, several insights can be gained concerning the construction of a suitable feature set for the classification of artifactual components in general. First, the spatial, the temporal and the spectral domain of the components contain non-redundant information. Second, features that quantify aspects of the pattern's activity distribution, not its single values, are discriminative. Features that were ranked high in our feature selection procedure were the range within the pattern, a feature based on the simplicity of a source separation, features that analyzed the spatial frequency and a binary feature which indicates if the maximal activation is on the border of the pattern. Third, features that model the shape of the power spectrum as a $1/f$-curve as well as the absolute spectrum in the α range are discriminative. Fourth, features that quantify outliers in the time series such as kurtosis, entropy, and mean local skewness, seem to be important but redundant. We analyzed 12 such features and only one obtained a high ranking in the feature selection. Last but not least, a linear classification method seems to be sufficient when the feature set is carefully constructed.

The classification difficulties of expert raters and of proposed automatic classification methods reflect the fundamental fact that any ICA-based artifact reduction method depends crucially on the quality of the source separation into clear artifactual and neuronal source components. A good source separation method avoids mixed components that contain both, neural and artifactual activity as well as arbitrary splits of a single source into several components. In the following, both type of errors are briefly discussed.

Blind source separation is a difficult problem by itself, and various approaches have been proposed to solve it (see, e.g. [39] for a review). In the context of EEG signals, the goal is to find a source separation that minimizes the amount of mixed components. The choice of TDSEP for the pre-processing of the EEG data was motivated by the ability of the algorithm to utilize temporal structure in the data. Although this is not a unique feature of TDSEP, this approach seemed to be suitable for the processing of EEG data, which is composed of multidimensional time series signals with temporal dependencies. Moreover, research indicates that methods based on second-order statistics might outperform methods based on higher-order statistics in the removal of ocular artifacts [10,22]. Although, as Fitzgibbon stated, "the quality of the separation is highly dependent on the type of contamination, the degree of contamination, and the choice of BSS algorithm" [9], a thorough test of various ICA methods is out of the scope of this paper.

The second kind of error, the arbitrary split of sources into several components, can partially be compensated by combining the ICA with a preceding PCA step for the dimensionality reduction. This procedure has the additional advantage of removing noise in the data. We chose to project the original data into a 30 dimensional space. The value of 30 was based on rough experience and on a quick visual inspection of the data, and was probably not the optimal choice. An improvement of the quality of separation might be possible by optimizing the dimensionality reduction, but the effort was not undertaken here. Future work is needed to analyze the influence of dimensionality reduction on source separation.

To conclude, we hope that the source component classification method presented in this study delivers a substantial contribution for the BCI community and the EEG community in general, as a reliable and practical tool for the removal of artifacts. To support the community, to encourage the reproduction of our results, or allow for re-labeling of data we provide the readily trained classifier, an implementation of the feature

extraction routines together with example scripts, the extracted features of the RT data, a visualization of 1770 components together with the expert labels used for the classifier training, and a visualization of components misclassified by our method (see Additional File 1 - MatlabCode; Additional File 2 - TrainComponents; Additional File 3 - TestComponents; Additional File 4 - Misclassifications).

Additional material

Additional file 1: MatlabCode. A Matlab implementation of the feature extraction routines together with example scripts, the readily trained classifier and the extracted features for all the components of the RT data set.

Additional file 2: TrainComponents. Visualization of the 690 independent components in the training RT data, together with the expert's labels.

Additional file 3: TestComponents. Visualization of the 1080 independent components in the RT test data, together with the expert's labels.

Additional file 4: Misclassifications. Visualization of the 75 + 21 misclassified components of the RT test data

Acknowledgements
The authors thank Martijn Schreuder for his help with recording and preparing the auditory ERP data set, Claudia Sannelli for her help with recording and preparing the RT EEG data set, the authors of [33] for providing the motor imagery data set.
Funding by the *European Community* under the PASCAL Network of Excellence (IST-2002-506778) and under the FP7 Programme (TOBI ICT-2007-224631 and ICT-216886), by the *Bundesministerium für Bildung und Forschung (BMBF)* (FKZ 01IBE01A, FKZ 16SV2231, 01GQ0850 and 01IB001A) and by the *Deutsche Forschungsgemeinschaft (DFG)* (VitalBCI MU 987/3-1) is gratefully acknowledged. (This publication only reflects the authors' views. Funding agencies are not liable for any use that may be made of the information contained herein.) Last but not least, we would like to thank our reviewers for their valuable comments.

Authors' contributions
The EEG studies were performed in cooperation with colleagues mentioned in the acknowledgments section. Authors MT and SH designed and carried out the RT EEG study. MT designed and carried out the ERP study. IW and MT designed the feature extraction and feature selection algorithms and the artifact classification method. SH provided the current density norm feature method and implementation. All other implementations were carried out by IW. IW and MT analyzed and evaluated the overall methodology and wrote the manuscript. All authors proof-read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 29 April 2011 Accepted: 2 August 2011
Published: 2 August 2011

References
1. del R, Millán J, Rupp R, Mueller-Putz G, Murray-Smith R, Giugliemma C, Tangermann M, Vidaurre C, Cincotti F, Kübler A, Leeb R, Neuper C, Müller KR, Mattia D: **Combining Brain-Computer Interfaces and Assistive Technologies: State-of-the-Art and Challenges.** *Frontiers in Neuroprosthetics* 2010, **4.**
2. Müller KR, Tangermann M, Dornhege G, Krauledat M, Curio G, Blankertz B: **Machine Learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring.** *Journal of neuroscience methods* 2008, **167**:82-90.
3. Iwasaki M, Kellinghaus C, Alexopoulos AV, Burgess RC, Kumar AN, Han YH, Lüders HO, Leigh RJ: **Effects of eyelid closure, blinks, and eye movements on the electroencephalogram.** *Clinical Neurophysiology* 2005, **116(4)**:878-885.
4. Goncharova II, McFarland DJ, Vaughan TM, Wolpaw JR: **EMG contamination of EEG: spectral and topographical characteristics.** *Clinical Neurophysiology* 2003, **114**:1580-1593.
5. Fatourechi M, Bashashati A, Ward RK, EBirch G: **EMG and EOG artifacts in brain computer interface systems: A survey.** *Clinical Neurophysiology* 2007, **118**:480-494.
6. Croft RJ, Barry RJ: **Removal of ocular artifact from the EEG: a review.** *Clinical Neurophysiology* 2000, **30**:5-19.
7. Makeig S, Bell AJ, Jung TP, Sejnowski TJ: **Independent Component Analysis of Electroencephalographic Data.** *Advances in neural information processing systems* 1996, **8**:145-151.
8. Jung TP, Makeig S, Humphries C, Lee TW, Mckeown MJ, Iragui V, Sejnowski TJ: **Removing electroencephalographic artifacts by blind source separation.** *Psychophysiology* 2000, **37**:163-178.
9. Fitzgibbon SP, Powers DMW, Pope KJ, Clark CR: **Removal of EEG Noise and Artifact Using Blind Source Separation.** *Clinical Neurophysiology* 2007, **24(3)**:232-243.
10. Romero S, Mañanas MA, Barbanoj MJ: **A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: A simulation case.** *Computers in Biology and Medicine* 2008, **38**:348-360.
11. Crespo-Garcia M, Atienza M, LCantero J: **Muscle artifact remval from human sleep EEG by using independent component analysis.** *Ann Biomed Eng* 2008, **36**:467-475.
12. McMenamin BW, Shackman AJ, Maxwell JS, Bachhuber DRW, Koppenhaver AM, Greischar LL, Davidson RJ: **Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG.** *NeuroImage* 2010, **49**:2416-2432.
13. Barbati G, Porcaro C, Zappasodi F, Rossini PM, Tecchio F: **Optimization of an independent component analysis approach for artifact identifaction and removal in magnetoencephalographic signals.** *Clinical Neurophysiology* 2004, **115**:1220-1232.
14. Delorme A, Makeig S, Sejnowski T: **Automatic artifact rejection for EEG data using high-order statistics and independent component analysis.** *Proceedings of third international independent component analysis and blind source decomposition conference, San Diego, CA* 2001, 457-462.
15. Romero S, Mañanas M, Riba J, Giménez S, Clos S, Barbanoj M: **Evaluation of an automatic ocular filtering method for awake spontaneos EEG signals based on Independt Component Analysis.** *26th Annual International Conference of the Engineering in Medicine and Biology Society (EMBS)* 2004, 925-928.
16. Shoker L, Sanei S, Chambers J: **Artifact Removal from Electroencephalograms using a hybrid BSS-SVM algorithm.** *IEEE Signal Processing Letters* 2005, **12(10)**:721-724.
17. James CJ, Gibson OJ: **Temporally constrained ICA: an application to artifact rejection in electromagnetic brain signal analysis.** *IEEE Trans Biomed Eng* 2003, **50**:1108-1116.
18. Joyce CA, Gorodnitsky IF, Kutas M: **Automatic removal of eye movement and blink artifacts from EEG data using blind component separation.** *Psychophysiology* 2004, **41**:331-325.
19. Viola FC, Thorne J, Edmonds B, Schneider T, Eichele T, Debener S: **Semi-automatic identification of independent components representing EEG artifact.** *Clinical Neurophysiology* 2009, **120**:868-877.
20. Mognon A, Jovicich J, Bruzzone L, Buiatti M: **ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features.** *Psychophysiology* 2010, 1-12.
21. LeVan P, Urrestarazu E, Gotman J: **A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification.** *Clinical Neurophysiology* 2006, **117**:912-927.
22. Halder S, Bensch M, Mellinger J, Bogdan M, Kübler A, Birbaumer N, Rosenstiel W: **Online Artifact Removal for Brain-Computer Interfaces Using Support Vector Machines and Blind Source Separation.** *Computational Intelligence and Neuroscience* 2007, **7(3)**:1-10.
23. Ziehe A, Laskov P, Nolte G, Müller KR: **A Fast Algorithm for Joint Diagonalization with Non-orthogonal Transformations and its**

Application to Blind Source Separation. *Journal of Machine Learning Research* 2004, **5**:801-818.
24. Blankertz B, Lemm S, Treder MS, Haufe S, Müller KR: **Single-trial analysis and classification of ERP components - a tutorial.** *NeuroImage* 2011.
25. Titchener MR: **T-entropy of EEG/EOG Sensitive to Sleep State.** *International Symposium on Nonlinear Theory and Applications (NOLTA)* 2006, 859-862.
26. Hämäläinen MS, Ilmoniemi R: **Interpreting magnetic fields of the brain: minimum-norm estimates.** *Med Biol Eng Comput* 1994, **32**:35-42.
27. Haufe S, Nikulin VV, Ziehe A, Müller KR, Nolte G: **Combining sparsity and rotational invariance in EEG/MEG source reconstruction.** *NeuroImage* 2008, **42**:726-738.
28. Bennett KP, Mangasarian OL: **Robust linear programming discrimination of two linearly inseparable sets.** *Optimizaion Methods and Software* 1994.
29. Chang CC, Lin CJ: *LIBSVM. a library for support vector machines* 2001.
30. Ledoit O, Wolf M: **A well-conditioned estimator for large-dimensional covariance matrices.** *Journal of Multivariate Analysis* 2004, **88(2)**:365-411.
31. Schäfer J, Strimmer K: **A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.** *Statistical Applications in Genetics and Molecular Biology* 2005, **4(32)**.
32. Blankertz B, Tomioka R, Lemm S, Kawanabe M, Müller KR: **Optimizing spatial filters for robust EEG single-trial analysis.** *IEEE Signal Proc Magazine* 2008, **25**:41-56.
33. Blankertz B, Sannelli C, Halder S, Hammer EM, Kübler A, Müller KR, Curio G, Dickhaus T: **Neurophysiological predictor of SMR-based BCI performance.** *NeuroImage* 2010, **51(4)**:1303-1309.
34. Klekowicz H, Malinow U, Niemcewi S, Wakarow A, Wolynczyk-Gmaj D, Piotrowski T, Durka P: **Automatic analysis of sleep EEG.** *Frontiers in Neuroinformatics. Conference Abstract. Neuroinformatics* 2008.
35. Lugaresi E, Coccagna G, Mantovani M, Lebrun R: **Some periodic phenomena arising during drowsiness and sleeping in humans.** *Electroenceph clin Neurophysiol* 1972, **32**:701-705.
36. Krauledat M, Dornhege G, Blankertz B, Müller KR: **Robustifying EEG data analysis by removing outliers.** *Chaos and Complexity Letters* 2007, **2(3)**:259-274.
37. Blankertz B, Kawanabe M, Tomioka R, Hohlefeld F, Nikulin V, Müller KR: **Invariant Common Spatial Patterns: Alleviating Nonstationarities in Brain-Computer Interfacing.** In *Advances in Neural Information Processing Systems 20.* Edited by: Platt J, Koller D, Singer Y, Roweis S. Cambridge, MA: MIT Press; 2008:113-120.
38. Kawanabe M, Vidaurre C, Scholler S, Blankertz B, Müller KR: **Robust Common Spatial Filters with a Maxmin Approach.** *EMBS-Conference* 2009, 2470-2473.
39. Choi S, Cichocki A, Park HM, Lee SY: **Blind Source Separation and Independent Component Analysis: A Review.** *Neural Information Processing - Letters and Reviews* 2005, **6**:1-57.

doi:10.1186/1744-9081-7-30
Cite this article as: Winkler *et al.*: Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals. *Behavioral and Brain Functions* 2011 **7**:30.

3.2 Robust artifactual independent component classification for the BCI practitioner

Irene Winkler, Stephanie Brandl, Franziska Horn, Eric Waldburger, Carsten Allefeld and Michael Tangermann. Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering*, 11(3):035013, 2014. `http://iopscience.iop.org/1741-2552/11/3/035013/`

Short summary. This paper presents a number of changes to MARA that make it more useful for practitioners. First, we present a simple strategy to make sure that MARA generalizes better to reduced electrode setups: we re-train the classifier using the desired electrode setup. Second, we validate the method on more data sets, notably a data set of 4473 labeled ICA components which was provided by Carsten Allefeld from the neuroimaging lab of Prof. Haynes. Third, we investigate the effect of artifact removal on single-trial BCI classification from 101 users and 3 paradigms. It turns out that ICA artifact cleaning has little influence on average BCI performance when analyzed by state-of-the-art BCI methods. Last but not least, we implement MARA as an EEGLAB plug-in.

Contributions. I wrote the majority of the paper and carried out the majority of the analysis. The EEGLAB plug-in was jointly developed by Eric Waldburger and I. Stephanie Brandl analyzed the CNT data set, under the supervision of Carsten Allefeld and I. Most of the LRP-MI-BCI analysis was carried out by Franziska Horn.

OPEN ACCESS
IOP Publishing
Journal of Neural Engineering
J. Neural Eng. **11** (2014) 035013 (10pp)
doi:10.1088/1741-2560/11/3/035013

Robust artifactual independent component classification for BCI practitioners

Irene Winkler[1], **Stephanie Brandl**[2], **Franziska Horn**[1], **Eric Waldburger**[1], **Carsten Allefeld**[2] **and Michael Tangermann**[3]

[1] Machine Learning Laboratory, Technical University of Berlin, Marchstr. 23, D-10587 Berlin, Germany
[2] Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, Philippstr. 13, Haus 6, D-10115 Berlin, Germany
[3] BrainLinks-BrainTools Excellence Cluster, University of Freiburg, Albertstr. 23, D-79104 Freiburg, Germany

E-mail: irene.winkler@tu-berlin.de

Received 30 August 2013, revised 11 November 2013
Accepted for publication 15 November 2013
Published 19 May 2014

Abstract

Objective. EEG artifacts of non-neural origin can be separated from neural signals by independent component analysis (ICA). It is unclear (1) how robustly recently proposed artifact classifiers transfer to novel users, novel paradigms or changed electrode setups, and (2) how artifact cleaning by a machine learning classifier impacts the performance of brain–computer interfaces (BCIs). *Approach.* Addressing (1), the robustness of different strategies with respect to the transfer between paradigms and electrode setups of a recently proposed classifier is investigated on offline data from 35 users and 3 EEG paradigms, which contain 6303 expert-labeled components from two ICA and preprocessing variants. Addressing (2), the effect of artifact removal on single-trial BCI classification is estimated on BCI trials from 101 users and 3 paradigms. *Main results.* We show that (1) the proposed artifact classifier generalizes to completely different EEG paradigms. To obtain similar results under massively reduced electrode setups, a proposed novel strategy improves artifact classification. Addressing (2), ICA artifact cleaning has little influence on average BCI performance when analyzed by state-of-the-art BCI methods. When slow motor-related features are exploited, performance varies strongly between individuals, as artifacts may obstruct relevant neural activity or are inadvertently used for BCI control. *Significance.* Robustness of the proposed strategies can be reproduced by EEG practitioners as the method is made available as an EEGLAB plug-in.

Keywords: EEG, artifact removal, independent component analysis (ICA), blind source separation (BSS), brain–computer interface (BCI)

(Some figures may appear in colour only in the online journal)

1. Introduction

Artifacts are omnipresent in recordings of the electroencephalogram (EEG) and other brain signals. For neuroscientific or clinical purposes the interpretation of EEG signals depends on relatively clean recordings. Thus, artifact avoidance during measurement and post-hoc artifact removal are important steps to enhance the signal-to-noise ratio (SNR) before scientific interpretation of the data. While task-independent artifacts may mask an existing effect, artifacts systematically locked to an experimental task are even more problematic: they may lead to misinterpretation of the data and spurious results.

The field of the brain–computer interface (BCI) not only makes use of offline analyses, but strives to interpret mental states on a single-trial basis in real-time and in closed-loop scenarios [1]. BCI research is especially sensitive to

task-locked artifacts, as the decoding of a user's intent by a BCI system should not rely on task-related non-neural signals. This requirement is most important when conducting research with healthy study participants on a novel paradigm or analysis method which should be transferable to severely motor-impaired patients, because they may not be physically capable of producing those artifacts [2–4]. Understandably, the role of artifacts is thus scrutinized during peer-reviewed publication processes.

The exclusive use of brain signals in BCI must typically be dropped when it comes to practical tests with end-users in need, as hybrid BCI approaches [5, 6] provide a richer and more reliable control than pure BCIs. Additionally, interest in novel types of studies is growing amongst EEG researchers. Such studies include users (inter-)acting in space [7–9] like in collaborative and social paradigms (for a review see [10]), the interaction between users and machines [11] and the non-medical use of BCI methods [12, 13].

From an EEG practitioner's point of view, a fully automatic algorithmic solution for the treatment of artifacts is desirable. It would put him or her in control of artifacts and enable him or her to either remove them or check their influence. Ideally, this would be realized by a global classifier which could be trained once and then reliably separates multiple types of artifactual components from neural components. The classifier should work robustly across data from different users and across domains. The latter includes changing experimental paradigms and tasks, different preprocessing methods and varying EEG electrode setups. It should do so without any need of re-training, and it should not require separate artifact recordings before it can be applied to novel scenarios.

1.1. State-of-the-art IC artifact classification

For an extensive review of artifact reduction techniques in the context of BCI-systems, we refer the reader to [14]. In our work, we concentrate on a class of popular artifact rejection approaches, which decompose the original EEG into independent source components (ICs) using independent component analysis (ICA). This method exploits the assumption that artifactual signal components and neural activity are generated independently. Artifactual ICs are hand-selected and then discarded. The remaining neural components are used to reconstruct the EEG [15, 16].

While assumptions for the application of ICA methods are only approximately met in practice (no systematic co-activation of artifactual and neural activity, linear mixture of independent components (ICs), stationarity of the sources and the mixture, prior knowledge about the number of components), their application usually leads to a good, albeit not perfect separation for common artifacts such as blinks, eye movements or scalp muscles [17–20]. ICA has successfully been applied to the removal of cochlea implant artifacts [21]. However, gait-related artifacts are reported to remain in most of the ICs in EEG recorded during mobile activities [9, 22].

Because a thorough analysis of the achievable separation performance is out of the scope of this paper, we refer the reader to [17, 23, 24] on the question of which ICA variants are well-suited for artifact rejection. Instead, we focus on practical tools which avoid the time-consuming hand-rating process of ICs by classifying ICs with the help of machine learning methods into artifactual and non-artifactual components. Most approaches concentrate on eye artifacts [25–31], but automatic classification has also been successful for heart-beat artifacts [28, 31], generic discontinuities [29], muscle artifacts [31–34] and even very specialized artifacts such as cochlear implants [21]. As most of these methods have a supervised basis, to some degree they reflect the specific conditions of the training set. The EEG practitioner is now faced with the question of how well supervised methods generalize to his or her data acquired under novel experimental conditions with different preprocessing.

Unsupervised methods successfully circumvent this problem for example by reverting to automatic thresholding strategies [29]. However, these methods are often limited to the use of one or two features and detect only certain types of artifacts. It is unclear how to extend them to more complex artifacts with a varying physiological fingerprint, such as muscle artifacts. For supervised or template-based approaches, first studies suggest that generalization to novel paradigms is possible [28, 30, 31, 34]; however, efforts have concentrated on eye artifacts [28, 30].

1.2. Robustness under novel paradigms and electrode setups

In this paper, we take a step forward by analyzing the generalization ability of a state-of-the-art supervised IC classification algorithm which we have recently proposed [34]. It is not restricted to the classification of eye or muscle artifacts, but is equally well suited to detect other artifacts such as loose electrodes. By comparing three strategies, we investigate this multi-artifact classifier wrt. new electrode setups and paradigms. We ask the following questions: How does a change of the electrode setup impact the IC classification performance? Is it necessary to hand-label components of the new data set and retrain the classifier based on those? How strong is the deterioration of IC classification performance without re-training? We investigate these questions for three data sets of 6303 labeled ICs from 35 participants in 3 experimental studies: a reaction time (RT) task embedded in a simulated-driving task, an auditory event-related potential study (ERP-BCI) and a study analyzing continuous EEG data (CNT) of subjects instructed to listen to short stories.

1.3. Effect on BCI performance

After having demonstrated the robustness properties of the IC classification, we are interested in the effects of automatic ICA artifact cleaning on the classification of EEG trials in BCI systems. As a first proof-of-concept, Halder *et al* [33] applied artifact cleaning to data from three participants who performed motor imagery. Depending on whether artifacts were systematically co-activated with the task or not, opposite effects of artifact cleaning on BCI classification performance were demonstrated. To the best of our knowledge, only small

data sets of one or two participants have been analyzed since then [35, 36].

To fill this gap, we extend our analysis from [34] by investigating the overall effect of ICA artifact cleaning on BCI performance to data of 101 participants wrt. 3 BCI paradigms: auditory event-related potentials, event-related (de-)synchronization and slow motor-related potentials due to motor imagery tasks.

1.4. Software for the EEG practitioner

Last but not least, we make our IC classification software available as an EEGLAB plug-in 'MARA' (Multiple Artifact Rejection Algorithm). EEGLAB [37] is a popular, Matlab-based open-source tool and used by a growing community of EEG researchers. As existing ICA-based plug-ins primarily focus on the detection of eye artifacts [27–29], we hope this will deliver a substantial contribution to the community by assisting EEG practitioners with the rejection of multiple type of artifacts.

2. Methods and materials

2.1. Processing chain for ICA artifact rejection

The typical process chain for artifact rejection with ICA consists of the following steps: first, a rough pre-cleaning of the data by channel rejection and trial rejection based on variance criteria may be performed. Second, a dimensionality reduction may help to avoid an unnatural splitting of (neural) sources. Unfortunately, the optimal number of components to extract remains unknown and has to be determined either by visual inspection or by a heuristic, such as retaining 99% of the explained variance or a fixed number of components. Third, ICA methods decompose the observed EEG data x into unknown source components s assumed to be mutually independent and following the generative linear model $x = A \cdot s$. Finally, artifactual source components are identified which allows the EEG signals to be reconstructed without them.

In manual classification of ICs, experts ratings are based on a component's time series, its power spectrum and spatial pattern (given by the respective column of A). Unfortunately, ICA frequently results in mixed components containing aspects of both neural and artifactual activity which cannot be rated unambiguously [38]. Consequently, such mixed components tend to be either retained or rejected depending on the specific application. The subjective nature of such expert decisions is reflected by the fact that experts disagree with each other as well as with themselves over time [39]. Nevertheless, the reliability of component classification is often not reported, and if it is, researchers use one of many metrics of inter-rater reliability statistics which are difficult to compare directly (e.g. *Krippendorff's alpha* in [20], *inter-class correlation coefficient* in [40], *degree of association phi* in [28], *mean-squared error (MSE)* or *average agreement* in [34, 39]).

Automatic classification of ICs based on Machine Learning methods offers a well-described algorithm which rates consistently over time. However, this algorithm, too, is of subjective nature in the sense that it is optimized to predict labels similar to those labeling strategies applied by human raters. The performance of the algorithm thus crucially depends on the quality of the training set and its labels. For all our IC data sets, experts were instructed to identify components which are predominantly driven by artifacts.

In this paper, automatic IC classification is realized by a linear pre-trained classifier. It is based on the following six features which were determined in a feature selection procedure described in [34]. One feature aims to detect outliers in the time series of an IC, three features are extracted from the spectrum, and two features extract information from the scalp pattern of an IC—the latter depending directly on the electrode layout.

(i) *Current density norm*. ICA itself does not provide information about the locations of the sources s. However, ICA patterns can be interpreted as EEG potentials for which the location of the sources can be estimated. We considered 2142 locations arranged in a 1 cm spaced 3D-grid, formulated the forward problem according to [41–43] and sought the source distribution with minimal l_2-norm (i.e. the 'simplest' solution) [44, 45]. Since this source distribution can model cerebral sources only, it is natural that artifactual signals originating outside the brain can only be modeled by rather complicated sources. Those are characterized by a large l_2-norm, which we use as a feature.

(ii) *Range within pattern*. The logarithm of the difference between the minimal and the maximal activation in a pattern.

(iii) *Mean local skewness*. The mean absolute local skewness of time intervals of 15 s duration. This feature aims to detect outliers in the time series.

(iv) $\boldsymbol{\lambda}$ *and fit error*. These two features describe the deviation of a component's spectrum from a prototypical $1/f$ curve and its shape. The parameters $k_1, \lambda, k_2 > 0$ of the curve

$$f \mapsto \frac{k_1}{f^{\lambda}} - k_2 \tag{1}$$

are determined by six points of the log spectrum: (1) the log power at 2 Hz, (2) the log power at 3 Hz, (3) the point of the local minimum in the band 5–13 Hz, (4) the point 1 Hz below the third point of support, (5) the point of the local minimum in the band 33–39 Hz, and (6) the point 1 Hz below the fifth point of support. Finally, the logarithm of λ and of the MSE of the approximation of f to the real spectrum in the 8–15 Hz range are used as features for the classifier.

(v) 8–13 *Hz*. The average log band power of the α band (8–13 Hz).

2.2. Data sets and experimental paradigms

Data sets of four experimental EEG paradigms (named RT, CNT, MI-BCI, ERP-BCI) were available for this study. For three of them, RT, CNT and ERP-BCI, expert-labeled ICs

(artifacts versus neural sources) were available. Two data sets (MI-BCI, ERP-BCI) stem from BCI experiments. As the trial-wise BCI tasks are known, the estimated single-trial BCI-classification performance provides a metric for the influence of a preceding artifact treatment.

RT. For this data set, labeled ICs were available. In a simulated-driving study, participants performed a forced-choice left or right key press RT task upon two auditory stimuli in an oddball paradigm [34]. EEG data was recorded from 121 approx. equidistant sensors and high-noise channels were rejected based on a variance criterion. We selected 43 runs of 10 min duration from eight participants that had 104 electrodes in common. Prior to the IC computation via TDSEP [46], a 2 Hz high-pass filter was applied, and dimensionality was reduced to 30 PCA components. Two experts hand-labeled the resulting 30 ICs per run into artifactual and neural components (1290 labeled ICs altogether).

Of these, 840 ICs (28 runs from 5 participants) were used to train a linear classifier C_{RT} to discriminate artifactual from neural components. Another 450 ICs (15 runs from 3 remaining subjects) were available for estimating the generalization performance of C_{RT}. The training set contained 52% of artifactual ICs, the test set contained 59%.

CNT. For this data set, labeled ICs were available. Nine participants continuously listened to audio–visual stories during short runs of an average duration of 3.77 min [40]. The resulting 71 recordings contained 62 EEG channels plus one EOG channel. The recording of each run was appended with a short eyes-closed and eyes-open recording and high-pass filtered at 0.16 Hz. No dimensionality reduction was applied, before ICs were estimated by FastICA [47] on the full set of electrodes. This decomposition yielded $63 \times 71 = 4473$ components, which were hand-rated by three experts into 47% artifactual and 53% neural source components.

ERP-BCI. For this data set, labeled ICs as well as labeled BCI-trials were available. In a spatial auditory BCI study which made use of auditory event-related potentials, participants underwent a calibration run of approx. 30 min duration and an online spelling run [48]. In the online run, subjects were asked to write a sentence while auditory and visual feedback was provided. EEG was recorded from 61 electrodes while the participants listened to a rapid sequence of 6 auditory stimuli and were instructed to silently count the number of appearances of a rare target tone.

For the classification of artifacts, data of 18 participants was analyzed. Their EEG signals were band-pass filtered between 0.1 and 40 Hz and the dimensionality was reduced to 30 PCA channels. Subsequently 30 ICs were computed per run using TDSEP. The resulting 540 source components were hand-labeled into 72% artifactual and 31% neural source components.

To assess the influence of artifact correction onto the BCI classification performance, data of the 21 BCI novices participating in the first session of the auditory ERP speller study of Schreuder *et al* [48] was re-analyzed. Their calibration measurement is used to train a shrinkage regularized linear classifier based on spatio-temporal ERP features [48, 49]. BCI performance evaluations are based on the re-analyzed online data of these participants.

MI-BCI. For this data set, labeled BCI-trials were available, but no labeled ICs. This data set was recorded with 119 EEG channels from 80 healthy BCI novices, who first performed motor imagery tasks (left hand, right hand and both feet) in a calibration run (i.e. without feedback). Every 8 s, the requested BCI task of the current trial was indicated by a visual cue. A CSP-based BCI-classifier (see below) was trained on the labeled calibration trials using the pair of classes which provided best discrimination. During the three online runs of 100 trials each participant controlled an application which provided continuous visual feedback in the form of a horizontally moving cursor [50].

Motor imagery data can be exploited by two different types of EEG features.

(i) CSP-MI-BCI: the most common strategy makes use of oscillatory features which describe event-related (de)-synchronization (ERD/ERS) in the alpha- and beta band of the EEG. After enhancing the SNR of these effects by individual data-driven spatial filters, which are derived by the common spatial patterns (CSP) analysis [51], CSP-features can be classified by a shrinkage-regularized linear classifier.

(ii) LRP-MI-BCI: the second strategy is based on slow motor-related potentials (e.g. the lateralized readiness potential (LRP)). Different classes of imagined movements are distinguished with an ERP-type analysis [49, 52]: EEG is band-pass filtered between 4 and 8 Hz, before a small number of class-discriminative intervals is determined on the calibration data. The average activity per interval and channel is used as features for a binary shrinkage-regularized linear classifier.

While the original online runs were performed with the CSP-MI-BCI classifier, without artifact rejection, the offline re-analysis makes use of both types of features in order to assess the influence of a preceding artifact removal.

2.3. Robustness under novel paradigms and electrode setups

For the classification of artifactual IC components, three classification strategies—*fixed*, *adapted* and *study-specific*—were compared on the ERP-BCI and the CNT data set. Figure 1 visualizes the strategies. In the *fixed* scenario, classifier C_{RT} is trained once on features of labeled ICs of the RT data set, and furthermore applied to ICs of any other data set. Neither hand-labeling of novel ICs nor re-calculation of features or any re-training of the classifier is necessary in this simplest scenario. While hand-labeling of novel ICs is also avoided successfully in the *adapted* strategy, a channel adaptation on the RT-data is performed by cutting the training patterns to the specific electrode layout of the test data set. Features then need to be re-calculated based on the reduced patterns and a

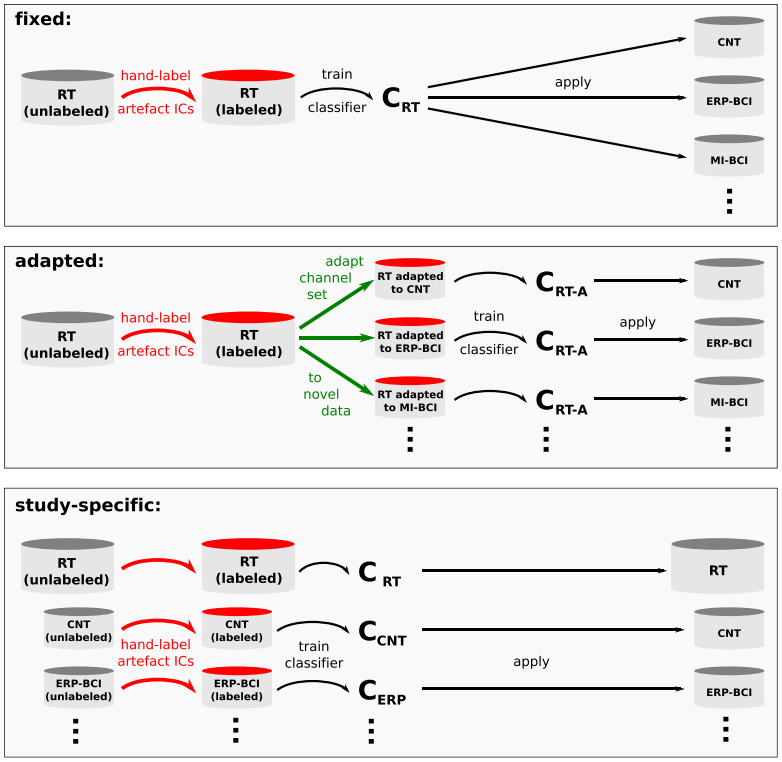


Figure 1. Schematic plot of the three transfer strategies *fixed*, *adapted* and *study-specific*. Expensive hand-labeling steps of ICs are marked with red arrows, cheap channel reduction and classifier training steps in green and black. Note that any self-application of classifiers in the *study-specific* strategy was performed exclusively in a leave-one-subject-out validation scenario.

re-training yields the *adapted* classifier $C_{\mathrm{RT-A}}$. All steps can be performed automatically and do not require user input. The third strategy, *study-specific*, requires the effort of experts every time a novel study is performed. The ICs of at least some subjects need to be hand-labeled, before a study-specific classifier (e.g. C_{CNT} or C_{ERP}) can be trained and applied to novel subjects. It's performance was evaluated by leave-one-subject-out cross-validation.

To explore the robustness of the artifact classifier against reduced EEG channel sets, we compared the *fixed* IC-classifier C_{RT} with the *adapted* IC-classifier $C_{\mathrm{RT-A}}$ on the RT and ERP-BCI test data sets with reduced setups (varying from 16 to 104 resp. 61 EEG channels). All electrode setups were approximately equidistant and covered the whole scalp.

2.4. Effect on BCI performance

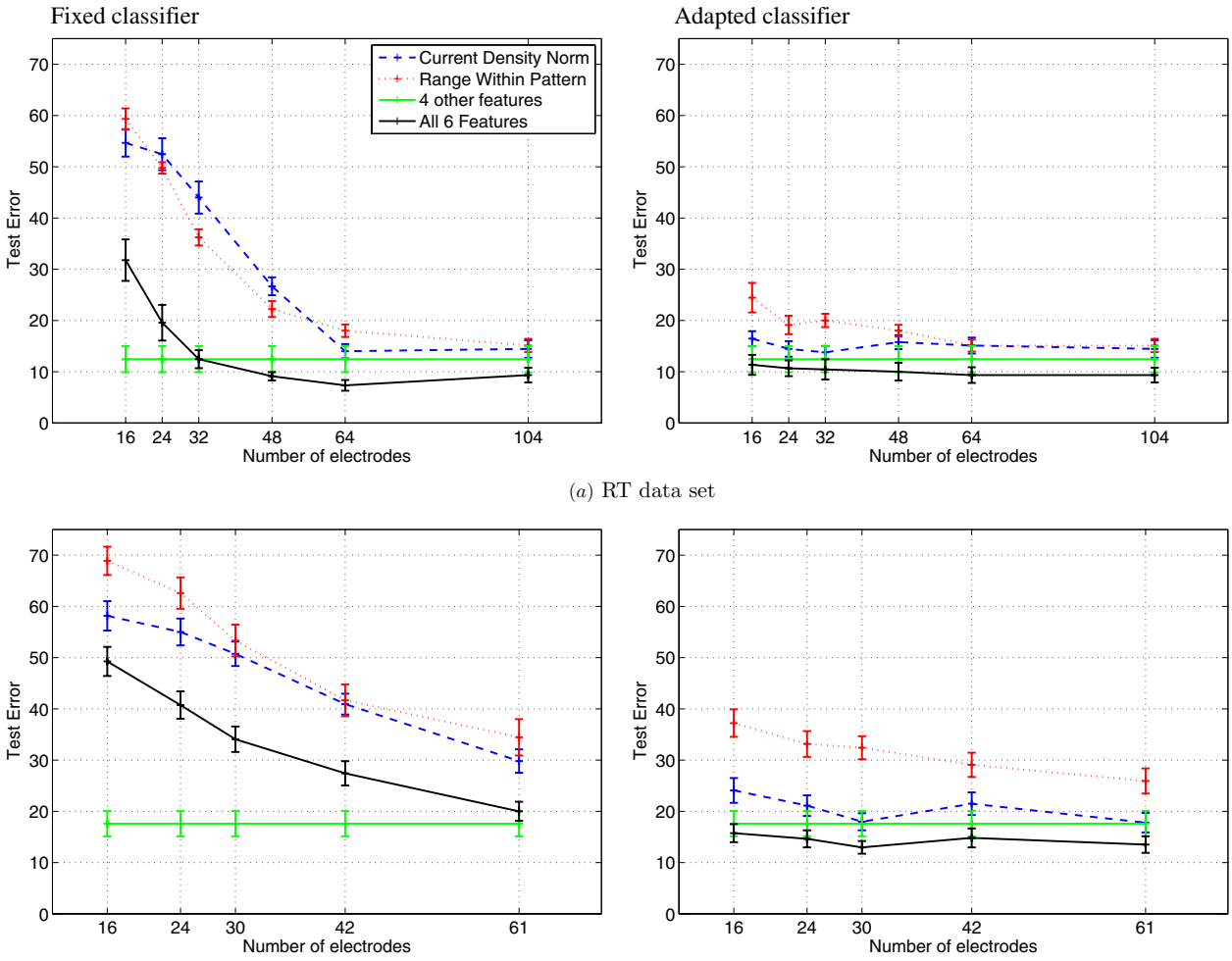
This offline re-analysis of three BCI paradigms described in section 2.2 compares standard BCI performance with and without a preceding ICA artifact cleaning. In both cases, artifactual channel and trial rejection based on a variance criterion was performed prior to BCI training. Training of the BCI-classifiers is based on the calibration runs only, and BCI performance tests are performed with the online runs of the participants.

ICA artifact cleaning is included in a manner that allows for real-time BCI applications. Prior to TDSEP, we estimated whether a PCA pre-processing to 99% explained variance would be useful via cross-validation on the calibration data. This was the case only for the LRP-MI paradigm. IC components were then derived by TDSEP and classified with the *adapted* classifier $C_{\mathrm{RT-A}}$ on the calibration data. The BCI is set up on the remaining ICs. On the online runs, un-mixing and component rejection is performed according to the de-mixing determined on the calibration data. The BCI classifier is applied to features extracted from the remaining components of the online runs.

3. Results

3.1. Robustness under novel electrode setups

Figure 2 shows the classification error for the *fixed* classifier C_{RT} and the *adapted* classifier $C_{\mathrm{RT-A}}$ for different channel setups on both the RT and the ERP-BCI test sets. On the RT test data with the full 104 channel setup, a classifier using all six features achieves a MSE of 9.3% only, which slightly outperforms the use of only four pattern-independent features (12.4% MSE). While C_{RT} generalizes robustly over the range of 104 to 48 electrodes in the RT test sets, its error increases up to 31.8% for the smallest set of 16 electrodes. On the ERP-BCI data set, the use of only four pattern-independent features is already outperforming the fixed classifier C_{RT} on the full 61 electrode setup. Classification performance of C_{RT} then breaks down to 50% on the smallest set of 16 electrodes. In both the RT and the ERP-BCI data set, the drop in overall performance is due to the bad performance of both pattern-based features of over 50%.



(*a*) RT data set

(*b*) ERP-BCI data set

Figure 2. Mean classification error ± standard error estimated on (*a*) the RT and (*b*) the ERP-BCI test sets for different channel setups. The left plot shows the results for a fixed classifier, the right plot for a classifier adapted to each channel setup.

For the *adapted* strategy (i.e. re-training the classifier on the patterns cut to the specific electrode setup), the error of the pattern features (*range within pattern* and *current density norm*) was much less pronounced in both data sets. The overall error of $C_{\mathrm{RT-A}}$ for 16 electrodes remained at 11.3% on the RT data set (compared with 9.3% on 104 channels) and at 15.9% for the ERP-BCI data set (compared with 13.3% on 61 channels). In both data sets, we slightly gain from using the pattern features. On the reduced electrode setup, the classifier weight of the *range in pattern* dropped, while the weight for *current density norm* remained stable.

3.2. Robustness under novel paradigms

The results for the three proposed classification strategies on the three labeled IC data sets are summarized in table 1. The *adapted* classifier $C_{\mathrm{RT-A}}$ (trained on the RT data set cut to the specific electrode montage of the ERP-BCI or CNT data set) achieves an error of 13.3% on the ERP-BCI data and an error of 14.0% on the CNT data set.

The classification performance can be improved by a re-training on labeled data from the same study, but the effect is small. We observe an error of 9.3% on the RT data set, an error of 9.6% on the ERP-BCI data set and an error of 13.1% on the CNT data set. This improved performance is due to two effects: first, adjusting feature thresholds for the specific study may improve the performance of each feature. For example, a re-training of the 8–13 Hz feature of the CNT data set decreased its error from 33.3% to 18.0%. Second, feature weights adjust such that more discriminative features obtain a higher weight. Interestingly, after re-training both C_{ERP} and C_{CNT} primarily use one of the two pattern features—C_{ERP} focuses mostly on the *current density norm* feature, while C_{CNT} is strongly based on the *range within pattern* feature.

3.3. Effect on BCI performance

The upper plots of figure 3 show scatter plots of BCI performance with and without preceding ICA artifact cleaning for the three analyzed BCI paradigms. For ERP-BCI, BCI performance decreased slightly from 69.4% to 68.3% ($t(20) = -2.43$, $p = 0.03$, $d = 0.21$). On average, 44 components were retained and 16 artifactual components were removed. There was no significant change in overall MI-CSP performance

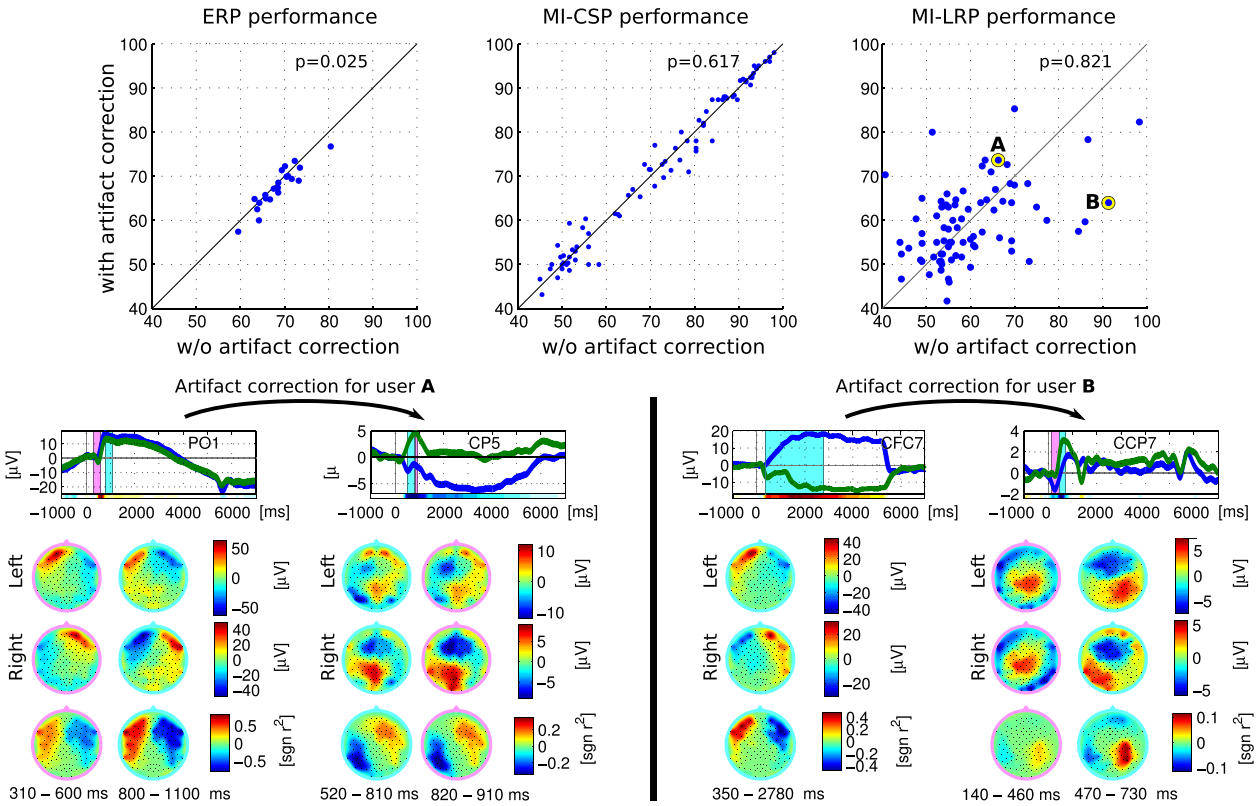


Figure 3. Upper plots: effect of artifact correction for three BCI paradigms. Dots over the diagonal indicate participants, whose data improved in classification performance (in per cent correct trials), dots below indicate participants whose performance decreased by the correction. Changes are strongest for the paradigm MI-LRP, which is most sensitive to eye artifacts. For this paradigm, participants (A) and (B) are highlighted, which undergo relatively strong changes. Lower plots: effect of artifact cleaning for participants (A) and (B). Top row: average activity of selected channels for left trials (blue) and right trials (green). The four upper scalp plots indicate the spatial distribution of average activity (in μV) for one or two time intervals (in columns) and for left and right trials (upper and lower scalp plots). Lowest scalp plots indicate the spatial distribution of class-discriminative information (as signed r^2 values) per interval. For participant A, a dominating eye artifact could be removed, which lead to an increase in the SNR and of classification performance. For participant B, very little class-discriminant signal remained after artifact cleaning.

Table 1. Feature weight vectors w and test errors (MSE) for three data sets (RT, ERP-BCI and CNT) and three classification strategies (fixed classifier C_{RT}, adapted classifier C_{RT-A} and study-specific classifiers C_{ERP}, C_{CNT}). Test errors are reported for the 6 single features and for the combined classification. The fixed classifier is trained on the RT train data set. The adapted classifier is trained on the RT train data set cut to the specific electrode montage. The study-specific classifiers are trained on data from the same study and evaluated with leave-one-subject-out CV.

			Current density norm	Range within pattern	Local skewness	λ	8–13 Hz	FitError	Combined
RT	C_{RT}	w	0.485	0.511	0.404	0.155	−0.522	−0.210	
		MSE	0.144	0.151	0.355	0.158	0.171	0.173	**0.093**
ERP-BCI	C_{RT}	MSE	0.296	0.289	0.459	0.244	0.154	0.357	**0.185**
	C_{RT-A}	w	0.454	0.463	0.384	0.235	−0.563	−0.247	
		MSE	0.178	0.259	0.459	0.244	0.154	0.357	**0.133**
	C_{ERP}	w	0.533	0.085	0.363	0.359	−0.650	−0.009	
		MSE	0.244	0.289	0.376	0.237	0.150	0.298	**0.096**
CNT	C_{RT}	MSE	0.421	0.198	0.275	0.190	0.323	0.489	**0.167**
	C_{RT-A}	w	0.341	0.498	0.417	0.234	−0.587	−0.251	
		MSE	0.265	0.214	0.275	0.190	0.323	0.489	**0.140**
	C_{CNT}	w	0.035	0.589	0.459	0.259	−0.602	−0.010	
		MSE	0.234	0.196	0.232	0.163	0.180	0.569	**0.131**

($t(79) = -0.50$, $p = 0.62$, $d = 0.04$) which remained constant at ≈72% after the removal of on average 18 artifactual components (69 components were kept). In both BCI systems, the effect per subject was small.

The strongest changes were observed for the MI-LRP paradigm, which is most prone to eye artifacts due to the focus on low-frequency signal components. Note that as feedback was provided with a moving cursor, eye activity

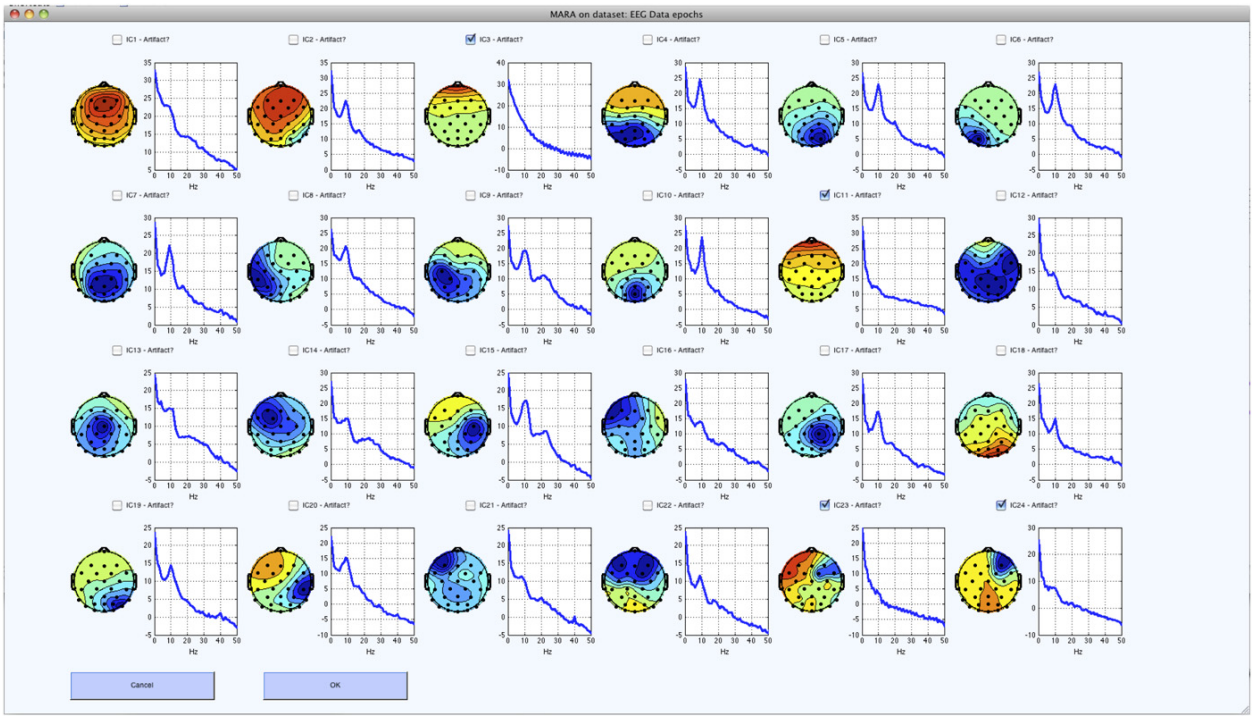


Figure 4. Screen shot of the MARA plug-in applied to EEGLAB sample data.

may be correlated with the two classes. On average, nine components were retained and ten artifactual components were removed. While the mean BCI accuracy remained constant at ≈60% ($t(79) = 0.23$, $p = 0.82$, $d = 0.03$), the performance of each participant varied considerably. The lower plots of figure 3 exemplarily highlight the effect of the artifact rejection for two participants. Without artifact rejection, both participants mainly use eye artifacts for BCI control (frontal class-discriminative activation). The effect of artifact removal can be twofold. For participant A, eye artifacts obstruct the underlying neural activity, and the system's accuracy improved upon artifact cleaning from 66.3% to 73.6% due to an improved signal-to-noise level. In participant B, very little class-discriminant activity remained after the eye activity was removed. BCI classification dropped considerably from 91.3% to 64.0%.

4. Discussion

To summarize, we have analyzed the robustness properties of our recently proposed artifact classification method and proposed a strategy to handle a wide range of electrode setups. The proposed *adapted* strategy fully automates the time-consuming rating of artifactual ICs and reliably identified multiple types of artifacts from 35 participants and 3 EEG paradigms.

IC classification performance of three strategies was evaluated against expert ratings. We showed that our simplest automatic *fixed* strategy (train the classifier once, then apply to other setups) exhibits sensitivity to drastically reduced electrode setups. As a solution, we proposed the *adapted* strategy which recomputes the training features based on the specific electrode montage of the test sets. Using this relatively inexpensive strategy—no hand-labeling is involved—artifact classification generalizes well even on very reduced electrode setups.

For comparison reasons, a re-training of the classifier using labor-intensively gained hand-labeled ICs from every new study was analyzed (strategy *study-specific*). While avoiding some generalization issues in theory, it is prohibitively expensive in most practical situations and only achieved a performance gain of a few per cent compared with the *adapted* strategy.

We therefore recommend the *adapted* strategy for artifact classification. It generalized robustly even to completely novel EEG paradigms, with its IC classification performance (13.3% MSE on auditory ERP data and 14.0% MSE on auditory listening data) staying on a similar level as inter-expert disagreements (often above 10% [34, 39]). This classification error is remarkably low given that the studies have been recorded with half the number of electrodes, used different ICA methods and contained different proportions of artifactual components.

We provide the ready-to-use artifact classifier to the community as an open-source EEGLAB plug-in called MARA (multiple artifact rejection algorithm). MARA automatically adapts to novel channel setups and its output is designed to support the experimenter in his or her decisions: a semi-automatic mode allows for visual inspection of components and for changing the classifier's proposed ratings. Figure 4 shows an example screen shot of the visual inspection menu. The plug-in is published under the

General Public License (GPL) and can be downloaded from www.user.tu-berlin.de/irene.winkler/artifacts/.

BCI practitioners may find the application of MARA on BCI data sets of particular interest. We used the *adapted* strategy to analyze how ICA artifact cleaning impacts on single-trial BCI performance of three different BCI paradigms. In all three paradigms, we were able to remove artifactual activity while maintaining the average BCI performance.

On the single subject level the effect of artifact cleaning depends on whether artifacts mask the relevant neural activity or serve as a control signal for BCI. While artifact cleaning had little influence on an auditory ERP speller and on oscillatory motor imagery data analyzed with CSP, we observed strong effects for a paradigm known to be heavily affected by eye artifacts, the use of slow motor-related potentials. Here our analysis suggests that artifact removal by MARA or similar tools may drastically improve the safety and reliability of results, as they guarantee that rejected artifacts are not utilized mistakenly to control the BCI system.

Acknowledgments

We would like to thank Stefan Haufe for providing the code for the Current Density Norm feature, Claudia Sanelli and Stefan Haufe for their help with recording and preparing the RT data set, Anna Kuhlen for providing the manual labels of the CNT dataset, the authors of [50] for providing the motor imagery data set, Martijn Schreuder for his help with recording and preparing the auditory ERP-BCI data set, and Klaus-Robert Müller, Daniel Bartz and Andrew Dowding for helpful comments on the manuscript. Last but not least, we would like to thank our reviewers for their valuable comments.

This work is supported by the European ICT Programme (Project FP7-224631 *TOBI*), by the German Federal Ministry for Education and Research (BMBF) (grant 01GQ0850), by the Federal State of Berlin, and by the BrainLinks-BrainTools Cluster of Excellence (DFG, grant number EXC 1086). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

[1] Tangermann M *et al* 2012 Review of the BCI competition IV *Front. Neurosci.* **6** 55

[2] Kübler A, Nijboer F, Mellinger J, Vaughan T M, Pawelzik H, Schalk G, McFarland D J, Birbaumer N and Wolpaw J R 2005 Patients with ALS can use sensorimotor rhythms to operate a brain–computer interface *Neurology* **64** 1775–7

[3] Hill N J *et al* 2006 Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects *IEEE Trans. Neural Syst. Rehabil. Eng.* **14** 183–6

[4] Conradi J, Blankertz B, Tangermann M, Kunzmann V and Curio G 2009 Brain–computer interfacing in tetraplegic patients with high spinal cord injury *Int. J. Bioelectromagn.* **11** 65–8

[5] Millán J del R *et al* 2010 Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges *Front. Neurosci.* **4** 161

[6] Pfurtscheller G, Allison B Z, Brunner C, Bauernfeind G, Solis-Escalante T, Scherer R, Zander T O, Mueller-Putz G, Neuper C and Birbaumer N 2010 The hybrid BCI *Front. Neurosci.* **4** 42

[7] Gramann K, Gwin J T, Ferris D P, Oie K, Jung T-P, Lin C-T, Liao L-D and Makeig S 2011 Cognition in action: imaging brain/body dynamics in mobile humans *Rev. Neurosci.* **22** 593–608

[8] Debener S, Minow F, Emkes R, Gandras K and de Vos M 2012 How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology* **49** 1617–21

[9] Castermans T, Duvinage M, Petieau M, Hoellinger T, De Saedeleer C, Seetharaman K, Bengoetxea A, Cheron G and Dutoit T 2011 Optimizing the performances of a P300-based brain–computer-interface in ambulatory conditions *IEEE J. Emerg. Sel. Top. Circuits Syst.* **4** 566–77

[10] Hari R and Kujala M V 2009 Brain basis of human social interaction: from concepts to brain imaging *Physiol. Rev.* **89** 453–79

[11] Tangermann M, Krauledat M, Grzeska K, Sagebaum M, Vidaurre C, Blankertz B and Müller K-R 2009 Playing pinball with non-invasive BCI *Advances in Neural Information Processing Systems 21, December 8–11, 2008* ed D Koller, D Schuurmans, Y Bengio and L Bottou (Vancouver, BC: MIT Press) pp 1641–8

[12] Blankertz B *et al* 2010 The Berlin brain–computer interface: non-medical uses of BCI technology *Front. Neurosci.* **4** 198

[13] van Erp J, Lotte F and Tangermann M 2012 Brain–computer interfaces: beyond medical applications *Computer* **45** 26–34

[14] Fatourechi M, Bashashati A, Ward R K and Birch G E 2007 EMG and EOG artifacts in brain computer interface systems: a survey *Clin. Neurophysiol.* **118** 480–94

[15] Makeig S, Bell A J, Jung T-P and Sejnowski T J 1996 Independent component analysis of electroencephalographic data *Advances in Neural Information Processing Systems* vol 8 ed D S Touretzk, M C Mozer and M E Hasselmo (Cambridge, MA: MIT Press) pp 145–51

[16] Jung T-P, Makeig S, Humphries C, Lee T-W, Mckeown M J, Iragui V and Sejnowski T J 2000 Removing electroencephalographic artifacts by blind source separation *Psychophysiology* **37** 163–78

[17] Fitzgibbon S P, Powers D M W, Pope K J and Richard Clark C 2007 Removal of EEG noise and artifact using blind source separation *Clin. Neurophysiol.* **24** 232–43

[18] Romero S, Mañanas M A and Barbanoj M J 2008 A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: a simulation case *Comput. Biol. Med.* **38** 348–60

[19] Crespo-Garcia M, Atienza M and Cantero J L 2008 Muscle artifact removal from human sleep EEG by using independent component analysis *Ann. Biomed. Eng.* **36** 467–75

[20] McMenamin B W, Shackman A J, Maxwell J S, Bachhuber D R W, Koppenhaver A M, Greischar L L and Davidson R J 2010 Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG *NeuroImage* **49** 2416–32

[21] Campos Viola F, De Vos M, Hine J, Sandmann P, Bleeck S, Eyles J and Debener S 2012 Semi-automatic attenuation of cochlear implant artifacts for the evaluation of late auditory evoked potentials *Hear. Res.* **284** 6–15

[22] Gwin J T, Gramann K, Makeig S and Ferris D P 2010 Removal of movement artifact from high-density EEG recorded during walking and running *J. Neurophysiol.* **103** 3526–34

[23] Meinecke F, Ziehe A, Kawanabe M and Müller K-R 2002 Resampling approach to estimate the stability of one- or multidimensional independent components *IEEE Trans. Biomed. Eng.* **49** 1514–425

[24] Choi S, Cichocki A, Park H-M and Lee S-Y 2005 Blind source separation and independent component analysis: a review *Neural Inform. Process. Lett. Rev.* **6** 1–57
[25] Romero S, Mañanas M A, Riba J, Morte A, Giménez S, Clos S and Barbanoj M J 2004 Evaluation of an automatic ocular filtering method for awake spontaneous EEG signals based on independent component analysis *EMBS: 26th Annu. Int. Conf. Engineering in Medicine and Biology Society* pp 925–8
[26] Shoker L, Sanei S and Chambers J 2005 Artifact removal from electroencephalograms using a hybrid BSS-SVM algorithm *IEEE Signal Process. Lett.* **12** 721–4
[27] Gómez-Herrero G, De Clercq W, Anwar H, Kara O, Egiazarian K, Van Huffel S and Van Paesschen W 2006 Automatic removal of ocular artifacts in the EEG without an EOG reference channel *NORSIG: Proc. 7th Nordic Signal Processing Symp.* pp 130–3
[28] Campos Viola F, Thorne J, Edmonds B, Schneider T, Eichele T and Debener S 2009 Semi-automatic identification of independent components representing EEG artifact *Clin. Neurophysiol.* **120** 868–77
[29] Mognon A, Jovicich J, Bruzzone L and Buiatti M 2010 ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features *Psychophysiology* **48** 229–40
[30] Bigdely-Shamlo N, Kreutz-Delgado K, Kothe C and Makeig S 2013 Eyecatch: data-mining over half a million EEG independent components to construct a fully-automated eye-component detector *Proc. 35th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society* pp 5845–8
[31] Frøhlich L, Andersen T S and Mørup M 2013 Classification of independent components of EEG into multiple artifact classes *BaCI Conf. Abstract: Conf. on Basic and Clinical Multimodal Imaging*
[32] LeVan P, Urrestarazu E and Gotman J 2006 A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification *Clin. Neurophysiol.* **117** 912–27
[33] Halder S, Bensch M, Mellinger J, Bogdan M, Kübler A, Birbaumer N and Rosenstiel W 2007 Online artifact removal for brain–computer interfaces using support vector machines and blind source separation *Comput. Intell. Neurosci.* **7** 1–10
[34] Winkler I, Haufe S and Tangermann M 2011 Automatic classification of artifactual ICA-components for artifact removal in EEG signals *Behav. Brain Funct.* **7** 30
[35] Asadi Ghanbari A, Kousarrize M R N, Teshnehlab M and Aliyari M 2009 An evolutionary artifact rejection method for brain computer interface using ICA *Int. J. Electr. Comput. Sci.* **9** 48–53
[36] Bartels G, Shi L-C and Lu B-L 2010 Automatic artifact removal from EEG—a mixed approach based on double blind source separation and support vector machine *EMBC'10: Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society* pp 5383–6
[37] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21
[38] Shackman A J, McMenamin B W, Slagter H A, Maxwell J S, Greischar L L and Davidson R J 2009 Electromyogenic artifacts and electroencephalographic inferences *Brain Topography* **22** 7–12
[39] Klekowicz H, Malinowska U, Piotrowska A J, Wolynczyk-Gmaj D, Niemcewi S and Durka P 2009 On the robust parametric detection of EEG artifacts in polysomnographic recordings *Neuroinformatics* **7** 147–60
[40] Kuhlen A K, Allefeld C and Haynes J-D 2012 Content-specific coordination of listeners' to speakers' EEG during communication *Front. Human Neurosci.* **6** 266
[41] Fonov V, Evans A C, McKinstry R C, Almli C R and Louis Collins D 2009 Unbiased nonlinear average age-appropriate brain templates from birth to adulthood *NeuroImage* **47** (Suppl. 1) S102
[42] Fonov V, Evans A C, Botteron K, Almli C R, McKinstry R C and Louis Collins D 2011 Unbiased average age-appropriate atlases for pediatric studies *NeuroImage* **54** 313–27
[43] Nolte G and Dassios G 2005 Analytic expansion of the EEG lead field for realistic volume conductors *Phys. Med. Biol.* **50** 3807–23
[44] Hämäläinen M S and Ilmoniemi R J 1994 Interpreting magnetic fields of the brain: minimum-norm estimates *Med. Biol. Eng. Comput.* **32** 35–42
[45] Haufe S, Nikulin V V, Ziehe A, Müller K-R and Nolte G 2008 Combining sparsity and rotational invariance in EEG/MEG source reconstruction *NeuroImage* **42** 726–38
[46] Ziehe A, Müller K-R, Nolte G, Mackert B-M and Curio G 2000 Artifact reduction in magnetoneurography based on time-delayed second-order correlations *IEEE Trans. Biomed. Eng.* **47** 75–87
[47] Hyvärinen A and Oja E 1997 A fast fixed-point algorithm for independent component analysis *Neural Comput.* **9** 1483–92
[48] Schreuder M, Rost T and Tangermann M 2011 Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI *Front. Neurosci.* **5** 112
[49] Blankertz B, Lemm S, Treder M S, Haufe S and Müller K-R 2011 Single-trial analysis and classification of ERP components—a tutorial *NeuroImage* **56** 814–25
[50] Blankertz B, Sannelli C, Halder S, Hammer E M, Kübler A, Müller K-R, Curio G and Dickhaus T 2010 Neurophysiological predictor of SMR-based BCI performance *NeuroImage* **51** 1303–9
[51] Blankertz B, Tomioka R, Lemm S, Kawanabe M and Müller K-R 2008 Optimizing spatial filters for robust EEG single-trial analysis *IEEE Signal Proc. Mag.* **25** 41–56
[52] Krauledat M, Dornhege G, Blankertz B, Losch F, Curio G and Müller K-R 2004 Improving speed and accuracy of brain–computer interfaces using readiness potential features *IEMBS'04: 26th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society* vol 2 pp 4511–5

3.3 On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP

Irene Winkler, Stefan Debener, Klaus-Robert Müller, Michael Tangermann. On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. *Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE*, 2015.

Short summary. Successful ICA-based artifact reduction relies on suitable pre-processing. In this conference paper, we systematically evaluated the effects of high-pass filtering at different frequencies, with the help of MARA. Analyses were based on event-related potential (ERP) data from 21 participants performing a task which induces well known auditory ERPs. As a pre-processing step for ICA, we found that high-pass filtering between 1-2 Hz consistently produced good results. Furthermore, ICA-based artifact removal with MARA outperformed a regression-based approach to remove eye artifacts.

Contributions. I wrote the majority of the paper and carried out all the analysis.

On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP

Irene Winkler, Stefan Debener, Klaus-Robert Müller, and Michael Tangermann

***Abstract*— Standard artifact removal methods for electroencephalographic (EEG) signals are either based on Independent Component Analysis (ICA) or they regress out ocular activity measured at electrooculogram (EOG) channels. Successful ICA-based artifact reduction relies on suitable pre-processing. Here we systematically evaluate the effects of high-pass filtering at different frequencies. Offline analyses were based on event-related potential data from 21 participants performing a standard auditory oddball task and an automatic artifactual component classifier method (MARA). As a pre-processing step for ICA, high-pass filtering between 1-2 Hz consistently produced good results in terms of signal-to-noise ratio (SNR), single-trial classification accuracy and the percentage of 'near-dipolar' ICA components. Relative to no artifact reduction, ICA-based artifact removal significantly improved SNR and classification accuracy. This was not the case for a regression-based approach to remove EOG artifacts.**

I. INTRODUCTION

Electroencephalography (EEG) measurements of brain activity are contaminated by undesired additional signals. These artifacts are caused by non-neural physiological activities of the subject, such as movements of the eyes and muscles, heart beat and pulse, and by external technical sources.

A common approach for artifact reduction is the transformation of EEG signals into a space of independent source components (ICs) using Independent Component Analysis (ICA). Ideally ICA separates artifactual and neural activity into distinct ICs, so that artifactual ICs can be identified [1] and a cleaner EEG can be constructed without them. In practice, ICs are often mixtures of both neural and artifactual sources. However, pre-processing of EEG data prior to ICA can improve the quality of the artifact separation [2], e.g. by removal of obvious, high-amplitude artifactual epochs, or by dimensionality reduction with Principal Component Analysis (PCA). Here we focus on the role of high-pass filtering.

Without prior high-pass filtering, ICA often produces visibly poor separation, with many mixed components, such as the ones displayed in the right part of Fig. 2. Recent research indicates that high-pass filtering improves reliability [3] and measures of independence and dipolarity [4] of the estimated independent components. Furthermore, trial-by-trial fluctuations of the blood-oxygen-level dependent (BOLD) signal were found to be positively correlated with high EEG gamma power when ICA de-mixing was obtained on gamma band-pass filtered EEG data, but not when 30 Hz low-pass filtered data was fed into ICA [5]. In this paper, we systematically analyze the effect of ICA-based artifact reduction on Event-Related Potentials (ERPs), as well as the percentage of 'near-dipolar' ICA-components [6], as a function of the high-pass filter frequency.

Our focus is not on how to obtain best classification performance. Instead, we will use single-trial classification performance as a proxy for successful artifact reduction. In addition, we focus on the signal-to-noise ratio (SNR) of ERPs. For varying cutoff frequencies, we train ICA on high-pass filtered data, automatically classify the resulting ICs with MARA [7], [8], and apply the obtained filter weights on the original non high-pass filtered data (see Fig. 1).

Similar to an analysis presented in [9], we also compare ICA-based artifact removal with a regression-based method, which uses electrooculogram (EOG) signals to partially remove ocular activity [10].

II. METHODS

A. Why high-pass filtering influences ICA decomposition

Given EEG signals $x_1, \ldots, x_K$ recorded from K electrodes over time, ICA methods linearly decompose the data into K source components $s_1, \ldots, s_K$. To solve this blind source separation problem, ICA assumes the mutual independence of source components and a linear generative model $x_j = \sum_{k=1}^{K} \mathbf{a}_k[j] \cdot s_k \quad (j \in \{1, \ldots, K\})$. Here $\mathbf{a}_k \in \mathbb{R}^K$ denotes the spatial activation pattern of source k and $\mathbf{a}_k[j]$ its jth element.

High-pass filtering is a linear transformation of the signals. Therefore, if the assumed generative model was true, filtering

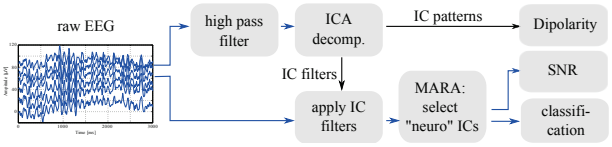


Fig. 1: Schematic workflow for high-pass filtering of EEG data prior to ICA decomposition. IC filters are applied on the unfiltered raw data before their classification by MARA.

This work was supported by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education and by the German Research Council (DFG) within the Cluster of Excellence BrainLinks-BrainTools (EXC 1086).

Irene Winkler (irene.winkler@tu-berlin.de) and Klaus-Robert Müller (klaus-robert.mueller@tu-berlin.de) are with the Machine Learning Group, Berlin Institute of Technology, Germany. Klaus-Robert Müller is also with the Department of Brain and Cognitive Engineering, Korea University, Republic of Korea. Stefan Debener (stefan.debener@uni-oldenburg.de) is with the Cluster of Excellence Hearing4all, University of Oldenburg and the Neurophysiology Lab, Department of Psychology, European Medical School, University of Oldenburg, Germany. Michael Tangermann (michael.tangermann@blbt.uni-freiburg.de) is with the Brain State Decoding Lab, BrainLinks-BrainTools Cluster of Excellence and Computer Science Dept., University of Freiburg, Germany.

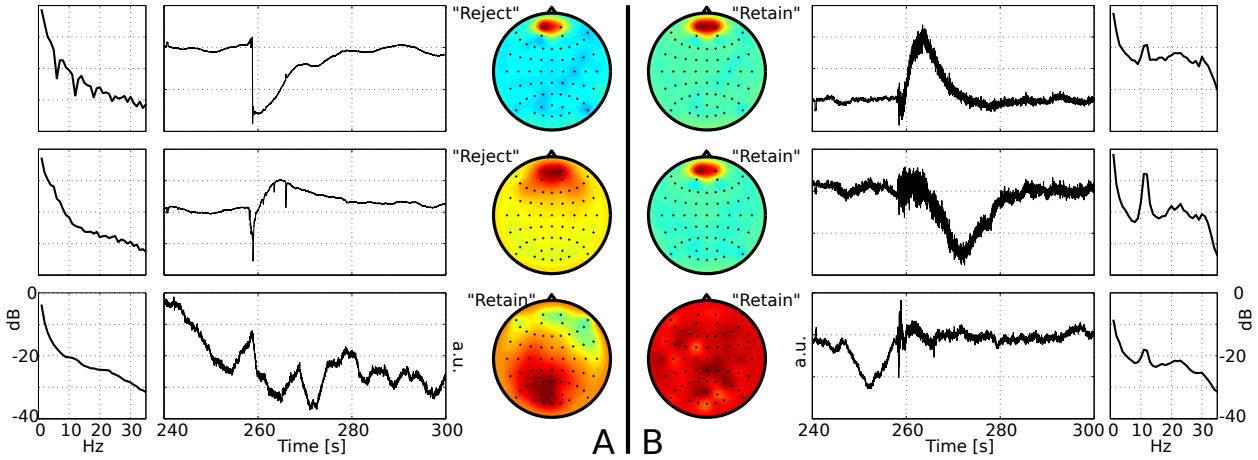


Fig. 2: Example for ICA decompositions of one study participant gained (A) with vs. (B) without 1 Hz high-pass preprocessing. Scalp patterns, spectrum and time course of the three components explaining most variance are shown together with labels (reject vs. retain) estimated by MARA, which labels components conservatively: mixed components, which contain both artifactual and neuronal activity tend to be retained in the data. In contrast to ICA after high-pass filtering, ICA on the unfiltered data did not separate oscillatory activity from the electrode pop artifact.

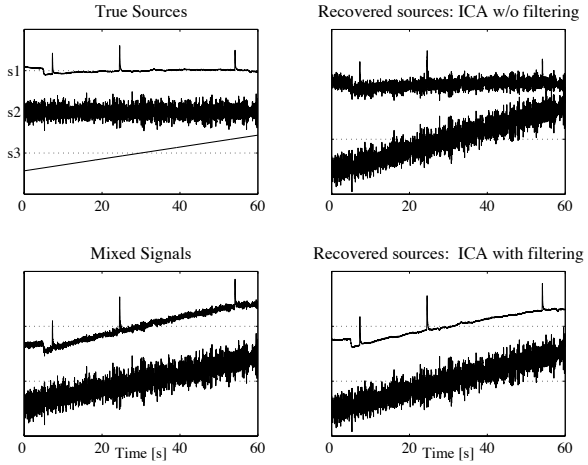


Fig. 3: Toy example of an underdetermined ICA problem with two sensors but three independent sources: (s1) eye blinks, (s2) oscillatory neural activity, and (s3) a linear drift. ICA without prior high-pass filtering separates the drift from the eye blink component, but both components contain neural activity. In contrast, ICA applied after 1 Hz high pass filtering separates the eye blink from the neural component.

would not change the ICA model coefficients: Under the model assumptions, it also holds for the filtered signals $h(x_j)$ that $h\left(x_j\right) = h\left(\sum_{k=1}^{K} \mathbf{a}_k[j] \cdot s_k\right) = \sum_{k=1}^{K} \mathbf{a}_k[j] \cdot h\left(s_k\right)$, where $h(\cdot)$ denotes linear filtering. The filtered source signals $h(s_k)$ remain mutually independent, and the coefficients of the mixing matrix $\mathbf{a}_k[j]$ are unchanged. It is therefore valid to use the filtered data for the estimation of ICA only, and then apply the obtained demixing matrix to the unfiltered data (see [2], [11] for more information).

In practice, high-pass filtering does make an important difference. It is, however, not entirely understood why this is the case. High-pass filtering can help ICA estimation by increasing the independence between sources, because slowly changing trends are not very independent [2], [11]. Furthermore, standard ICA assumptions such as the limited number of sources are at best approximately met in practice. Filtering 'guides' the ICA decomposition towards extracting the components that explain the activity we are interested in and may help to better satisfy ICA's stationarity assumption. The low-frequency parts of an EEG signal contain a large portion of its variance, that we are typically not so interested in. It is thus often beneficial to remove them. A simple toy example, which illustrates this point, is presented in Fig. 3.

B. Data

Data of 21 healthy subjects were recorded during a standard auditory oddball paradigm as part of an auditory Brain Computer Interface (BCI) study [12]. The experiment was conducted according to the Declaration of Helsinki, participants provided their written informed consent prior to participation. They were asked to avoid blinking while attending rare high-frequent target tones and disregard frequent non-target tones. This measurement lasted approx. 10 minutes. Per participant, 400 non-target and 100 target stimuli were presented in randomized order at a Stimulus Onset Asychnrony (SOA) of 1 s. It can be expected, that attended target tones lead to a more negative ERP component around 100 ms post stimulus (N1) compared to non-target responses, and to a positive ERP component (P3) at approx. 300 ms post stimulus. EEG was recorded with nose reference from 61 scalp channels and one electrode below the right eye (EOGvu). Signals were downsampled to 100 Hz

and low-pass filtered at 45 Hz (forward-backward two-pass 5th order Butterworth filter) for all variants of the following offline analysis. As no rejection of any epoch or channel was performed, artifacts remained in the data.

C. Evaluation metrics

Visual inspection: The tested artifact removal variants were independently applied to the continuous data. For plotting grand average ERP responses and further ERP processing, data was epoched around the stimulus, and baseline activity was removed according to an interval of 150 ms duration pre-stimulus.

SNR: Suitable artifact processing has the potential to decrease single-trial noise around the average ERP response. For each artifact removal variant, we assessed this effect by computing the signal-to-noise ratio (SNR). We measure SNR as proposed by [13]: Given N epochs, $y_1, \ldots, y_N \in \mathbb{R}^T$, each measured over T time points at one channel, the SNR is given as

$$\text{SNR} = \frac{\text{Var}_t\{\bar{y}(t)\}}{\frac{1}{N}\sum_{n=1}^{N}\text{Var}_t\{y_n(t) - \bar{y}(t)\}} \tag{1}$$

where $\bar{y}(t) = \frac{1}{N}\sum_{n=1}^{N} y_n(t)$ is the ERP averaged over epochs at time t. This defines the ratio of the variance of the ERP (*signal*) and the mean variance of residual deviation (*noise*).

We report SNR values for the N1–P2 complex at channel FC3 in the interval [100–250 ms], and separately for target and non-target epochs. We use the target class to compute SNR for P3 at channel Cz in the interval [250–550 ms].

Classification: The accuracy for the binary target vs. non-target classification task was estimated by chronological 10-fold cross-validation of a linear classifier (shrinkage regularized Linear Discriminant Analysis) trained on features which consist of windowed means derived from specific non-overlapping consecutive intervals [14]. Classification was performed in three conditions, which used different ERP feature intervals (all values in ms): (1) the N1-P2 complex (100–130, 140–170, 180–210, 220–250 $\Rightarrow 4 \cdot 61 = 244$ features), (2) the P3 component (260–340, 350–430, 440–550 $\Rightarrow 3 \cdot 61 = 183$ features) and (3) combined ($\Rightarrow 7 \cdot 61 = 427$ features). Accuracy is defined as the percentage of correctly classified epochs and is reported class-wise normalized, i.e. the accuracies were calculated for both classes separately and the results were averaged.

Dipolarity: We also compare ICA decompositions with a measure that does not depend on the classification of artifactual components or subsequent EEG analysis: the percentage of 'near-dipolar' components as proposed by [6]. It is defined as the percentage of components whose scalp patterns can be explained by a single equivalent dipole (which we identify with MUSIC [15]) with less than a specified error variance (which we set to 20%). As discussed in [6] in detail, this 'dipolarity' of the ICA decomposition is a very informative, albeit simplistic measure of physiological plausibility.

D. ICA and the effect of high-pass filtering

To evaluate the effect of high-pass filtering, we computed the ICA demixing on high-pass filtered data at 37 different cutoff frequencies (0.1, 0.2, ..., 2, 2.5, ..., 6, 7, ..., 10, 15, ..., 40 Hz). Filtering was carried out with a second order Butterworth filter. We chose FastICA [16] as it is frequently used ICA method for the analysis of EEG data.

The obtained de-mixing coefficients were then applied to the *unfiltered* data, as proposed e.g. in [17] and illustrated in Fig. 1. In this way, we only consider the effect of filtering on the ICA decomposition. (Note that de-mixing coefficients are often applied to the filtered data, in which case filtering affects all subsequent analysis. Filtering may distort the shape of ERP components, including peak amplitudes and onset latencies [18], but can improve classification performance [19], [20]. This is not the focus of our attention here.)

The resulting source components were labeled with MARA [7], [8]. MARA is a heuristic, that solves the binary classification problem 'reject vs. accept' fast and objectively. It is able to handle eye artifacts, muscular artifacts and loose electrodes. When confronted with mixed components, MARA decides conservatively and retains them in the data, as shown in Fig. 2. Subsequently, (cleaner) EEG data was reconstructed by omitting components labeled as artifacts.

E. Comparison with regression

We compare ICA-/MARA-based artifact cleaning (with 1 Hz and 2 Hz pre-filtering) to a regression approach for the removal of eye activity measured at additional electrooculogram (EOG) channels [10]. A standard procedure is to subtract part of the vertical (VEOG) and horizontal EOG (HEOG) from each recorded EEG electrode x_j as

$$z_j(t) = x_j(t) - \hat{\alpha}_j VEOG(t) - \hat{\beta}_j HEOG(t) - \hat{\gamma}_j \tag{2}$$

where z_j denotes the 'cleaned' EEG signal at electrode j, and $\hat{\alpha}_j, \hat{\beta}_j$ and $\hat{\gamma}_j$ are regression coefficients estimated by ordinary least squares.

Regression-based methods are limited by the fact that EOG is contaminated by brain activity which is removed as well. To alleviate bidirectional contamination, EOG channels are typically low-pass filtered prior to the regression step. Motivated by [9], [21], we used a cut-off at 7.5 Hz.

For this analysis EOG was derived as a post-hoc bipolar derivation from channels (F9, F10) for horizontal EOG and (Fp2, EOGvu) for vertical EOG. Channels F9, F10 and Fp2 were excluded from the set of EEG channels.

III. RESULTS

A. ICA and the effect of high-pass filtering

Grand-average ERPs for several artifact removal variants are depicted in Fig. 5. The influence of high-pass filtering on the ICA decomposition is not strongly reflected in the shape of ERPs. Peak amplitudes were only slightly attenuated even when ICA was trained on 30 Hz high-pass filtered data. Similarly, strong drifting components were removed both when ICA was trained on filtered and unfiltered data.

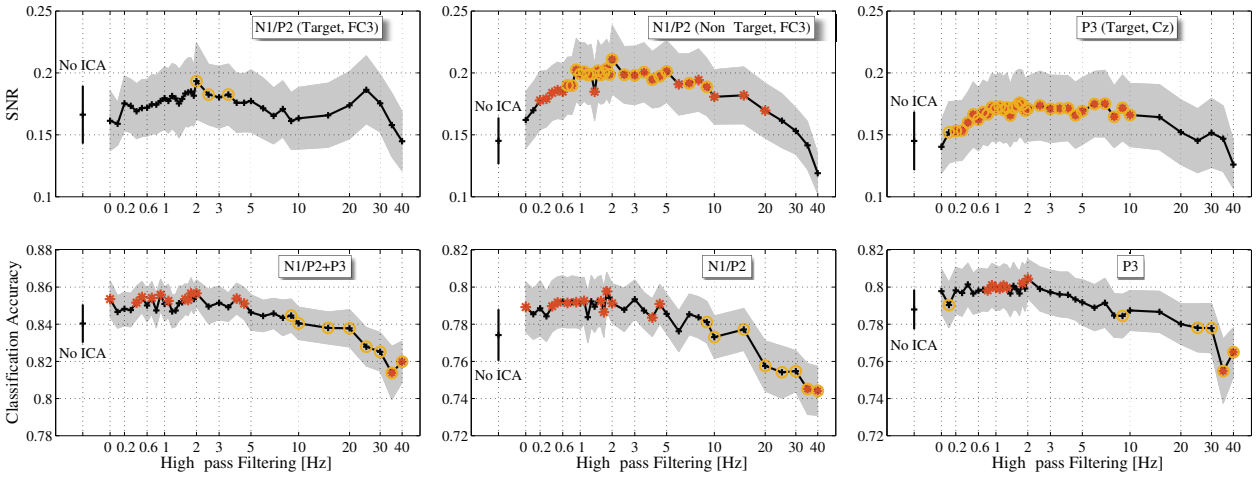


Fig. 4: SNR values (top row) and performance of target vs. non-target classification (bottom row) including means and ± standard errors (s.e.) as a function of high-pass filtering cut-off frequency applied before ICA. Condition '*0*' refers to no high-pass filtering before ICA artifact removal, and condition '*No ICA*' omits the ICA artifact removal completely. A red star indicates values which significantly differ from condition '*No ICA*', a yellow circle indicates values which differ from condition '*0*', according to a Wilcoxon signed rank test at $p < 0.05$. No multiple-testing corrections were applied.

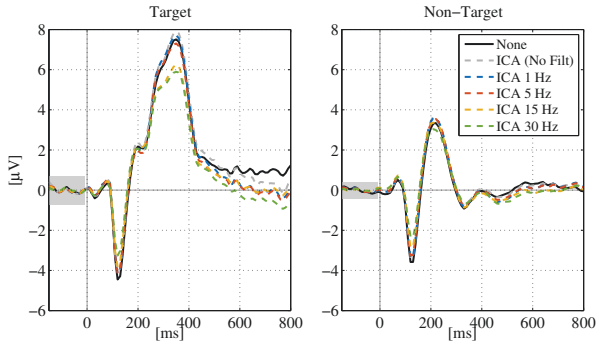


Fig. 5: Target- and non-target grand average (n=21) ERP responses at channel Cz from either raw data or after ICA-based cleaning. ICA weights were always applied to the unfiltered data, however ICA was trained on high-pass filtered data at different cutoff frequencies.

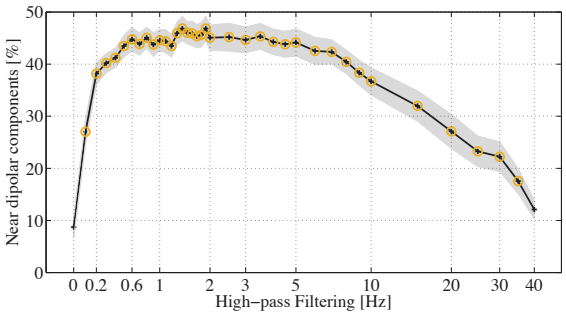


Fig. 6: Percentage of components whose scalp patterns can be explained from the scalp projection of one equivalent dipole with less than 20% error variance (± s.e.), as a function of high-pass filtering frequency applied before ICA. A yellow circle indicates values which significantly differ from condition '*0*' (no high-pass filtering), as in Fig. 4.

Obtained SNR values, classification performances and percentages of near-dipolar components are depicted in Fig. 4 and Fig. 6. Values on the x-axis indicate the type of high-pass filtering applied. For N1-P2 SNR (non-targets), P3 SNR (targets) and the percentage of near-dipolar components we see a consistent increase for small frequencies. The effect is particularly strong (and highly significant) for the percentage of near-bipolar ICA components, which ranges from 9% without to 47% with 1.9 Hz pre-filtering. Similarly, best SNR values are achieved between 1 and 5 Hz. ICA artifact removal applied to high-pass filtered data in these ranges significantly improves SNR as compared to no artifact reduction and compared to ICA without high-pass filtering.

We observe no strong differences between early and late ERP components. However, SNR values of the N1-P2 complex of the target class are less sensitive than for the non-target class. This may be because there are four times as many epochs in the non-target class, which allows for a more accurate SNR estimate. Classification accuracy is also not very sensitive to the artifact reduction variants we analyzed. Nevertheless, ICA artifact removal significantly improved over no artifact reduction mostly when applied to high-pass filtered data in frequencies between 0.5 and 2 Hz.

B. Comparison with regression

A comparison of ICA-based (with 1 Hz and 2 Hz pre-filtering) with regression-based artifact removal in terms of SNR and classification performance is summarized in Table I. ICA with 2 Hz pre-filtering yielded higher SNR and classification accuracy than 1 Hz pre-filtering, however this

TABLE I: SNR and overall classification accuracy (mean $\pm$ s.e.) for different artifact reduction methods. $<, >$ indicate significant differences ($p < 0.05$, Wilcoxon signed rank).

	None (N)	**ICA (I1)** -1 Hz-	**ICA (I2)** -2 Hz-	**Regr. (R)**
Accuracy (in %)	84.2±1.0 $<$ I1 I2	85.3±0.9 $>$ N	**85.7**±0.7 $>$ N	84.7±1.0
SNR N1/P2 (Target)	0.17±0.02	0.17±0.03 $<$ I2	**0.19**±0.03 $>$ I1 R	0.16±0.02 $<$ I2
SNR N1/P2 (Non-Target)	0.15±0.02 $<$ I1 I2	0.19±0.02 $>$N R , $<$ I2	**0.21**±0.03 $>$ N I1 R	0.15±0.02 $<$ I1 I2
SNR P3 (Target)	0.15±0.02 $<$ I1, I2	0.17±0.03 $>$ N	**0.17**±0.03 $>$ N R	0.15±0.02 $<$ I2

effect is not consistently significant. Classification accuracy increased after ICA for both pre-filtering variants, which was not the case for the regression-based approach. In terms of SNR, ICA with 2 Hz pre-filtering significantly improved over regression-based artifact removal.

IV. DISCUSSION

In this paper, we have quantified the impact of high-pass filtering on artifact reduction, focussing on ERPs obtained from a standard auditory oddball task resulting in typical target and non-target event-related responses. With adequate pre-filtering, artifact cleaning based on ICA and MARA improved both, classification accuracy and signal-to-noise ratio (SNR). Consistent with [9], this was not the case for the regression-based method we analyzed in addition.

In general, SNR was more sensitive to variations in artifact removal than classification accuracy or the shape of the ERP. As the obtained high classification rates are not strongly influenced by cleaner data, we conjecture that a relative large distance between class means dominates the investigated classification problem. Of course, this situation may be different with other ERP data sets. An example is the data analyzed in [8], which was recorded under much lower stimulus onset asynchrony (175 ms instead of 1000 ms in the oddball data). ICA-based artifact reduction of this BCI data increased the classification error slightly, but significantly. This result may be explained by the fact, that target and non-target responses have lower amplitudes and are not separated as clearly as in the oddball data. Thus any class-discriminative activity may be contained in small-variance signal components which are harder to separate from artifactual components using ICA.

The fact that pre-filtering strongly influences the unmixing quality of ICA is highlighted by its strong impact on the percentage of near-dipolar ICA components. Consistent with the results obtained from the SNR, the dipolarity measure indicates that pre-filtering at very small frequencies < 0.5 Hz may not be optimal. In our analysis, high-pass filtering between 1 and 2 Hz consistently produced good results in terms of SNR, classification accuracy, and 'dipolarity' of the ICA decomposition. If information is contained in slow signal components, ICA can be trained on filtered data, and the learned weights can be applied to the unfiltered data.

ACKNOWLEDGMENT

We thank the authors of [12] for providing data and Stefan Haufe and Guido Nolte for advise on and implementation of MUSIC, respectively.

REFERENCES

[1] L. Frølich, T. S. Andersen, and M. Mørup, "Classification of independent components of EEG into multiple artifact classes," *Psychophysiology*, vol. 52, no. 1, pp. 32–45, 2015.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[3] D. M. Groppe, S. Makeig, and M. Kutas, "Identifying reliable independent components via split-half comparisons," *NeuroImage*, vol. 45, no. 4, pp. 1199 – 1211, 2009.

[4] Z. Zakeri, S. Assecondi, A. Bagshaw, and T. Arvanitis, "Influence of signal preprocessing on ICA-based EEG decomposition," in *XIII MEDICON 2013*, pp. 734–737.

[5] R. Scheeringa, P. Fries, K.-M. Petersson, R. Oostenveld, I. Grothe, D. G. Norris, P. Hagoort, and M. C. Bastiaansen, "Neuronal dynamics underlying high- and low-frequency EEG oscillations contribute independently to the human BOLD signal," *Neuron*, vol. 69, no. 3, pp. 572 – 583, 2011.

[6] A. Delorme, J. Palmer, J. Onton, R. Oostenveld, and S. Makeig, "Independent EEG sources are dipolar," *PLoS ONE*, vol. 7, no. 2, p. e30135, 02 2012.

[7] I. Winkler, S. Haufe, and M. Tangermann, "Automatic classification of artifactual ICA-components for artifact removal in EEG signals," *Behavioral and Brain Functions*, vol. 7, p. 30, 2011.

[8] I. Winkler, S. Brandl, F. Horn, E. Waldburger, C. Allefeld, and M. Tangermann, "Robust artifactual independent component classification for BCI practitioners," *J. Neural. Eng.*, vol. 11, no. 3, p. 035013, 2014.

[9] F. Ghaderi, S. K. Kim, and E. A. Kirchner, "Effects of eye artifact removal methods on single trial P300 detection, a comparative study," *Journal of Neuroscience Methods*, vol. 221, pp. 41 – 47, 2014.

[10] R. J. Croft and R. J. Barry, "Removal of ocular artifact from the EEG: a review," *Clinical Neurophysiology*, vol. 30, pp. 5–19, 2000.

[11] J. M. Pignat, O. Koval, D. V. D. Ville, S. Voloshynovskiy, C. Michel, and T. Pun, "The impact of denoising on independent component analysis of functional magnetic resonance imaging data," *Journal of Neuroscience Methods*, vol. 213, no. 1, pp. 105 – 122, 2013.

[12] M. Schreuder, T. Rost, and M. Tangermann, "Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI," *Frontiers in Neuroprosthetics*, vol. 5, no. 112, 2011.

[13] S. Lemm, G. Curio, Y. Hlushchuk, and K.-R. Müller, "Enhancing the signal-to-noise ratio of ICA-based extracted ERPs," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 4, pp. 601–607, 2006.

[14] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – a tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[15] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, 1986.

[16] A. Hyvärinen and E. Oja, "A fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 7, pp. 1483–1492, 1997.

[17] S. Debener, J. Thorne, T. R. Schneider, and F. C. Viola, "Using ICA for the analysis of multi-channel EEG data," in *Simultaneous EEG and fMRI: Recording, Analysis, and Application*, M. Ullsperger and S. Debener, Eds. Oxford University Press, 2010.

[18] A. Widmann, E. Schröger, and B. Maess, "Digital filter design for electrophysiological data - a practical approach," *Journal of Neuroscience Methods*, 2014, in press.

[19] L. Bougrain, C. Saavedra, and R. Ranta, "Finally, what is the best filter for P300 detection?" in *TOBI Workshop lll-Tools for Brain-Computer Interaction*, 2014, pp. 734–737.

[20] J. Farquhar and N. Hill, "Interactions between pre-processing and classification methods for event-related-potential classification," *Neuroinformatics*, vol. 11, no. 2, pp. 175–192, 2013.

[21] S. Romero, M. Mañanas, and M. Barbanoj, "Ocular reduction in EEG signals based on adaptive filtering, regression and blind source separation," *Annals of Biomedical Engineering*, vol. 37, no. 1, pp. 176–191, 2009.

3.4 Removal of muscular artifacts for the analysis of brain oscillations: Comparison between ICA and SSD

Irene Winkler, Stefan Haufe, Klaus-Robert Müller. Removal of muscular artifacts for the analysis of brain oscillations: Comparison between ICA and SSD. *ICML Workshop on Statistics, Machine Learning and Neuroscience*, 2015.
`http://sites.google.com/site/stamlins2015/accepted-papers`

Short summary. In this workshop paper we explore an alternative for ICA when we are interested in clean *oscillatory* EEG activity. We compare ICA with the recently proposed spatio-spectral decomposition (SSD) method. SSD is designed to extract components that explain oscillation-related variance, and is faster to compute than ICA. We investigate EEG data from 18 subjects performing self-paced foot movements with respect to event-related desynchronization (ERD) in the beta band. Results indicate that SSD recovers cleaner signals than ICA on this data set.

Contributions. I wrote the paper and carried out all the analysis.

Removal of muscular artifacts for the analysis of brain oscillations: Comparison between ICA and SSD

Irene Winkler IRENE.WINKLER@TU-BERLIN.DE
Machine Learning Group, Berlin Institute of Technology, Marchstr. 23, 10587 Berlin, Germany

Stefan Haufe STEFAN.HAUFE@TU-BERLIN.DE
Laboratory for Intelligent Imaging and Neural Computing, Columbia University, New York, NY, USA, 10027
Machine Learning Group, Berlin Institute of Technology, Marchstr. 23, 10587 Berlin, Germany

Klaus-Robert Müller KLAUS-ROBERT.MUELLER@TU-BERLIN.DE
Machine Learning Group, Berlin Institute of Technology, Marchstr. 23, 10587 Berlin, Germany
Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Republic of Korea

Abstract

The electroencephalogram (EEG) is contaminated by undesired signals of non-neural origin, such as movements of the eyes and muscles. The most common approach for muscle artifact reduction is the linear transformation of EEG signals into source components using Blind Source Separation (BSS) techniques, to separate artifactual and neuronal sources.

Here we present a case study in which we are interested in clean oscillatory EEG activity. We compare the frequently used Independent Component Analysis (ICA) approach with the recently proposed spatio-spectral decomposition (SSD) method. SSD is designed to extract components that explain oscillations-related variance, and is faster to compute than ICA. We investigate EEG data from 18 subjects performing self-paced foot movements with respect to event-related desynchronization (ERD) in the beta band. Results indicate that SSD recovers cleaner signals than ICA on this data set.

1. Introduction

As the interpretation of electroencephalographic (EEG) signals depends on relatively clean recordings, artifact reduction is an important step in EEG signal processing. These artifacts are caused by non-neural physiological activities of the subject, such as movements of the eyes and muscles, heart beat and pulse, or by external technical sources.

The most common approach for muscle artifact reduction is the linear transformation of EEG signals into source components with techniques of Blind Source Separation (BSS), the most frequently used being Independent Component Analysis (ICA). If artifactual and neural activity are contained in separate components, artifactual components can be identified and a cleaner EEG can be reconstructed.

The assumptions for the application of ICA methods are only approximately met in practice (linear mixture of independent components, stationarity of the sources and the mixture, no systematic co-activiation of artifacts and neuronal signals). Nevertheless, their application usually leads to a good separation. However, separation is usually not perfect and a number of mixed components contain both neural and artifactual activity. While several methods try to alleviate this issue, ICA remains the state-of-the-art (see e.g. (Vigario & Oja, 2008; Urigüen & Garcia-Zapirain, 2015) for a review).

In this paper, we are interested in obtaining clean *oscillatory* EEG activity. We present evidence that in some cases, the recently developed spatio-spectral decomposition (SSD) method (Nikulin et al., 2011), which extracts components explaining oscillations-related variance, may achieve better artifact reduction than ICA.

We investigate Event-Related Desychronization (ERD), that is, the suppression of brain rhythms in response to an event, in a data set that is heavily contaminated by muscle artifacts. 18 subjects performed self-paced foot movements, which are well known to be preceded by an ERD of 8-13 Hz (mu band) and 15-30 Hz (beta band) rhythms over

ICML Workshop on Statistics, Machine Learning and Neuroscience

corresponding sensorimotor areas (Neuper & Pfurtscheller, 2001). Here we focus on beta ERD, which is thought to be related to movement preparation and execution (Kilavik et al., 2013).

2. Methods

2.1. Data

Data stem from a pre-measurement of a simulated driving experiment described in (Haufe et al., 2011). 18 healthy participants were instructed to perform self-paced right foot movements (i.e. to press the brake pedal) once per second for five minutes. EEG data were recorded with 64 Ag/AgCl electrodes at 1000 Hz. Furthermore, an electromyographic (EMG) signal was recorded using a bipolar montage at the tibialis anterior muscle and the knee of the right leg. For the presented offline-analysis, EEG data were decimated to 200 Hz, broad-band filtered between 2 and 45 Hz, and artifactual electrodes were rejected using a variance criterion.

2.2. Compared methods

We compare two methods of blind source separation (BSS) for artifact reduction. BSS is the task of recovering underlying signals $S \in \mathbb{R}^{K \times T}$ from multivariate recordings $X \in \mathbb{R}^{M \times T}$ generated from the linear model $X = AS$, with very little information about the underlying source signals S or the mixing process $A \in \mathbb{R}^{M \times K}$. Here K denotes the number of source signals, M denotes the number of electrodes and T denotes the number of available time points. The problem is underdetermined and can only be solved using assumptions about the signals to be recovered. A demixing matrix $\hat{W} \in \mathbb{R}^{K \times M}$ is estimated such that the estimated sources

$$\hat{S} = \hat{W}X \tag{1}$$

best fulfill pre-defined assumptions.

In BSS-based artifact reduction we then hope that artifactual and neuronal activity are contained in different source components, so that cleaner EEG signals can be reconstructed by omitting the artifactual signals.

2.2.1. ICA

The most common approach for artifact reduction is Independent Component Analysis (ICA), which solves the BSS problem under the assumption of mutually statistically independent sources. Here we use TDSEP (Ziehe & Müller, 1998), which relies on second-order statistics by taking the temporal structure of the time series into account. TDSEP amounts to finding a demixing $\hat{W}$ which leads to minimal cross-covariances over several time-lags between all pairs of components of $\hat{S}$.

We applied TDSEP to EEG signals whose dimensionality was reduced with Principal Component Analysis (PCA) to 99.9% explained variance. We then manually selected artifactual components, based on the pattern, spectrum and time course of each component. On average per subject, 14 components were identified as artifacts, and cleaner EEG signals were reconstructed with the remaining 17 components. For a description of typical artifact components we refer the reader to (Chaumon et al., 2015).

2.2.2. SSD

The purpose of spatio-spectral decomposition (SSD) (Nikulin et al., 2011) is to extract brain oscillations in a frequency band of interest. It maximizes the signal power in a frequency band of interest (here: 15 - 30 Hz) while simultaneously minimizing it at the neighboring frequency bins (here: 13-14 Hz, 31-32 Hz). SSD seeks spatial filters $\mathbf{w} \in \mathbb{R}^k$ which maximize

$$\mathrm{SNR}(\mathbf{w}) = \frac{\mathbf{w}^\top \Sigma_{\mathrm{sig}} \mathbf{w}}{\mathbf{w}^\top \Sigma_{\mathrm{noise}} \mathbf{w}} \tag{2}$$

where Σ_{sig} is the covariance of the data filtered in the frequency band of interest and Σ_{noise} is the covariance of the data filtered in the sidebands. The entire SSD demixing matrix can be computed by solving a generalized eigenvalue problem in a matter of seconds (Nikulin et al., 2011; Haufe et al., 2014).

For the subsequent analysis, we retained the 10 components with the highest SNR, as in (Dähne et al., 2014; Winkler et al., 2015). This choice of 10 SSD components was based on prior experience.

2.3. Event-Related Desynchronization

SSD and ICA were independently applied to the continuous EEG data. To compare the methods, we plot grand-average Event-related (de-)synchronization (ERD/ERS) in the beta band (15 - 30 Hz), aligned to EMG peak activity.

ERD is computed as the relative difference in signal power of a certain frequency band compared to a reference period (Pfurtscheller & Aranibar, 1979; Blankertz et al., 2008):

$$\mathrm{ERD}(t) = \frac{\mathrm{Power}(t) - \mathrm{Reference\ power}}{\mathrm{Reference\ power}} \tag{3}$$

where $\mathrm{Power}(t)$ denotes the average power over all trials at time point t. We use the interval of [-1200 -800 ms] prior to EMG peak activity as the reference interval.

3. Results

Grand-average ERDs for the three different preprocessing variants (Nothing, SSD and ICA) are depicted in Fig. 1.

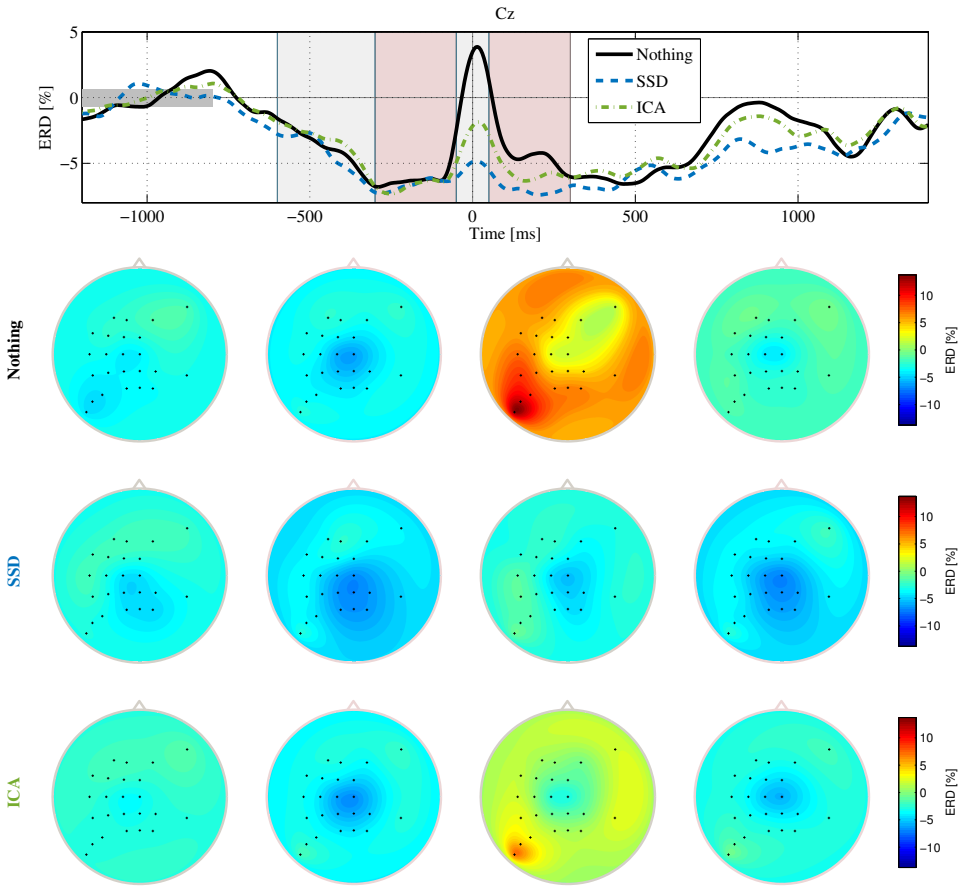


Figure 1. Grand-average ERD/ERS for 18 subjects recorded during self-paced foot movements in the beta band (15-30 Hz), aligned to EMG peak activity. The plots show time courses at channel Cz and series of ERD maps in the marked intervals ([-600 -300], [-300 -50], [-50 50], [50 300]) for three conditions: No pre-processing, SSD pre-processing (10 components with highest SNR as defined in Eq. (2) were retained) and ICA-based artifact removal (TDSEP; manually selected neural components were retained). The maps represent a top view on the head with nose pointing upwards, + indicate electrodes.

Prior to foot movement, we see a typical foot ERD over central sensorimotor areas as expected (cf. (Neuper & Pfurtscheller, 2001)). During movement, the ERD is contaminated by a muscular artifact which spans the whole scalp. This is probably due to subjects moving their heads along with the fairly rhythmical foot movement once per second.

This muscle artifact is reduced after a preceding ICA-based artifact removal step. ERD is even cleaner when SSD was applied. SSD is able to almost completely eliminate the artifact, without removing neural beta activity.

The computational time was estimated on a MacBook 2.66 GHz, 8 GB RAM, using Matlab R2012a. For the presented data set, computing SSD took on average 1.5 s (including filtering). In contrast, the average computation time for TDSEP was 4.0 s and FastICA (Hyvärinen & Oja, 1997) took 36.5 s –both after the number of components was reduced to explain only 99.9% variance.

4. Discussion

We presented preliminary evidence that SSD may be a powerful tool for the removal of artifacts in EEG data, when the neural signals of interest are of oscillatory nature. Compared to ICA, SSD is faster to compute, and recovered a cleaner grand-average ERD on the self-paced movement data set we analyzed here.

Our findings are in line with findings in Motor Imagery based Brain-Computer-Interfaces (BCI), where SSD im-

proved classification performance (Haufe et al., 2014). In contrast, applying two different ICA methods (Bell & Sejnowski, 1995; Ziehe & Müller, 1998) in combination with two different automatic artifactual component classifiers (Frøhlich et al., 2015; Winkler et al., 2014) on the same data set, did not (Winkler et al., 2014; Frølich et al., 2015).

SSD is especially designed to increase the signal-to-noise ratio of oscillatory sources, and it is therefore not surprising that it can be more suitable to separate artifacts (=noise) from oscillatory neuronal signals. In the EEG data presented here, the observed muscle artifacts are also not occurring independently from motor planning neuronal activity, which violates ICA's assumptions. However, the coactiviation of artifacts and neuronal activity is quite common, and often poses the most serious problems in practice. In those cases, ICA is often applied anyway, due to lack of better alternatives and/or satisfactory performance. SSD also assumes uncorrelated sources, but seemed to be able to better separate correlated artifacts by using information about the neuronal sources' expected frequency content.

The findings presented here are subject to future research. Interestingly, while SSD outperforms ICA in terms of identifying the subspace that contains the relevant neural activity, ICA was better at extracting a single motor preparatory source component on the same data set (cf. (Winkler et al., 2015)). Another point is that ICA separation quality depends on pre-processing steps such as high-pass filtering. It remains to be seen whether stronger high-pass filtering would improve ICA's performance. An open question is also how many components to choose for SSD.

Acknowledgments

SH was supported by a Marie Curie International Outgoing Fellowship (grant No. PIOF-GA-2013-625991) within the 7th European Community Framework Programme. KRM acknowledges support by the BK21 Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology.

References

Bell, Anthony J. and Sejnowski, Terrence J. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

Blankertz, Benjamin, Tomioka, Ryota, Lemm, Steven, Kawanabe, Motoaki, and Müller, Klaus-Robert. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Proc Magazine*, 25(1):41–56, 2008.

Chaumon, Maximilien, Bishop, Dorothy V.M., and Busch, Niko A. A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of Neuroscience Methods*, 2015. in press.

Dähne, Sven, Nikulin, Vadim V., Ramírez, David, Schreier, Peter J., Müller, Klaus-Robert, and Haufe, Stefan. Finding brain oscillations with power dependencies in neuroimaging data. *NeuroImage*, 96:334–348, 2014.

Frøhlich, Laura, Andersen, Tobias S., and Mørup, Morten. Classification of independent components of EEG into multiple artifact classes. *Psychophysiology*, 52(1):32–45, 2015.

Frølich, Laura, Winkler, Irene, Müller, Klaus-Robert, and Samek, Wojciech. Investigating effects of different artefact types on motor imagery bci. In *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

Haufe, Stefan, Treder, Matthias S, Gugler, Manfred F, Sagebaum, Max, Curio, Gabriel, and Blankertz, Benjamin. EEG potentials predict upcoming emergency brakings during simulated driving. *Journal of Neural Engineering*, 8(5), 2011.

Haufe, Stefan, Dähne, Sven, and Nikulin, Vadim V. Dimensionality reduction for the analysis of brain oscillations. *NeuroImage*, 101:583–597, 2014.

Hyvärinen, Aapo and Oja, Erkki. A fixed-point algorithm for independent component analysis. *Neural Computation*, 7:1483–1492, 1997.

Kilavik, Elisabeth, Zaepffel, Manuel, Brovelli, Andrea, MacKay, William A., and Riehle, Alexa. The ups and downs of beta oscillations in sensorimotor cortex. *Experimental Neurology*, 245:15 – 26, 2013.

Neuper, Christa and Pfurtscheller, Gert. Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. *International Journal of Psychophysiology*, 43:41–58, 2001.

Nikulin, Vadim V., Nolte, Guido, and Curio, Gabriel. A novel method for reliable and fast extraction of neuronal EEG/MEG osciallations on the basis of spatio-spectral decomposition. *NeuroImage*, 55:1528–1535, 2011.

Pfurtscheller, Gert and Aranibar, A. Evaluation of event-related desynchronization preceding and following voluntary self-paced movement. *Electroencephalogr. Clin. Neurophysiol*, 46: 138–146, 1979.

Urigüen, Jose Antonio and Garcia-Zapirain, Begoña. EEG artifact removal – state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001, 2015.

Vigario, Ricardo and Oja, Erkki. BSS and ICA in neuroinformatics: From current practices to open challenges. *Biomedical Engineering, IEEE Reviews in*, 1:50–61, 2008.

Winkler, Irene, Brandl, Stephanie, Horn, Franziska, Waldburger, Eric, Allefeld, Carsten, and Tangermann, Michael. Robust artifactual independent component classification for BCI practitioners. *J. Neural. Eng.*, 11(3):035013, 2014.

Winkler, Irene, Haufe, Stefan, Porbadnigk, Anne K., Müller, Klaus-Robert, and Dähne, Sven. Identifying granger causal relationships between neural power dynamics and variables of interest. *NeuroImage*, 111:489 – 504, 2015.

Ziehe, Andreas and Müller, Klaus-Robert. TDSEP - an efficient algorithm for blind source separation using time structure. In *ICANN 98*, pp. 675–680, 1998.

3.5 Conclusion

Summary

In this chapter, we worked on signal processing techniques for the removal of artifacts from EEG signals. We constructed the subject-independent component classification method MARA (Multiple Artifact Rejection Algorithm) that automates the process of tedious handselection of artifactual independent components (ICs). MARA delivers physiologically interpretable results, generalizes well over different experimental setups, and has been made available as an open-source EEGLAB plugin. It is not limited to a specific type of artifact and should be able to handle muscle artifacts, eye artifacts, and loose electrodes equally well.

MARA realizes automatic artifactual independent component classification using a linear pre-trained classifier. It is based on the following six features which were determined in a feature selection procedure (Section 3.1):

1. Two features are extracted from the scalp pattern of an IC. They attempt to quantify whether the pattern is neurophysiologically plausible. The first feature, Current Density Norm, quantifies how easy it is to model the pattern by an underlying source distribution over 2142 locations arranged in a 1 cm spaced 3D-grid. The second feature, Range Within Pattern, is simply the difference between the minimal and the maximal activation in a standardized pattern. Spatially localized scalp maps stemming from muscle artifacts or loose electrodes are typically characterized by a high Range Within Pattern.

2. One feature is extracted from the time series of an IC. The mean absolute local skewness of time intervals of 15 s duration aims to detect outliers in the time series.

3. Three features are extracted from the spectrum of an IC. One spectral feature, λ, measures how steep the 1/f spectrum is. Muscle artifacts are characterized by unusual high values in the 20-50 Hz range, which is reflected by a comparatively steep curve. Two spectral features aim to detect the typical α peak in components of neuronal origin. One is the average log band power of the α band (8-13 Hz), and one is the deviation of a 1/f approximation to the spectrum in the 8-15 Hz range.

MARA is trained on labeled components of one study. Its generalization ability has been thoroughly tested:

- On labeled IC components (Section 3.2): On data of the same study, MARA misclassified 9.3 % of 450 unseen tested components. On data of an auditory ERP-BCI study (Schreuder et al., 2011), MARA misclassified 13.3 % of

540 tested components. On data from an auditory listening task (Kuhlen et al., 2012), MARA misclassified 14.0 % of 4473 tested components.

The performance of the classifier has to be judged in the light of the fact that inter-expert disagreements on EEG signals can be above 10%. We asked one expert to re-label 690 components of the training set, two years after the original labeling. The disagreement between the new and the former rating was 13.2%. For the auditory ERP-BCI data, two persons who labeled all components disagreed on 10.6% of the components. This imperfect agreement between human expert has also been reported by (Klekowicz et al., 2009; Viola et al., 2009).

Thus, the classification error of MARA is relatively low, especially considering that some studies have been recorded with half the number of electrodes, used different ICA methods and contained different proportions of artifactual components.

- Using well known ERPs (Section 3.3): We analyzed a data set of 21 healthy subjects which performed a task that induces well-known ERPs. In a standard auditory oddball paradigm, participants were asked to attend rare high-frequent target tones and disregard frequent non-target tones (Schreuder et al., 2011). Artifact cleaning based on ICA and MARA improved single-trial classification accuracy (target vs. non-target) and a signal-to-noise ratio (SNR) measure for ERPs proposed in (Lemm et al., 2006). It is also notable that this was not the case for a regression-based method we analyzed, which is consistent with a similar finding in (Ghaderi et al., 2014).

Furthermore, we used MARA to evaluate ICA-based artifact removal in the following settings:

- In BCI studies (Section 3.2): We analyzed how ICA artifact cleaning with MARA impacts on single-trial BCI performance of three different BCI paradigms using data from 101 participants (Blankertz et al., 2010a; Schreuder et al., 2011).

  ICA-based artifact removal had no significant influence on oscillatory motor imagery data analyzed with CSP. For offline data from an auditory ERP speller, it increased the classification error slightly, but significantly.

  However, we could show on an individual level the effect of artifact cleaning can be very ambiguous — artifacts can either mask the relevant neuronal activity, or serve as a control signal for BCI. We observed a strong influence for a paradigm known to be strongly affected by eye artifacts: the use of slow motor-related potentials. Here our analysis suggests that artifact removal by

MARA or similar tools may improve the reliability of results, as they guarantee that rejected artifacts are not utilized mistakenly to control the BCI system.

- Impact of high-pass filtering on ICA (Section 3.3): It has to be pointed out that the success of ICA-based artifact removal strategies crucially depends on the quality of the obtained decomposition. It is well known that pre-processing of EEG data prior to ICA, such as high-pass filtering, can improve the quality of the artifact separation (Hyvärinen et al., 2001; Scheeringa et al., 2011; Pignat et al., 2013). Using MARA, we systematically analyzed the effect of ICA-based artifact reduction on Event-Related Potentials (ERPs), as well as the percentage of 'near-dipolar' ICA-components (Delorme et al., 2012), as a function of the high-pass filter frequency. We found that, as a pre-processing step for ICA, high-pass filtering between 1-2 Hz consistently produced good results in terms of signal-to-noise ratio (SNR), single-trial classification accuracy and the percentage of 'near-dipolar' ICA components.

Limitations and Future Work

We would like to stress that ICA + MARA is not a black-box method. We call it 'fully automatic', because it can be applied in a fully automatic mode. Nevertheless, expert knowledge is needed to judge whether the algorithms perform well on the particular data set at hand. The quality of the obtained ICA decomposition as well as MARA's classification should still be inspected at least for a few subjects of a study. MARA, just like any other automatic IC classifier, is a heuristic, as there is no guarantee that the correct components are identified.

From our experience, MARA works relatively well on ERP data sets. However, on the self-paced braking data sets we analyzed for the workshop paper presented in Section 3.4, MARA labeled components too conservatively. This is why we manually labeled the components on this data set. In general, if the data are heavily contaminated by artifacts, ICA will have problems with extracting clean neuronal components. MARA then tends to label components conservatively, that is, mixed components which contain both artifactual and neuronal activity tend to be retained in the data. In this case, MARA may not reject enough components.

In summary, MARA yields sufficient performance for it to be a useful tool in many cases. However, we have encountered data sets in which its performance could be better. There are several factors which limit MARA's performance. Since artifacts are a major problem for EEG analysis, future work should improve upon the following aspects:

- *Limited number of training data points.* At the moment, MARA is trained on 1290 components which were labeled by only two persons in our group.

This number is small, especially if we want to generalize over a large variety of experimental setups, some of which may generate more artifactual signals then others. Furthermore, labeling independent components is not a clear-cut task, and is subject to human judgement. It would therefore be good to have labels from different research groups.

MARA could be made more reliable if more training components were available. However, to the best of our knowledge, almost no labeled component data is currently publicly available. For eye artifacts, an interesting plugin was recently developed by (Bigdely-Shamlo et al., 2013). They provide a database of 3452 eye-related IC scalp patterns, to which new input ICs are compared against. These scalp maps were obtained by data-mining over half a million IC scalp maps obtained from about 80 000 EEG datasets. Research efforts in this direction seem very promising.

- *Applicability to general electrode setups.* The applicability of MARA to general electrode setups is still an issue. So far, MARA is only applicable to EEG data which have been recorded using electrode positions from the 10-20-system, just as the training data. This is not always the case, as some researchers use dense array EEG. An interpolation between the electrodes of the training data is therefore desirable.

- *A classification pipeline which identifies eye artifacts first.* Eye artifacts are much easier to identify than other types of artifacts (Halder et al., 2007; Winkler et al., 2011; Frølich et al., 2015a). This is because eye components have characteristic scalp patterns, which do not vary strongly over subjects and sessions. Also, eye components are typically quite high-variance signals with very non-Gaussian distributions (due to many outliers in the time series), and are therefore typically well extracted with ICA (cf. (Hyvärinen et al., 2010a)).

  The classification problem 'eye artifact vs. not an eye artifact' is therefore an easier classification problem than 'artifact vs. not an artifact'. Thus, it might be more reliable to solve the eye artifact classification problem separately (for example with EyeCatch, the method developed by (Bigdely-Shamlo et al., 2013)), before identifying other artifact components with a classifier like MARA.

Last but not least, the key to BSS-based artifact removal is to obtain a good decomposition into artifactual and neuronal components. Unfortunately, separation is usually not perfect and a number of mixed components contain both neuronal and artifactual activity. This existence of mixed components is the limiting factor for BSS-based artifact removal in general, because the EEG researcher –or the automatic classifier– is forced to decide between removing too much information (when

removing a mixed component) or retaining artifacts in the data (when retaining a mixed component).

While several methods try to alleviate this issue, ICA has been the state-of-the-art for many years. This might indicate that it is not possible to drastically improve over ICA in general. Nevertheless, decomposition may be improved using the right preprocessing settings (Section 3.3), and interesting alternatives may exist for special cases. For example, the recently proposed spatio-spectral decomposition (SSD) method (Nikulin et al., 2011) seems very well suited for the removal of artifacts/noise for the analysis of oscillatory signals (Haufe et al., 2014a).

In Section 3.4, we therefore compared ICA with SSD. SSD is designed to extract components that explain oscillations-related variance (cf. Section 2.2.3). We investigate EEG data from 18 subjects performing self-paced foot movements (Haufe et al., 2011) with respect to event-related desynchronization (ERD) in the beta band. The data contain a typical foot ERD over central sensorimotor areas which is clearly contaminated by an event-locked muscle artifact. This contamination is probably due to subjects moving their heads along with the fairly rhythmical foot movement once per second. The clearly visible artifact may allow us to investigate the performance of more artifact removal methods in the future.

The results so far indicate that SSD recovers cleaner signals than ICA on this data set. We therefore use SSD as a pre-processing step for the analysis of Granger causal brain oscillations in the next chapter.

4 Extracting Granger causal brain oscillations

4.1 Identifying Granger causal relationships between neural power dynamics and variables of interest

Irene Winkler, Stefan Haufe, Anne Porbadnigk, Klaus-Robert Müller, and Sven Dähne. Identifying Granger causal relationships between neural power dynamics and variables of interest. *NeuroImage*, 111: 489-504, 2015.
`http://www.sciencedirect.com/science/article/pii/S1053811914010647`

Short summary. The functional role of oscillatory activity and its causal effects on behavior is a field of intense research. In this paper, we investigate which methods are best suited to reveal Granger causal links between the power of brain oscillations and experimental variables. Using both simulated and real EEG recordings, we compare Granger causal analysis on power dynamics obtained from a) sensor directly, b) two state-of-the-art multivariate methods and c) a novel multivariate method, referred to as Granger Causal Power Analysis (GrangerCPA), that directly optimizes for Granger causality.

We find that all three multivariate approaches are better suited than sensor analysis. For the multivariate approaches, comparison must be undertaken in more detail. GrangerCPA slightly outperforms ICA in some data sets, but this is not always the case. In general, both algorithms arrive at similar solutions using an entirely different set of assumptions.

Contributions. I wrote the majority of the paper and carried out all the analysis.

NeuroImage 111 (2015) 489–504

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Identifying Granger causal relationships between neural power dynamics and variables of interest

Irene Winkler [a,*], Stefan Haufe [a,b,c], Anne K. Porbadnigk [a,c,d], Klaus-Robert Müller [a,c,d,e,*], Sven Dähne [a,d,*]

[a] Machine Learning Laboratory, Berlin Institute of Technology, Marchstr. 23, 10587 Berlin, Germany
[b] Neural Engineering Group, Department of Biomedical Engineering, The City College of New York, New York City, NY, USA
[c] Bernstein Focus Neurotechnology, Berlin, Germany
[d] Bernstein Center for Computational Neuroscience, Berlin, Germany
[e] Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Republic of Korea

ARTICLE INFO

Article history:
Accepted 19 December 2014
Available online 30 December 2014

Keywords:
EEG
MEG
Oscillations
Granger causality

ABSTRACT

Power modulations of oscillations in electro- and magnetoencephalographic (EEG/MEG) signals have been linked to a wide range of brain functions. To date, most of the evidence is obtained by correlating bandpower fluctuations to specific target variables such as reaction times or task ratings, while the causal links between oscillatory activity and behavior remain less clear. Here, we propose to identify causal relationships by the statistical concept of Granger causality, and we investigate which methods are bests suited to reveal Granger causal links between the power of brain oscillations and experimental variables.

As an alternative to testing such causal links on the sensor level, we propose to linearly combine the information contained in each sensor in order to create virtual channels, corresponding to estimates of underlying brain oscillations, the Granger-causal relations of which may be assessed. Such linear combinations of sensor can be given by source separation methods such as, for example, Independent Component Analysis (ICA) or by the recently developed Source Power Correlation (SPoC) method.

Here we compare Granger causal analysis on power dynamics obtained from i) sensor directly, ii) spatial filtering methods that do not optimize for Granger causality (ICA and SPoC), and iii) a method that directly optimizes spatial filters to extract sources the power dynamics of which maximally Granger causes a given target variable. We refer to this method as Granger Causal Power Analysis (GrangerCPA).

Using both simulated and real EEG recordings, we find that computing Granger causality on channel-wise spectral power suffers from a poor signal-to-noise ratio due to volume conduction, while all three multivariate approaches alleviate this issue. In real EEG recordings from subjects performing self-paced foot movements, all three multivariate methods identify neural oscillations with motor-related patterns at a similar performance level. In an auditory perception task, the application of GrangerCPA reveals significant Granger-causal links between alpha oscillations and reaction times in more subjects compared to conventional methods.

Introduction

Oscillatory neural activity is a fundamental property of neuronal networks and has widely been linked with distinct brain functions (Jensen et al., 2007; Nikulin et al., 2007; Rieder et al., 2011; Başar, 2012). Bandpower fluctuations in electro- and magnetoencephalography (EEG/MEG), as well as electrocorticography (ECoG), signals have been shown to be correlated with behavioral measures of task performance or perceptual experience in humans and have been related to a variety of cognitive phenomena, including attention (Debener et al., 2003; Bauer et al., 2006; Womelsdorf and Fries, 2007; Haegens et al., 2011), memory (Klimesch, 1999; Osipova et al., 2006), vigilance (Oken et al., 2006; Berka et al., 2008) and perception (Kaiser et al., 2006; Thut et al., 2006; Babiloni et al., 2006; Schubert et al., 2009). As most of the evidence is of correlative nature, the functional role of oscillatory activity and its causal effects on behavior remain a field of intense research (Buzski and Draguhn, 2004; Thut and Miniussi, 2009).

An intriguing way to investigate the functional role of oscillations is to induce them with brain stimulation techniques such as repetitive Transcranial Magnetic Stimulation (rTMS) and Transcranial Alternating Current Stimulation (TACS) (Thut et al., 2012; Herrmann et al., 2013). Accumulating evidence suggests that rhythmic stimulation induces behavioral consequences, for instance on visual perception (Romei et al., 2010), motor performance (Joundi et al., 2012), mental rotation

* Corresponding authors.
E-mail addresses: irene.winkler@tu-berlin.de (I. Winkler), shaufe@ccny.cuny.edu (S. Haufe), klaus-robert.mueller@tu-berlin.de (K.-R. Müller), sven.daehne@tu-berlin.de (S. Dähne).

http://dx.doi.org/10.1016/j.neuroimage.2014.12.059
1053-8119/

(Klimesch et al., 2003), working memory (Zaehle et al., 2011), and sleep stages (Massimini et al., 2007).

A fundamentally different approach to studying the causal effects of oscillations that does not require direct intervention in the nervous system is the following: identification of causal relationships based on temporal precedence as revealed by a concept called 'Granger causality' (Granger, 1969). Granger causality is a standard statistical method from the field of econometrics and has been applied in neuroscience to infer functional brain connectivity (e.g. Roebroeck et al., 2005; Astolfi et al., 2007; Bressler and Seth, 2011). Assume we simultaneously measure EEG bandpower ϕ and a target variable z over time. Then ϕ is said to Granger cause z if ϕ helps to predict the future z above what is predicted by the past of z alone. Here, z can be any signal of interest, such as a behavioral output (e.g., reaction time, sensory detection, task rating, evoked potentials), a physiological measure (e.g., muscular activity, heart rate variability) or a second power time course.

The advantage of non-invasiveness warrants further pursuit of the Granger causality idea, as applied to power dynamics of EEG recordings. For example, an actively researched question in the field of Brain–Computer Interfaces (BCIs) is whether (and how) oscillatory sources influence the control performance of a user during a BCIs experiment (Grosse-Wentrup et al., 2011; Dähne et al., 2011; Maeder et al., 2012). Existing results suggest a causal role of gamma power in the modulation of BCIs control performance (Grosse-Wentrup, 2011). In the Granger causal setting, the target variable z would thus be the BCIs control performance per trial, while the goal would be to identify a neural source whose power time course Granger-causes z. Due to high inter-trial variability as well as low signal-to-noise ratio (particularly in high-frequency ranges such as the gamma band), finding predictive sources is a challenging task.

The simplest approach to testing for Granger causality is to consider each channel separately. However, the physics of EEG implies that the activity measured at a given channel is a mixture of contributions from several neuronal sources, whose activity is spread across the EEG channels due to volume conduction in the head (Baillet et al., 2001; Parra et al., 2005; Nunez and Srinivasan, 2006). This leads to a low signal-to-noise (SNR) ratio and may hinder physiological interpretation of the results, because the activity of a Granger causal neural source is not guaranteed to be best observable even in the sensors that are closest to the neural source. These considerations imply that testing Granger causality on sensor-level computed power time courses is potentially suboptimal.

The complications outlined above can be avoided by recovering the underlying neural source signals from scalp recordings prior to the computation of bandpower dynamics and the test for Granger causality. The task of recovering underlying signals from multivariate recordings is called (blind) source separation (BSS), and can only be solved using prior knowledge about the signals to be recovered. In the field of neuroscience, one of the most popular BSS algorithms is Independent Component Analysis (ICA), which seeks maximally statistically independent sources. However, here we are interested in recovering sources whose power dynamics Granger cause an external variable. Thus, we might benefit from basing the reconstruction of source activity exactly on this assumed dependency. This is especially important since an oscillatory source may Granger cause a behavioral output variable by modulating other brain rhythms — which contradicts the assumption of independence to all other sources. Moreover, a benefit of directly optimizing for the quantity of interest rather than statistical independence has been demonstrated recently in the context of correlation analysis (Dähne et al., 2013, 2014a,b).

In this paper, we investigate which methods are bests suited to reveal a Granger causal effect from neural oscillations to a given external target variable. To this end, we compare channel-wise Granger causality testing with three source separation methods. We propose a novel analysis method which extracts a source whose bandpower maximally Granger causes the target variable, and we compare it with ICA and the recently proposed SPoC method (Dähne et al., 2014a) which extracts neural sources whose bandpower is maximally correlated with the target variable. This comparison is carried out both in simulations and on two real EEG data sets.

Methods

Granger causality

Granger causality (Granger, 1969) is a statistical method to infer causality between time series based on the temporal argument that the cause should precede the effect. It has been widely applied to the study of economic variables and recently been adopted in the field of neuroscience (Roebroeck et al., 2005; Astolfi et al., 2007; Bressler and Seth, 2011). While Granger causality has gained popularity as a simple testable definition of causality, note that the scientific methodology for the inference of cause–effect relationships from data is subject to intense research. A significant Granger test is often thought not to reflect 'true causality', but what is sometimes termed 'predictive causality' or simply 'Granger causality'.

Let us consider two univariate time series ϕ and z (representing, for example, the EEG power and target variable time course). According to Granger's definition, ϕ is said to Granger cause z, if we are better able to predict z using 'all the information in the universe' than if all information apart from ϕ has been used (Granger, 1969). In practice, it is common to consider only the information in the past of ϕ and z (Hamilton, 1994; Bressler and Seth, 2011). The statistical test is then given by the comparison of the goodness of fit of two autoregressive (AR) models. First, z is modeled as a function of a predefined number P of its most recent past values. Second, z is modeled as a function of both its own past values and the past values of ϕ. Finally, Granger causality tests whether the second regression model explains significantly more variance of z than the first regression model.

In the linear case, the two regression models are given as

$$z(t) = \sum_{p=1}^{P} h_{res}(p) z(t-p) + \epsilon_{res}(t) \tag{2.1}$$

and

$$z(t) = \sum_{p=1}^{P} h_{full}(p) z(t-p) + \sum_{p=1}^{P} h_{full}(P+p) \phi(t-p) + \epsilon_{full}(t), \tag{2.2}$$

where P denotes the number of time lags, $h_{res} \in \mathbb{R}^P$ and $h_{full} \in \mathbb{R}^{2P}$ denote the regression coefficients, and ϵ_{res} and ϵ_{full} denote the residuals.

ϕ is said to Granger cause z if the variance of the residuals ϵ_{res} of the restricted model is significantly larger than the variance of the residuals ϵ_{full} of the full model. Granger causality from ϕ to z can be captured with Geweke's Granger causality index (Geweke, 1982), defined as

$$\mathcal{G}_{\phi \to z} = \log \frac{\mathrm{Var}\left(\epsilon_{res}\right)}{\mathrm{Var}\left(\epsilon_{full}\right)}. \tag{2.3}$$

Under the assumption of Gaussian distributed residuals, $\mathcal{G}_{\phi \to z}$ is asymptotically χ^2 distributed. Under the same assumption, an exact test is given by the F-test for regression (see Bressler and Seth, 2011, for instance). Under the null hypothesis of no Granger causality,

$$F_{\phi \to z} = \frac{\mathrm{Var}\left(\epsilon_{res}\right) - \mathrm{Var}\left(\epsilon_{full}\right)}{\mathrm{Var}\left(\epsilon_{full}\right)} \cdot \frac{N-2P}{P} \tag{2.4}$$

will have an F distribution with $(P, N-2P)$ degrees of freedom, where N denotes the number of available data points. If the distribution of the residuals is unknown, non-parametric methods such as permutation

testing can be used for significance testing. In this paper, we use the *F*-test for regression in our simulation study, and additionally a permutation test for the real EEG data. For the permutation test, we use 500 random permutation sequences generated by the method of Freedman and Lane (Freedman and Lane, 1983; Anderson and Robinson, 2001; Barnett and Seth, 2011).

In the neuroscientific context considered here, z is a measured variable of interest such as reaction time or task performance, which we refer to as the 'target variable'. Our goal is to find neural oscillations whose power ϕ Granger causes the target variable. For example, ϕ could be the power at one channel in a specific frequency band of measured EEG or MEG data. Throughout this paper, we will always consider ϕ to be the logarithm of the power because logarithmized EEG power time courses can assume negative values and are approximately Gaussian distributed, which is more consistent with autoregressive modeling.

Generative model

Due to volume conduction, neural signals generated inside the brain are spatially smeared while propagating to the sensors. In a simulation study (Nunez et al., 1997), only half of the signals picked up by a scalp electrode came from sources within a 3 cm radius. Because the superposition of neural sources is instantaneous and linear in the sources (Baillet et al., 2001; Parra et al., 2005; Nunez and Srinivasan, 2006), the electrophysics of EEG and MEG can be modeled as

$$X = A \cdot S + \eta. \tag{2.5}$$

Here X denotes the surface potentials measured at M sensors, S denotes the time-courses of K underlying neural sources, and the matrix $A \in \mathbb{R}^{M \times K}$ describes the influence of each source to each sensor. Each column of A is called a spatial *pattern* of the respective source and depends on both the spatial distribution and orientation of the source as well as the geometry and conductivity of different brain tissues. Any contribution that is not described by A is summarized in an additive sensor noise term η.

Throughout this paper, we consider time-windowed and band-pass filtered measurements $X = (X_1,\ldots, X_N) \in \mathbb{R}^{M \times (T \cdot N)}$ where M is the number of channels, N is the number of time windows and T is the number of data points per window. This notation allows to accommodate trial-based as well as continuous data. For trial-based data, the time windows may be defined in relation to the trial structure, i.e. the window could encompass the entire trial or pre-trial time segments, for example. For continuous data on the other hand, the time windows can simply be consecutive data segments. From here on we refer to the time-windowed and band-pass filtered data as *epoched* data.

The underlying sources are denoted $S = (s_1, s_2,\ldots, s_{K-1}, s^*)^T \in \mathbb{R}^{K \times (T \cdot N)}$ and mixed into X via their corresponding patterns $A = (\mathbf{a}_1, \mathbf{a}_2,\ldots, \mathbf{a}_{K-1}, \mathbf{a}^*) \in \mathbb{R}^{M \times K}$. The external target variable $z \in \mathbb{R}^N$ is acquired once per epoch and the epoch-wise logarithmized bandpower of s^* is assumed to Granger cause z. As the data is band-pass filtered in a band-of-interest already, we can use the logarithm of the variance of the band-passed signal as a proxy for the log-power of the band-of-interest. The log-variance is computed within epochs and thus results in epoch-wise time courses of bandpower. The other biological sources $s_1,\ldots, s_{K-1}$ may or may not be related to z, but their power time courses are assumed not to Granger cause z. Fig. 1 illustrates the generative model.

Channel-wise Granger causality

Given EEG or MEG data X and an external target variable z, we would like to determine whether oscillatory brain activity is Granger causing z. The simplest approach is the univariate approach of computing epoch-wise bandpower for each channel and testing for Granger causality for each channel. However, the activity picked up from each sensor reflects a superposition of several neural sources. Thus, the signal-to-noise ratio at the sensor level is low, which may cause failure to detect Granger causality.

Can the superposition of neural sources also introduce spurious causality in the sense that Granger causality that is only due to source mixing is detected? Recent research points out problems of Granger causality when studying connectivity: If only a single source is active, and measured with noise at two channels, the channels Granger cause each other. This is because the autoprediction of each channel is improved if the prediction model is augmented by another channel that measures the same signal with a different noise realization (Haufe et al., 2013). Spurious causality is not as problematic in our application scenario, because volume conduction is not affecting the external target variable z. However, if both ϕ and z are driven by a third common cause, the problem may occur.

Consider the following simple toy example. Suppose we measured a linear superposition x of two bandpower time courses ϕ^* and ϕ_1,

$$x = \phi^* + c \cdot \phi_1 \tag{2.6}$$

where $c \in \mathbb{R}$ governs the signal-to-noise ratio. Note that this is a strongly simplified case as bandpower computation is not a linear operation.

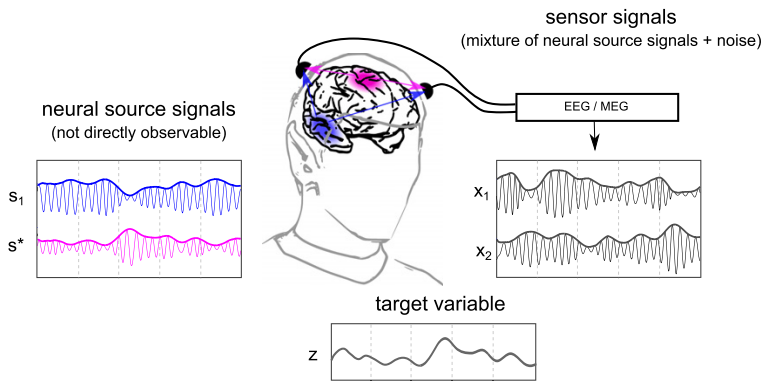


Fig. 1. Problem illustration. In this simple cartoon example, there are two active brain sources (s_1 and s^*). The power dynamics of source s^* are causally related to an externally observed target variable z, while the power dynamics of s_1 are unrelated to z. Sensors on the scalp record a linear mixture of the source activity, resulting in channel signals x_1 and x_2. Note that the channel-wise computed envelopes are not simply a linear mixture of the source envelopes. Even if a causal relation between sensor power dynamics and z can be detected, the spatial layout of such detected interaction may be misleading because the activity of source s^* is present not only in sensors in close proximity. The appropriate course of action is to recover the time course of s^* before the computation of bandpower. Information contained in the target variable z can be used to recover s^*.

How does this linear superposition influence a test for Granger causality from x to an external target variable z?

For different values of c, we simulated a target variable z and epoch-wise bandpower ϕ^* from a stable bivariate AR process of order $P = 1$, for $N = 1000$ epochs. ϕ_1 is generated independently from a stable univariate AR process of order $P = 1$. We consider two cases: (1) ϕ^* Granger causes z, but z does not Granger cause ϕ^*, and (2) ϕ^* does not Granger cause z, but z Granger causes ϕ^*. We perform 300 repetitions of the experiment for each value of c.

The frequency of estimated Granger causal significant relationships from x to z is shown in Fig. 2. We see that if ϕ^* does not Granger cause z, we will correctly infer no Granger causality for any value of the signal-to-noise parameter c. There is no problem of spurious causality here. This is because the additional noise in x will not help to predict z. Note however, that we need to assume that the target variable z is measured with a relatively low noise level to avoid spurious causality.

If ϕ^* Granger causes z, we will correctly identify Granger causality if c is small. However, the power of the Granger causality test from x to z decreases when c increases. The Granger causal relationship is thus masked by superimposed noise. This consideration implies that Granger causality testing on channel-wise computed power time-courses is potentially suboptimal. We may gain statistical power by trying to recover ϕ^* from the sensor measurements using multivariate source separation approaches.

Multivariate approaches

When analyzing a neural phenomenon of interest, it is often advisable to recover the unknown neural signals from the sensor measurements X in the linear model in Eq. (2.5), as this leads to a higher signal-to-noise ratio (Parra et al., 2005; Blankertz et al., 2008). To do so, various multivariate signal processing algorithms have been proposed that linearly combine channels to extract signals of interests. In this paper, we analyze the ability of three multivariate algorithms to extract neural oscillations whose power modulations Granger cause the external target variable.

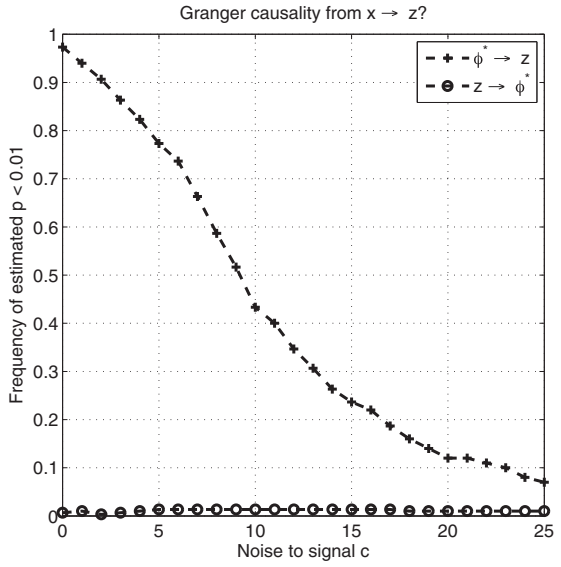


Fig. 2. Simple illustrative example of Granger causality under linear superposition $x = \phi^* + c \cdot \phi_1$. We plot the frequency of estimated significant Granger causal relationships from x to z depending on c in two scenarios: (1) ϕ^* Granger causes z, but z does not Granger cause ϕ^* and (2) ϕ^* does not Granger cause z, but z Granger causes ϕ^*.

Independent Component Analysis (ICA)

The blind source separation (BSS) problem consists in recovering the unknown neural signals $S \in \mathbb{R}^{K \times (T \cdot N)}$ from the sensor measurements $X \in \mathbb{R}^{M \times (T \cdot N)}$, i.e. of inverting the linear model in Eq. (2.5). A common approach is Independent Component Analysis (ICA), which solves the BSS problem under the assumption of mutually statistically independent sources. A demixing matrix $W \in \mathbb{R}^{K \times M}$ is estimated such that the estimated sources

$$\hat{S} = W \cdot X \tag{2.7}$$

are maximally statistically independent. Each row of W extracts one source and is called a spatial filter.

A number of measures can be used to assess whether two time series are statistically independent, and each lead to a different algorithm for determining W (Hyvärinen and Oja, 2000). We chose FastICA (Hyvärinen and Oja, 1997) as one ICA variant often used for the analysis of EEG and MEG data. Granger causality from the bandpower of each source to the target variable can then be computed.

ICA approaches are in line with the generative model of EEG and MEG and have become a standard tool for the analysis of such data. However, ICA methods are unsupervised and do not make use of the information of the target variable and typically assume the number of sources to be equal to the number of channels. We thus may gain from supervised methods that use the information in the target variable to guide the decomposition.

Source Power Correlation (SPoC)

Source Power Correlation Analysis (SPoC) is a recently proposed method which extracts sources whose bandpower maximally *correlates* with an external target (Dähne et al., 2014a). Specifically, SPoC seeks a spatial filter $\mathbf{w} \in \mathbb{R}^M$ such that the epoch-wise bandpower of the projected signal $\mathbf{w}^\top X$ is maximally correlated with z. As X has been band-pass filtered, this epoch-wise bandpower can be computed as the variance of the projected signal $\mathrm{Var}(\mathbf{w}^\top X_i)$ within each epoch i. We denote the epoch-wise bandpower by $\phi'_{\mathbf{w}}(i) := \mathrm{Var}(\mathbf{w}^\top X_i)$.

Here we consider an extension of SPoC where we maximize the correlation of the target variable with its prediction from the bandpower of several time lags P. Its objective can be summarized as

$$\begin{aligned} \underset{\mathbf{w}\in\mathbb{R}^M, h_{\mathrm{spoc}}\in\mathbb{R}^P}{\text{maximize}} \quad & \mathrm{Cov}\left(z(i),\ \sum_{p=1}^{P} h_{\mathrm{spoc}}(p)\cdot \phi'_{\mathbf{w}}(i-p)\right) \\ \text{subject to} \quad & \mathbf{w}^\top \cdot \mathrm{Cov}(X)\cdot \mathbf{w} = 1 \\ & h_{spoc}^\top \cdot B_{\mathbf{w}} \cdot h_{spoc} = 1 \end{aligned} \tag{2.8}$$

where $\phi'_{\mathbf{w}}(i) = \mathrm{Var}\left(\mathbf{w}^\top X_i\right)$ is the bandpower of the projected signal in epoch i, $\mathrm{Cov}(X)$ denotes the covariance of X and $B_{\mathbf{w}}$ denotes the autocorrelation matrix of $\mathrm{Var}(\mathbf{w}^\top X_i)$ computed over P time lags. The mathematical details of the optimization can be found in Dähne et al. (2013).

Optimizing Granger causality

We propose to directly optimize the quantity of interest, namely Granger causality rather than correlation. That is, given electrophysiological data X (for example EEG recordings) and the target variable z, we seek a spatial filter $\mathbf{w} \in \mathbb{R}^M$ such that the epoch-wise log power of the estimated source $\mathbf{w}^\top X$ maximally Granger causes the target variable. We refer to this algorithm as 'GrangerCPA' (Granger Causal Power Analysis). Combining channels to maximize Granger causality as measured on the projected data has recently been proposed by Ashrafulla et al. (2013); however, here we are interested in oscillatory phenomenon and thus the power of the projected signals.

To derive GrangerCPA's objective function, we introduce the following notation. Denote with

$$\check{\phi}_{\mathbf{w}}(i) := \log\left(\frac{\mathrm{Var}\left(\mathbf{w}^\top X_i\right)}{\mathbf{w}^\top \mathrm{Cov}(X)\mathbf{w}}\right) = \log\left(\frac{\mathbf{w}^\top \mathrm{Cov}(X_i)\mathbf{w}}{\mathbf{w}^\top \mathrm{Cov}(X)\mathbf{w}}\right) \tag{2.9}$$

the normalized log bandpower of the source extracted by the spatial filter $\mathbf{w}$ in epoch i and let

$$\phi_{\mathbf{w}}(i) := \check{\phi}_{\mathbf{w}}(i) - \frac{1}{N}\sum_{j=1}^{N}\check{\phi}_{\mathbf{w}}(j) \tag{2.10}$$

be the mean-free log bandpower. We also assume that z has been transformed to have zero mean. Let $\widetilde{Z}\in\mathbb{R}^{P\times(N-P)}$ contain z along with its lagged values, and let $\widetilde{\Phi}_{\mathbf{w}}\in\mathbb{R}^{P\times(N-P)}$ contain $\phi_{\mathbf{w}}$ along with its lagged values, i.e.

$$\widetilde{Z} := \begin{bmatrix} z(P) & \cdots & z(N) \\ \vdots & & \vdots \\ z(1) & \cdots & z(N-P) \end{bmatrix}, \quad \widetilde{\Phi}_{\mathbf{w}} := \begin{bmatrix} \phi_{\mathbf{w}}(P) & \cdots & \phi_{\mathbf{w}}(N) \\ \vdots & & \vdots \\ \phi_{\mathbf{w}}(1) & \cdots & \phi_{\mathbf{w}}(N-P) \end{bmatrix}.$$

Denote with $\widetilde{T}_{\mathbf{w}} := \begin{bmatrix} \widetilde{Z} \\ \widetilde{\Phi}_{\mathbf{w}} \end{bmatrix}\in\mathbb{R}^{2P\times N-P}$ the vertical concatenation of $\widetilde{Z}$ and $\widetilde{\Phi}_{\mathbf{w}}$.

We are now ready to calculate the coefficients of the restricted and full autoregressive models given in Eqs. (2.1) and (2.2) using Ridge Regression:

$$h_{res} = \left(\widetilde{Z}\widetilde{Z}^\top + \lambda_r I\right)^{-1}\widetilde{Z}z^\top \tag{2.11}$$

$$h_{full} = \left(\widetilde{T}_{\mathbf{w}}\widetilde{T}_{\mathbf{w}}^\top + \lambda I\right)^{-1}\widetilde{T}_{\mathbf{w}}z^\top \tag{2.12}$$

where λ_r and λ denote the respective regularization parameters. Plugging the regression's residuals, given as $\epsilon_{res} = z - h_{res}^\top\widetilde{Z}$ and $\epsilon_{full} = z - h_{full}^\top\widetilde{T}_{\mathbf{w}}$, in Eq. (2.3) yields a Granger causality index of

$$\mathcal{G}_{\phi_{\mathbf{w}}\to z} = \log\frac{\mathrm{Var}\left(\epsilon_{res}\right)}{\mathrm{Var}\left(\epsilon_{full}\right)} = \log\frac{\left\|z - z\widetilde{Z}^\top\left(\widetilde{Z}\widetilde{Z}^\top + \lambda_r I\right)^{-1}\widetilde{Z}\right\|^2}{\left\|z - z\widetilde{T}_{\mathbf{w}}^\top\left(\widetilde{T}_{\mathbf{w}}\widetilde{T}_{\mathbf{w}}^\top + \lambda I\right)^{-1}\widetilde{T}_{\mathbf{w}}\right\|^2}. \tag{2.13}$$

Fortunately the numerator in Eq. (2.13) does not depend on $\mathbf{w}$. Thus, maximizing $\mathcal{G}_{\phi_{\mathbf{w}}\to z}$ with respect to $\mathbf{w}$ reduces to minimizing the following cost function:

$$\mathcal{L}(\mathbf{w}) = \left\|z - z\widetilde{T}_{\mathbf{w}}^\top\left(\widetilde{T}_{\mathbf{w}}\widetilde{T}_{\mathbf{w}}^\top + \lambda I\right)^{-1}\widetilde{T}_{\mathbf{w}}\right\|^2. \tag{2.14}$$

This objective is a non-convex, higher order nonlinear function of the spatial filter $\mathbf{w}$ and a local minimum can be obtained by means of standard nonlinear optimization techniques such as the L-BFGS algorithm. We use the minFunc function of Mark Schmidt (2012). For the simulation and real world data analysis shown in this paper we restart the procedure five times with different random initializations. The gradient of Eq. (2.14) is given in Appendix A.

Optimizing the objective function stated in Eq. (2.14) yields a single weight vector that extracts neural source activity. If desired, further weight vectors can be obtained using a so-called *deflation scheme* which is also outlined in Appendix A.

Two parameters have to be set at the start of the optimization: the number of time lags P and the regularization parameter λ. We select the number of lags P as the optimizer of Schwarz's Bayesian Information Criterion (BIC) (Schwarz, 1978; Schneider and Neumaier, 2001) when fitting the restricted model given in Eq. (2.1). The number of lags P which optimizes the BIC will be a number as small as possible which still achieves a low residual variance. The regularization parameter λ used in Ridge Regression is needed to prevent overfitting on the training data. We select λ in a 5-fold (nested) chronological cross-validation procedure (see (Lemm et al., 2011), for instance) from $\{0, 10^3, 10^6\}$ in the simulation study and from $\{0, 10^3, 10^6, 10^9\}$ in the real EEG analysis.

The runtime depends on the number of sensors M, the number of epochs N, and the number of time lags P. On the realistic dimensionality of our simulation study ($M = 26$, $N = 600$, $P = 5$) and for a given regularization parameter λ, the runtime is in the order of seconds only. However, λ has to be estimated by cross-validation. The 5-fold cross-validation procedure from the simulation study took an average processing time of 60.0 s (10.8 second standard deviation) on a 3.6 GHz cluster node.

Computation of the spatial pattern

While the spatial filter $\mathbf{w}$ extracts a source $\hat{s}^* = \mathbf{w}^\top X$, we need to derive its corresponding spatial pattern $\hat{\mathbf{a}}^*$ to be able to interpret the neurophysiological origin of this source. The pattern $\hat{\mathbf{a}}^*$ gives an estimate of $\mathbf{a}^*$ in the generative model in Eq. (2.5) and contains the projection strength of $\hat{s}^*$ onto the scalp sensors. Visualizing the spatial filter $\mathbf{w}$ can lead to misinterpretations, because filters aim to suppress noise sources and thus depends on the spatial distribution of signal and noise sources (Blankertz et al., 2011; Haufe et al., 2014b). The spatial pattern $\hat{\mathbf{a}}^*$ is given by

$$\hat{\mathbf{a}}^* = \mathrm{Cov}(X)\cdot\mathbf{w}, \tag{2.15}$$

where $\mathrm{Cov}(X)$ denotes the covariance of X (Haufe et al., 2014b; Parra et al., 2005; Blankertz et al., 2011).

Simulations

We compare the ability of channel-wise Granger causality testing, ICA, SPoC and GrangerCPA to identify oscillatory signals whose power modulation Granger causes an external target variable. We first carry out simulations, where the target source is known, and we can compare the extracted source to the ground truth.

Data generation

We simulated epoched EEG recordings with 26 channels according to the generative model in Eq. (2.5). A target variable z and epoch-wise log bandpower ϕ^* which Granger causes z were generated from a stable bivariate AR process of order $P = 5$. The oscillatory target source s^* with log power modulation ϕ^* and 150 additional oscillatory sources were created and mixed into the simulated EEG. The signal-to-noise ratio between them is controlled by a parameter denoted γ which is approximately the percentage of variance explained by the target source s^*. The simulation protocol is outlined in more detail in Appendix B.

Simulation 1: one Granger-causal source

In Simulation 1, we compare the performance of channel-wise Granger causality testing, ICA, SPoC and GrangerCPA in different noise settings. We vary the variance explained by the target source γ while keeping the number of training epochs N fixed at 600. Furthermore, we vary the number of training epochs while keeping γ fixed at 6 %. For each value of γ and N we generate 200 data sets. Using the setting of $N = 600$ and $\gamma = 6\%$ we also analyze the effect of noise on the external target variable z. We add varying degrees of measurement noise κ according to $z + \kappa\cdot\epsilon$, where $\epsilon\sim\mathcal{N}(0,1)$ and z has been standardized to variance 1.

In our noisy simulations, a perfect reconstruction of the target source s^* is in general not possible. While s^* is not known in practice, we use it in this simulation to quantify an upper limit of possible source reconstruction accuracy, which we call 'MSE optimum'. We compute a filter

which yields a minimal mean square error (MSE) between the time course of the target source s^* and its reconstruction from the EEG data $\mathbf{w}^T X$. This is done by Ordinary Least Squares regression of s^* on the EEG data X.

While SPoC and GrangerCPA directly compute a weight vector which optimizes their respective objective functions, FastICA yields an entire set of unordered weight vectors with no information about which one extracts the target source. We identify the target source on the training data by selecting the component whose power time course shows strongest Granger causality to the target variable. Similarly, for the channel-wise analysis, we select the channel whose power time course shows strongest Granger causality to the target variable.

We measure the performance of the five methods on 100 unseen test epochs as (1) the mean correlation between the log power of the recovered and the true source and (2) the frequency of detecting the Granger-causal relationship.

Simulation 2: one Granger-causal and one correlated source

We conduct a second simulation to analyze the ability of channel-wise Granger causality, ICA, SPoC and GrangerCPA to identify causal sources when a confounding, merely correlated source is also present. To this end, we include a source s^c whose log power is perfectly correlated with the target variable z to the components of the artificial EEG. Note that this perfectly correlated source is not Granger-causal to z as it does not contain any additional information about z. Therefore, we expect GrangerCPA to identify the target source s^*, and SPoC to find the higher correlated source s^c. We investigate a setting of 600 training epochs in which the Granger causal source s^* and correlated source s^c each explain 6% of EEG variance.

Simulation 3: three Granger-causal sources

In a third simulation, we compare FastICA, deflation-based SPoC and deflation-based GrangerCPA using simulated EEG data in which three underlying sources are Granger causally connected to the external target variable. To do so, we generate the target variable z and epoch-wise log bandpower ϕ^*_A, ϕ^*_B, ϕ^*_C from a stable 4-dimensional AR process of order $P = 5$. By allowing the corresponding AR coefficients to be non-zero, we investigate the following two scenarios: (1) ϕ^*_A, ϕ^*_B, and ϕ^*_C are pairwise independent, but each Granger-causes z. (2) ϕ^*_C is Granger causing ϕ^*_B, ϕ^*_B is Granger causing ϕ^*_A, and ϕ^*_A is Granger causing z. For each scenario, we investigate a setting of 600 training epochs in which the three target sources each explain 6% of EEG variance.

Real EEG data

Experiment 1: self-paced braking

We apply channel-wise Granger causality testing, ICA, SPoC and GrangerCPA on data from 18 subjects performing self-paced foot movements. It is well known that brief, self-paced movements of specific body parts are preceded by an event-related desynchronization (ERD) of 8–13 Hz (mu band) and 15–30 Hz (beta band) rhythms over corresponding sensorimotor areas (Neuper and Pfurtscheller, 2001). Here we focus on beta ERD which starts as much as 1.5–2 s prior to movement onset and is thought to be related to movement preparation and execution (Pfurtscheller, 1981; Kilavik et al., 2013).

The experimental data we use here stems from a pre-measurement of a simulated driving experiment described in Haufe et al. (2011). 18 healthy participants were instructed to perform self-paced right foot movement (i.e. press the brake pedal) once per second for 5 min. EEG data were recorded with 64 Ag/AgCl electrodes at 1000 Hz. Furthermore, an electromyographic (EMG) signal was recorded using a bipolar montage at the tibialis anterior muscle and the knee of the right leg.

For the following offline-analysis, EEG data were decimated to 200 Hz. Artifactual electrodes were rejected using a variance criterion. As a subsequent preprocessing, the dimensionality of the data was reduced to 10 using spatio-spectral decomposition (SSD) (Nikulin et al., 2011). SSD suppresses artifacts and increases the signal-to-noise ratio of neuronal oscillations by maximizing the signal power in the frequency band of interest, here the beta band, while simultaneously minimizing it at neighboring frequencies. This makes SSD an effective tool for preprocessing, because it extracts a low dimensional subspace which captures oscillatory activity in the frequency range of interest (Haufe et al., 2014a). The choice of 10 SSD components was based on experience.

After dimensionality reduction, EEG data were band-pass filtered between 15 and 30 Hz (5th-order causal Butterworth filter) and segmented into consecutive epochs of 20 ms length yielding on average 17 880 epochs per subject. Similarly, the EMG signal was band-pass filtered between 2 and 45 Hz and segmented into consecutive epochs of 20 ms length. We use the logarithm of the variance in each EMG epoch as the target variable z, thus yielding a 50 Hz resolution for the target function as well as the EEG power time-series.

We test the ability to detect a motor preparation source of SPoC, GrangerCPA, FastICA and channel-wise statistical testing in a 5-fold chronological cross-validation procedure (see e.g. Lemm et al., 2011). On the training folds, each method estimates a spatial filter as described in Section 2.5. On the test fold, we test whether the log power of the estimated source Granger causes the target variable z. On each test fold, Granger causality is evaluated at a lag order which is chosen as the optimizer of Schwarz's Bayesian Information Criterion (BIC) (Schwarz, 1978). We check for residual autocorrelation using a Ljung–Box test (Ljung and Box, 1978), and we increase the lag order in case of a remaining significant autocorrelation at $p < 0.05$. The Ljung–Box statistic was based on log(number of epochs) time lags, as recommended by Tsay (2005).

The training/test split was repeated such that each fold became the test fold once, yielding a p-value for each split. These five p-values were then combined using the Z-transform test (Whitlock, 2005; Stouffer et al., 1949). The Z-transform test converts the p-values from each of the five tests, p_j, into standard normal deviates Z_j. Under the null hypothesis of no Granger causality the sum of the five Z_j's is normal distributed with mean zero and standard deviation $\sqrt{5}$. The final p-value of cumulative evidence for Granger causality over the five folds is thus given by comparing the test statistic $T_Z = \sum_{j=1}^{5} Z_j/\sqrt{5}$ to the standard normal distribution.

Experiment 2: predicting reaction times

In the second real-world example, we demonstrate how Granger causality analysis can be used to find oscillatory alpha components that are predictive of reaction times. A large literature links performance with ongoing alpha oscillations and reports effects that vary extensively depending on the investigated task. Low reference alpha power is related to low performance in cognitive and memory tasks (Klimesch et al., 1999), but is also often associated with high perception performance (Klimesch et al., 2007; Jensen and Mazaheri, 2010).

Here, we apply channel-wise Granger causality, ICA, SPoC and GrangerCPA to data from an auditory perception task. While the ground truth is not known, we expect to obtain physiological interpretable patterns and we will compare the methods' abilities to discover significant effects.

The behavioral and EEG data were recorded in series of studies where participants were asked to assess the signal quality of auditory signals (Antons et al., 2012; Porbadnigk et al., 2013). For the given study, nine participants were presented with a forced choice task, where they had to indicate by button press whether the presented speech stimulus was of high quality or degraded. As stimulus material, two words ('Haus' (engl. house) and 'Schild' (engl. sign)) were used, spoken by two different speakers. These stimuli were presented in an oddball paradigm, with a majority of stimuli in wideband quality (non-targets), interspersed with stimuli that were degraded with four different levels of bit rate reduction (targets). The inter-stimulus

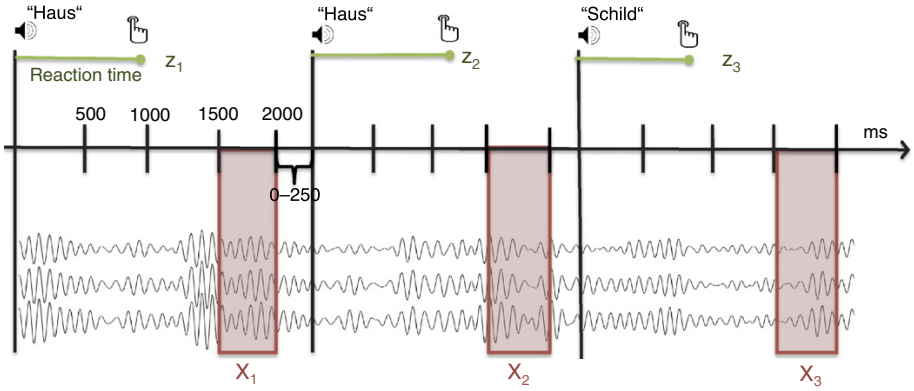


Fig. 3. Schematic illustration of Experiment 2. Participants had to rate whether the presented words ('Haus', 'Schild') were of maximal quality or degraded by bit rate reduction. We use the reaction time as a target variable, $z = (z_1, z_2, z_3 \ldots, z_N)$, which we try to predict using epoched, alpha band-pass filtered EEG data $X = (X_1, X_2, X_3, \ldots, X_N)$. Finding a source whose log power is Granger causing z means that the log power of that source in X_1 and X_2 in addition to past reaction times z_1 and z_2 predicts z_3 better than z_1 and z_2 alone.

interval varied between 2000 and 2250 ms. An experimental session lasted approximately 1.5 h (plus additional time for electrode application and removal) during which on average 2388 stimuli were presented. EEG data were recorded using a 64-channel EEG system at 1000 Hz.

For offline analysis, EEG data were downsampled to 100 Hz and artifactual electrodes were rejected using a variance criterion. The alpha frequency band was individually defined in relation to the individual alpha peak frequency (IAF) from IAF − 3 Hz to IAF + 2 Hz (Doppelmayr et al., 1999; Klimesch, 1999). SSD was used to project the data onto a 10 dimensional subspace that captures oscillatory activity in the individual alpha band.

Subsequently, EEG data were band-pass filtered in the individual alpha band and divided into epochs ranging from 1500 ms to 2000 ms with respect to stimulus onset. The reaction times are used as the target variable z. We apply channel-wise Granger causality, ICA, SPoC and GrangerCPA to test whether we can identify sources whose alpha power predicts reaction times above and beyond its own past values. The experimental setup is illustrated in Fig. 3. Given the inter-stimulus interval of about 2 s, the resulting time series of reaction times as well as the EEG (channel-/component-) power time-courses have a sampling rate of about 0.5 Hz. Consequently, a single time-lag corresponds to about 2 s. For significance testing, we use the same 5-fold cross-validation scheme as in Experiment 1.

Results

Simulations

Simulation 1: one Granger-causal source

Fig. 4 depicts the scalp plots obtained from an example simulation run using $N = 600$ training epochs with a target source explaining $\gamma = 6\%$ of EEG variance. Note how little the channel-wise Granger causality p-values resembles the spatial pattern of the true target source. The p-values for each electrode thus do not recover the pattern. On average over 200 simulations, the mean correlation of the estimated pattern with the true pattern is 0.94 for GrangerCPA, 0.87 for SPoC, 0.84 for FastICA, but only 0.29 for the channel-wise Granger-causality p-values.

Fig. 5 shows the results from the simulations in which we vary the noise in the data and the available number of training epochs and test the ability of GrangerCPA, SPoC, FastICA, and electrode-wise Granger causality to recover the target source. All reported correlations where obtained on test data that was not used to train the algorithms. In order to assess the specificity of this approach, we additionally performed the simulation without any Granger causal source present ($\gamma = 0$, $N = 600$ training epochs, 200 runs). As is to be expected, the performance of the four methods did not differ significantly from each other ($p > 0.05$, Wilcoxon signed rank tests on the z-transformed achieved p-values). A Granger causal source was falsely detected with $p < 0.01$ in 1.5% of runs by GrangerCPA, FastICA, and the most Granger-causal electrode, and in 0.5% of runs for SPoC. This is close to the theoretical expected value of 1%.

When the target source is present, the recovery of the source improves for all methods as the signal-to-noise ratio increases and more training epochs become available. For a target source explaining 6% of EEG variance, GrangerCPA outperforms the other three methods when at least 200 training epochs are available. For all methods, at least 200 training epochs are needed for a good reconstruction of the target source. For GrangerCPA, SPoC and FastICA at least 500 training epochs are recommendable. Out of all methods, FastICA is the most vulnerable to an insufficient amount of training data, while channel-wise Granger causality testing needs the least amount of data to reach its (low) peak performance. As the number of training epochs increases, GrangerCPA plateaus earlier than SPoC and FastICA.

For a target source explaining 6% of EEG variance, added measurement noise on the external target variable z has only a small effect for moderate noise levels of $\kappa \leq 20$ %, with GrangerCPA being slightly more vulnerable than the compared methods. As the noise level

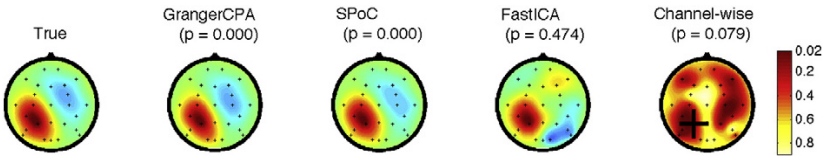


Fig. 4. Example simulation run using $N = 600$ training epochs and a target source explaining $\gamma = 6\%$ EEG variance. The left scalp map shows the true spatial pattern of the simulated target source. The second, third and fourth scalp maps show the estimated scalp pattern of GrangerCPA, SPoC and FastICA. The right scalp map shows the Granger-causality p-values between channel-wise log bandpower and the target variable estimated on the test data. The cross indicates the electrode chosen on the training data. For each method, the p-value related to the Null Hypothesis of no Granger causality from the extracted bandpower to the target variable is computed on the test data and presented above the scalp map.

increases, the Granger causal relationship of the target source starts to break down, and so does the performance of all the considered methods.

Concerning the signal-to-noise level γ of the target source, we can see that all three multivariate approaches outperform channel-wise statistical testing in all signal-to-noise regimes. GrangerCPA outperforms FastICA and channel-wise Granger causality in all signal-to-noise regimes, and SPoC for signal-to-noise levels above 1.5% of explained variance. This is due to the fact that the Granger causal target source power may only be very weakly correlated with the target variable, making it difficult for SPoC to identify it.

Fig. 6 highlights the difference between SPoC and GrangerCPA. For a source explaining 6% of EEG variance, we sorted the 200 simulation runs according to the correlation between the target variable and its prediction by past values of the true source power. In our simulation, this correlation varies as a function of the randomly generated

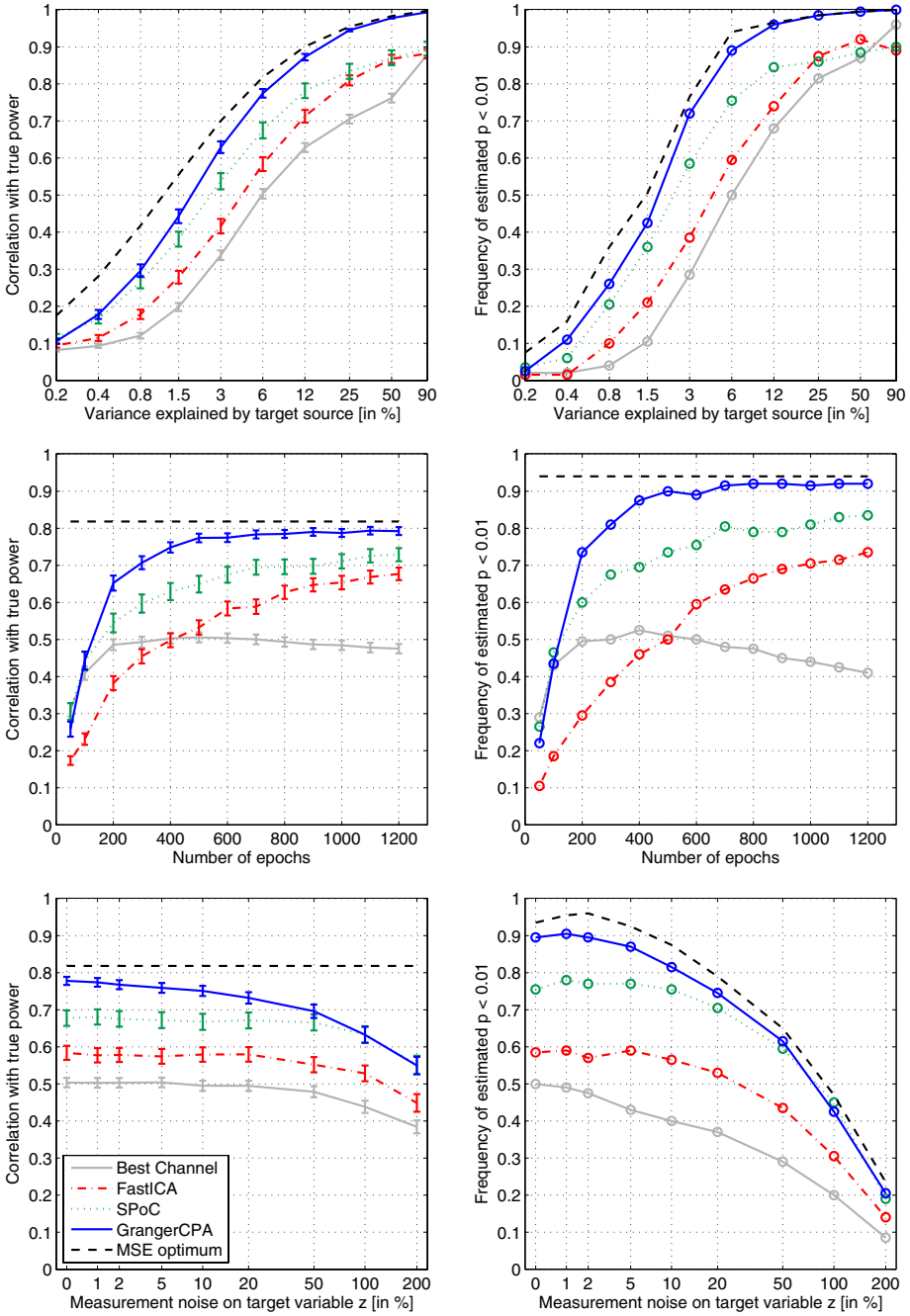


Fig. 5. Results of Simulation 1 in which we vary the variance explained by the target source γ (for $N = 600$), the number of training epochs N (for $\gamma = 6\%$) and add varying degrees of measurement noise κ to the target variable z (for $N = 600$, $\gamma = 6\%$). (1st column) Mean $\pm$ standard error of the correlation of the log power of the extracted source with the log power of the true source. (2nd column) frequency of estimated significant relationships.

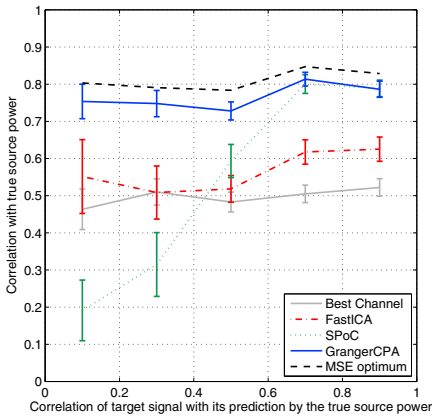


Fig. 6. Results of Simulation 1: Extraction performance – measured as the correlation of the estimated with the true source power – against the correlation of the target variable with its prediction by the true source power on 200 simulations runs with a target source explaining $\gamma = 6\%$ of EEG variance and using $N = 600$ training epochs.

coefficients in the bivariate AR model, and Fig. 6 plots the performance of channel-wise Granger causality testing, ICA, SPoC and GrangerCPA against this correlation. We can see that SPoC performs well if this correlation is high, but its performance declines when the correlation decreases. In contrast, the performance of GrangerCPA, ICA and the most Granger causal channel does not depend on this correlation.

This is to be expected because SPoC addresses the optimal extraction of a neural source whose power modulation is highly correlated with the target variable. However, a power modulation is Granger causal only if it predicts the target variable above what is predicted by its own past. For Granger causality, the bandpower itself does not need to be strongly correlated with the target variable.

Simulation 2: one Granger-causal source and one correlated source

Fig. 7 depicts the result for our second simulation in which both the target source (whose log power Granger causes the target variable) as well as a confounding non-target source (whose log power is perfectly correlated with the target variable) were mixed into the simulated EEG. As expected, GrangerCPA extracted the target source in the majority of cases (170 out of 200 simulation runs), while SPoC extracted the correlated non-target source in 150 out of 200 runs. FastICA and the most Granger causal channel are also not mislead by the correlated source, however they identify the target source in much fewer cases than GrangerCPA.

Simulation 3: three Granger-causal sources

Fig. 8 shows the results for the third simulation in which three target sources were mixed into the simulated EEG. We see that in the scenario of three sources whose power independently Granger causes z with equal strength, GrangerCPA is able to retrieve all three sources and their topographies with higher accuracy than SPoC or FastICA. SPoC is more affected by the presence of several Granger causal sources, with its performance now similar to FastICA. In comparison to the results with one single target source, GrangerCPA's performance decreases slightly (cf. Fig. 5). This might be due to the fact that each of the three sources exhibits weaker Granger causality to the target variable z than one single Granger-causal source, because the frequency of detectable Granger causal relationships is also reduced.

In the second scenario, source power C Granger-causes source power B, B Granger-causes source power A, and A Granger-causes the external target variable z. Consequently, the three sources do not Granger cause z with equal strength, as is reflected by the decreasing frequency of detectable Granger causal relationships. While GrangerCPA and SPoC are able to reliably extract a first source, their performance decreases for the 2nd and the 3rd source. GrangerCPA slightly outperforms SPoC and FastICA.

Real EEG data

Experiment 1: self-paced braking

Fig. 9 shows the results of applying channel-wise Granger causality, ICA, SPoC and GrangerCPA to EEG data from 18 subjects performing self-paced braking. For each subject, the spatial activations pattern of the source extracted by each method is displayed and the result of Granger causality significance testing is indicated. Subjects are ordered by the strength of Event-Related Desynchronization (ERD) over central electrodes prior to EMG peak activity. We can see that, while only the first 5 subjects have a prominent ERD pattern over motor areas, both GrangerCPA and FastICA are able to extract a motor-related source in 11 out 18 subjects. These sources are characterized by neurophysiologically plausible patterns which suggest a source located in motor foot-related areas and a power time course that significantly Granger causes motor activity at $p < 0.01$ (i.e. predicts it better than what is predicted by its own past). Granger causality analysis was based on average on 8 time lags, corresponding to 160 ms.

For all spatial filtering methods, the correspondence between neurophysiologically plausible patterns and Granger causality of the extracted sources is high: Sources whose power significantly Granger causes motor activity are characterized by patterns that indicate a foot-related area, while sources associated with unclassifiable or artifactual patterns do not significantly Granger cause the EMG. The channel-

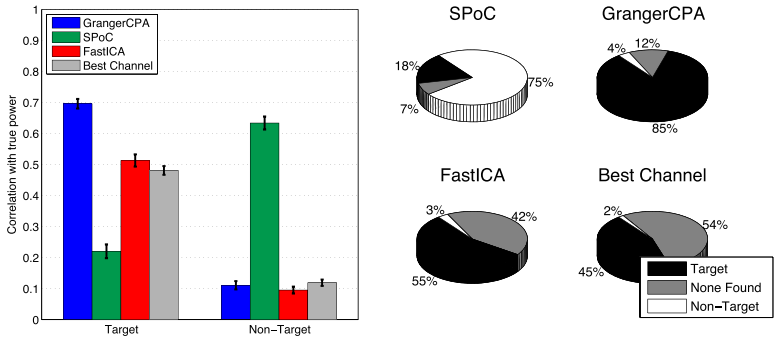


Fig. 7. Results of Simulation 2. Both a target source (whose log power Granger causes the target variable) as well as a confounding non-target source (whose log power is perfectly correlated with the target variable) were mixed into EEG toy data. (Left) Mean ± standard error of the correlation of the log power of the extracted source with the log power of the target and the non-target source. (Right) Pie charts showing which source was identified by the algorithms. The identified source was classified as 'no source extracted' when its estimated log power correlated with neither the target nor the non-target source above 0.5.

Table 1
Test statistic of cumulative evidence T_Z for Granger causality according to the *F*-test in both real EEG experiments (mean ± standard error): (a) for subjects for whom at least one method found a significant effect, and (b) for all subjects. A more negative test statistic T_Z indicates higher Granger causality.

		GrangerCPA	SPoC	FastICA	Best channel	SSD
Self-paced	(a)	−5.96 ± 1.3	−5.08 ± 1.6	−5.93 ± 1.4	−1.76 ± 0.4	−3.80 ± 1.0
	(b)	−3.95 ± 1.1	−3.37 ± 1.2	−4.42 ± 1.0	−1.30 ± 0.3	−2.62 ± 0.8
Reaction times	(a)	−2.69 ± 0.1	−1.94 ± 0.4	−1.64 ± 0.4	−1.01 ± 0.8	−1.41 ± 0.6
	(b)	−1.85 ± 0.4	−1.57 ± 0.3	−1.56 ± 0.3	−0.40 ± 0.5	−1.47 ± 0.4

wise significant values show little resemblance with the spatial activation patterns obtained using multivariate methods.

In the following, we restrict our analysis to the 12 subjects for whom at least one of the investigated methods was able to identify a significant motor source. We find that both GrangerCPA and FastICA yield a significantly larger additional effect of the power time course on EMG prediction than channel-wise statistical testing or the most Granger-causal filter from the SSD preprocessing ($p < 0.01$, Wilcoxon signed rank test on the test statistic T_Z). The performance of GrangerCPA, ICA and SPoC is statistically indistinguishable. The achieved Granger causality of each method is summarized in Table 1.

Nevertheless, we can identify subjects where GrangerCPA and FastICA find more meaningful sources than SPoC, notably subject s4: Both GrangerCPA and FastICA extract a motor related source, but SPoC gets stuck on artifactual activity. The source extracted by SPoC does not Granger cause the EMG signal, but is highly correlated with the EMG signal. Here, looking for a source that contains information beyond the autocorrelation of the target EMG signal, prevents us from extracting artifactual sources that reflect muscular activity. Note that EMG activity was significantly autocorrelated in each of the subjects, with past values of EMG explaining 45.8% to 73.9% of its variance.

To summarize, the results obtained in this analysis show that the multivariate approaches we analyzed significantly outperformed channel-wise statistical testing and reliably extract motor-related sources in subjects showing Event-Related Desynchronization prior to motor activity.

Experiment 2: predicting reaction times

Fig. 10 depicts the results obtained in the reaction time analysis of data acquired during an auditory perception task. For each subject, the spatial patterns obtained by FastICA, SPoC and GrangerCPA are displayed and the results of Granger causality significance testing are indicated. It can be seen that GrangerCPA identifies a Granger causal source in seven out of nine subjects at $p < 0.05$. In contrast, SPoC and ICA were able to extract sources whose power significantly Granger causes reaction times in three resp. four subjects. In sensor space, channel-wise Granger causality testing was only able to extract a Granger causal source in two subjects. The achieved Granger causality of each method is summarized in Table 1.

GrangerCPA finds significant effects for the same subjects as the other methods. However, it is capable to also identify additional ones. Note that reaction times were significantly autocorrelated in each of the subjects, with past reaction times explaining 1.9% to 37.2% of its variance.

The majority of spatial patterns seem neurophysiologically plausible and suggest sources in the parietal and occipital areas. Please note that the polarity of the spatial patterns is arbitrary. For each pattern, the polarity of the pattern was set such that the maximal activity is positive.

In 6 out of the 7 subjects in which GrangerCPA extracted a source with significant Granger causality, high pre-stimulus alpha power predicted a slow reaction. This is in line with the results of SPoC and FastICA, only that GrangerCPA reveals this effect for additional

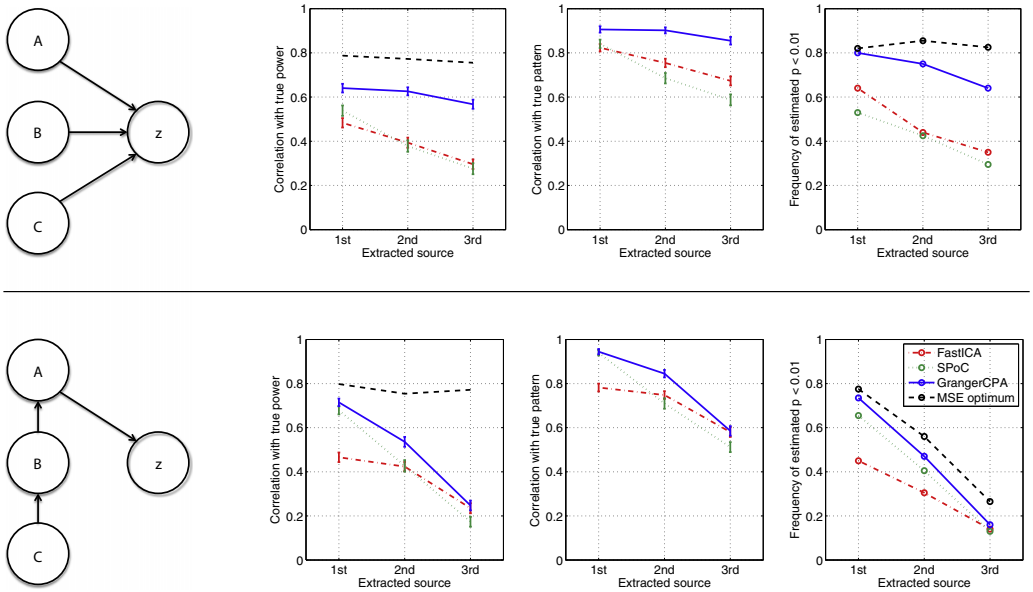


Fig. 8. Results of Simulation 3. Three sources *A*, *B*, and *C* are Granger-causally connected to the target variable *z* in two different schemes (1st column). For each scheme, the three most Granger-causal FastICA components, and the first three extracted sources of SPoC and GrangerCPA are evaluated in terms of their correlation to the true source log power (2nd column), their correlation to the true scalp topography/pattern (3rd column), and the frequency of detected Granger causal relationship (4th column).

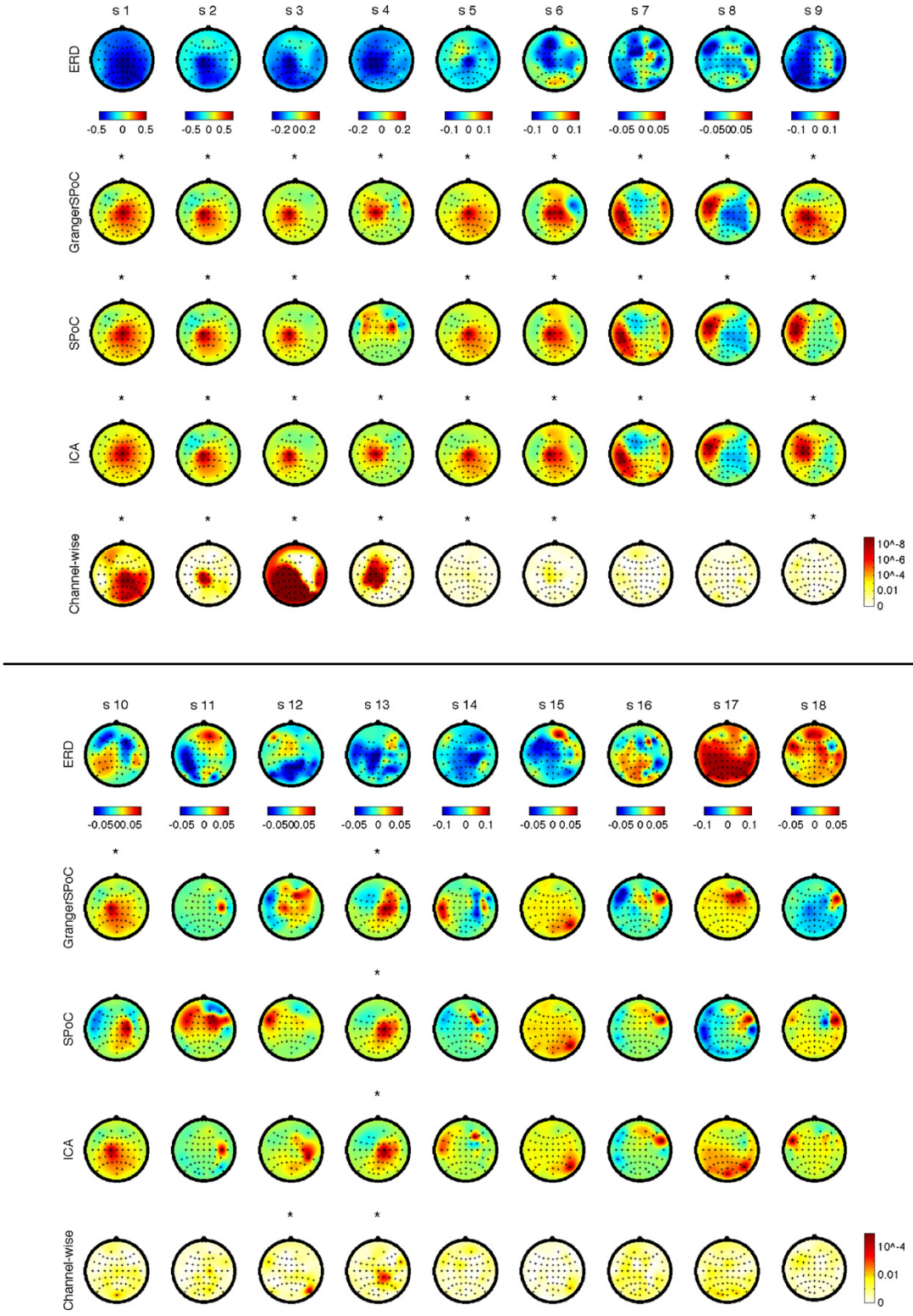


Fig. 9. Results on Self-paced Braking data. Each column shows the result for a single participant. Participants are sorted according to the strength of Event-Related Desynchronization (ERD) over central electrodes. 1st row: ERD in the [−400 ms, −200 ms] interval using EMG as a trigger. 2nd–4th row: Spatial patterns as obtained by GrangerCPA, SPoC and FastICA respectively. 5th row: Channel-wise p-values of Granger causality significance test. If estimated alpha power significantly Granger causes the EMG signal at $p < 0.01$ according to the permutation test, the corresponding pattern is marked with a star.

participants. Moreover, subject s7 which displays the opposite effect, also has a deviant component topography over the motor cortex, rather than the occipito-parietal cortex. Note that for each subject and each method, Granger causality testing was based on average on two time lags. This corresponds to information from the past 3 s, as only the last 500 ms of each epoch with length 2 s is used (cf. Fig. 3). For the extracted sources whose power significantly Granger causes reaction times based on several lags, all lags pointed in the same direction except for subject 8.

Summarizing, the analysis of reaction times during an auditory perception task reveals that reaction times can be predicted by pre-stimulus individual alpha power on a single-trial basis in the majority of the subjects. GrangerCPA was able to discover such a significant relationship in considerably more subjects than any other method tested here.

Discussion

In this paper, we have analyzed how to identify Granger causal links from brain oscillations to target variables of interest. We have shown that computing Granger causality in sensor space suffers from poor signal-to-noise ratio, while multivariate spatial filtering approaches such as ICA or SPoC alleviate this issue. Moreover, we presented a novel method called GrangerCPA which optimizes for Granger causality, and demonstrated its ability to reliably extract oscillations that Granger cause a given external target variable. Possible application scenarios include a wide range of target variables, such as behavioral output (e.g., reaction times, sensory detection, task rating, evoked potentials), another physiological measure (e.g., muscular activity, heart rate variability) or a second power time course.

Comparison of methods

We compared channel-wise Granger causality analysis with three spatial filtering methods: FastICA, SPoC, and our newly proposed GrangerCPA algorithm. Particularly striking is the poor performance of channel-wise Granger causality testing. While neuronal oscillations are often analyzed in sensor space, this approach disregards the physics of EEG and yields a poor signal-to-noise ratio.

For the spatial filtering approaches, comparison must be undertaken in more detail. SPoC is a method which reliably extracts a source whose band-power maximally correlates with the target variable. Thus, if a source is both Granger causal and highly correlated with the target variable, SPoC will identify a causal source. However, if the Granger causal source is only weakly correlated with the target variable or if a second merely correlated source is present, SPoC fails to recover the causal source.

ICA algorithms demix EEG data under the assumption that the underlying sources are independent of each other and have been successfully applied to a wide range of EEG (as well as MEG) data analysis problems. GrangerCPA outperformed FastICA both in our simulated EEG data and on the reaction time EEG data set, while both methods performed equally well on the self-paced braking EEG data set. Thus, we see that both algorithms may arrive at similar solutions using an entirely different set of assumptions. Yet, ICA algorithms are not explicitly designed to find power modulations which Granger cause a given external target variable. ICA essentially finds the most non-Gaussian components. Thus, if the source of interest is approximately Gaussian distributed, ICA will be less successful. Importantly, this can be the case for amplitude-modulated oscillations. An interesting remedy, which we have not explored further here, consists of applying ICA to short-time Fourier transforms of EEG signals (Hyvärinen et al., 2010). ICA optimization then translates into optimizing the sparseness of the Fourier coefficients, which should separate oscillatory signals at different frequencies. As ICA, Fourier-ICA is an unsupervised approach that does not make use of the information in the target variable.

Experimental results on real EEG data

The first experiment (self-paced braking) served as a proof of concept. We show that spatial filtering approaches outperform channel-wise statistical testing and that all three spatial filtering approaches reveal significant results with similar statistical power and for the same subjects. We could show that sources whose power significantly Granger causes the EMG show neurophysiologically plausible motor-related patterns. In contrast, components with patterns that are ambiguous or artifactual do not significantly Granger cause the motor activity.

In the second experiment, we applied Granger causal analysis to find oscillatory alpha components that are predictive of reaction times. As reaction times are autocorrelated (cf. (Aue et al., 2009)), we can hope to identify brain activity which reflect this variation. Indeed it has been found that ongoing pre-stimulus alpha power modulates perception performance (see (Jensen and Mazaheri, 2010) for a review), and is negatively correlated with subjective attentional state (Macdonald et al., 2011). Recent studies also demonstrate increased alpha band activity up to 20 resp. 10 s prior to the occurrence of an error (O'Connell et al., 2009; Martel et al., 2014).

The reaction time experiment shows that GrangerCPA can be more sensitive than the other spatial filtering methods. GrangerCPA finds statistically significant effects for the same participants as the other methods, but is also able to reveal effects in additional participants. The direction of the identified effect is the same across all participants except one (s7), who also shows a different pattern. We find that high pre-stimulus alpha power predicts slow reaction times. This is in line with findings that relate low pre-stimulus alpha to high perception performance of visual (Klimesch et al., 2007; van Dijk et al., 2008; Romei et al., 2010) or somatosensory stimuli (Schubert et al., 2009). However, the literature is not consistent in this respect (Linkenkaer-Hansen et al., 2004; Babiloni et al., 2006). GrangerCPA may prove helpful in this respect, allowing to (a) investigate such effects with higher statistical power and (b) to take dependencies of previous trials into account and (c) to find out how far back into time this dependency lasts. Concerning the latter, GrangerCPA reveals a time lag of on average two trials for the given data set, suggesting that the effect is not caused by longer-term changes in vigilance, but rather by the immediate past which may reflect modulation of attentional processes.

Application scenarios and limitations

Most work on Granger causality in the field of neuroscience has focused on understanding interactions of activated brain areas (e.g., (Astolfi et al., 2007; Haufe et al., 2010, 2013; Michalareas et al., 2013)). As sensor-level connectivity analysis is severely limited by volume conduction, it has often been suggested to study neuronal interactions in source space (Schoffelen and Gross, 2009; Gómez-Herrero et al., 2008). Conveniently, multivariate AR models in sensor space can be related analytically to AR models in source space through an inverse source reconstruction operator (Michalareas et al., 2013).

In contrast to studying neuronal interactions, here we relate amplitude modulated oscillations in one brain area to cognitive performance or other target variables. This may allow us to gather further crucial evidence to answer the question of whether rhythmic brain activity is causally shaping (and not merely correlating with) perception and cognition.

Causal and correlative relationships between two variables are not mutually exclusive, as the co-occurrence of both can be caused by auto-correlation. Thus, sources whose bandpower is highly correlated with a target variable may also have a causal influence on the target variable. As an example, consider an oscillatory beta motor source, which drives peripheral muscular activity as measured by the EMG. Now, if beta power is auto-correlated, past beta power will influence both the

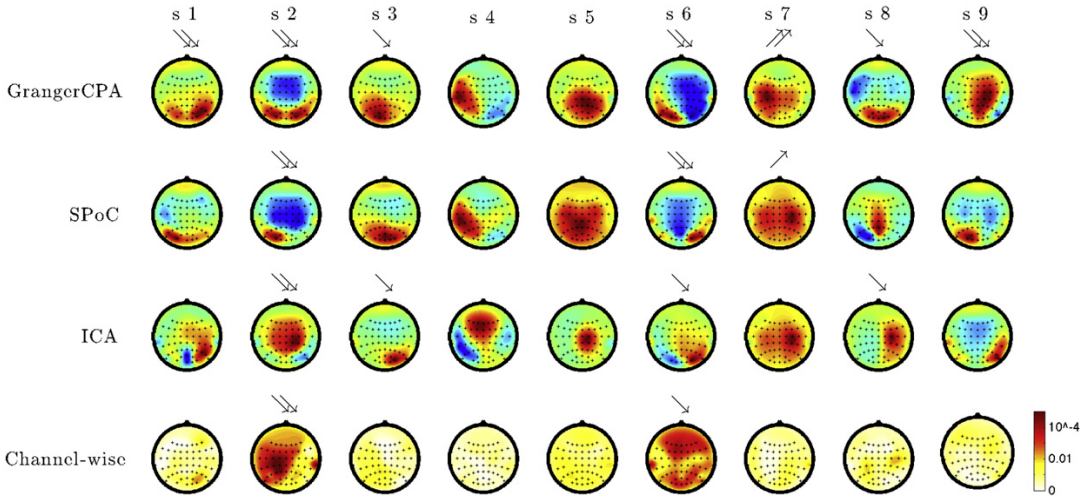


Fig. 10. Results of reaction time analysis: Spatial patterns as obtained by GrangerCPA, SPoC, and FastICA, respectively, and channel-wise *p*-values of Granger causality significance test. Each column shows the result for a single participant. Sources whose alpha power significantly Granger cause reaction time according to the permutation test are marked with one ($p < 0.05$) or two ($p < 0.01$) arrows. ↘ indicates that high pre-stimulus alpha power correctly predicts a slow reaction. ↗ indicates that high pre-stimulus alpha power correctly predicts a fast reaction.

present EMG signal as well as the present beta power, thereby inducing a correlation between both. While it is mathematically possible to construct cases in which a variable Granger-causes another variable without being correlated to it, these are special cases that may or may not arise in physiology. In fact, the power dynamics of neural oscillations exhibit strong auto-correlations up to the range of minutes (Linkenkaer-Hansen et al., 2001; Nikulin and Brismar, 2005). It is therefore likely that a neural source's power that Granger-causes a target variable will also be correlated with that target variable.

However, the framework of Granger causal analysis requires temporal precedence by definition. Since solely instantaneous effects cannot be interpreted as causal, they are explicitly excluded, while non-instantaneous effects are the focus of our considerations. Furthermore, in order to use the power dynamics of extracted sources to *predict* the target variable, temporal precedence is necessary.

Compared to stimulation techniques such as Transcranial Magnetic Stimulation, the advantage of the Granger causality approach is that it is simple and does not require intervention in the nervous system. However, it provides weaker evidence and suffers from some limitations: First, temporal precedence does not necessarily imply causality.[1] Furthermore, as long as not all relevant variables are incorporated into the autoregressive model, one cannot detect whether brain activity and target variable are driven by a third common cause. Last but not least, spurious results of Granger causality analysis can occur due to noise, because two correlated signals superimposed with noise will help to predict each other (Nalatore et al., 2007; Nolte et al., 2008). Nevertheless, we consider Granger causality to be a useful tool if these critical points are kept in mind. Moreover, some of the mentioned shortcomings can be alleviated. GrangerCPA can easily be extended to find sources predictive above and beyond several variables, and more noise-robust causality statistics such as PSI (Nolte et al., 2008) or time inversion testing (Haufe et al., 2012, 2013) can be applied to the extracted source.

At the very least, a Granger-causality analysis tells us whether an extracted neural source contains useful information for improving the predictions of another variable. In some applications, this may be all we care about. For example, a growing field of research tries to continuously monitor fatigue, attention, task engagement and mental workload in operational environments (Gevins et al., 1995; Berka et al., 2008; Müller et al., 2008; Blankertz et al., 2010) or during driving (Lal and Craig, 2001; Haufe et al., 2011; Dong et al., 2011). Here, it may be desirable to find neural oscillations that are maximally predictive beyond what can be predicted by other physiological sources and past task performance.

Another potential application scenario is Brain–Computer Interfaces (BCIs). Here an actively researched question is whether oscillatory sources influence the control performance of a BCIs user during a BCIs experiment (Grosse-Wentrup et al., 2011; Maeder et al., 2012).

Finally, the concept of Granger causality can be applied in the context of functional connectivity to improve our understanding of the concerted activity of spatially distinct, yet functionally connected brain areas. This is of particular interest with respect to bandpower dynamics of neural oscillations, as these have been shown to form functional networks defined via correlation of bandpower time courses (Hipp et al., 2012; Engel et al., 2013). Given a source bandpower time course of interest obtained from, say, an unsupervised source separation method such as ICA or SSD, GrangerCPA can be used to find a corresponding source the bandpower of which maximally predicts the bandpower of the first source. Such an analysis could help to unravel causal relations between functional networks in the brain.

Future research

Future effort will be required to directly optimize statistics such as PSI (Nolte et al., 2008), or to extend GrangerCPA to learn a weighting of multivariate target variables such as NIRS or fMRI measurements. Optimizing the weighting of a multivariate target variable is technically challenging, because our current optimization is simplified by the fact that the target variable is fixed (see Eq. (2.13)). Similarly, it would be interesting to optimize spatial filters for the extraction of sources, the power dynamics of which are causally influenced by an external target variable. This represents the opposite causal direction of the GrangerCPA method presented in this paper.

An open question is whether it is possible to reliably identify and infer the causality structure of several sources. Here we presented a preliminary simulation, in which three sources A, B, and C were causally connected to the target variable z in two different schemes, [$A \to z$, $B \to z$, $C \to z$] and [$C \to B \to A \to z$]. In the first case, GrangerCPA was often able to identify all three sources, while in the second scenario,

[1] Consider the following example from economics: interest rates or consumer confidence ratings predict economic development, but they reflect human forward-looking behavior and a causal relationship cannot be inferred (Hamilton, 1994).

performance decreased from *A* to *B* to *C*. The extraction performance might improve if optimization was based on multivariate AR modeling, rather than our deflation-based approach. However, the question that should be tackled first is whether we can distinguish both scenarios in case we found all three sources. While it is straightforward to apply post-hoc multivariate AR modeling, *A*, *B*, and *C* will be estimated with noise. The remaining volume artifact may still generate spurious causality (Schoffelen and Gross, 2009; Haufe et al., 2013), the extent of which will have to be analyzed in the future. Proposed remedies have so far focused on two variables (Nolte et al., 2008; Vicente et al., 2011; Haufe et al., 2013), but might be applied pairwise. The development of multivariate causality techniques more robust to noise and volume conduction remains an important challenge in the future.

Finally, the data sets used in this paper were from EEG experiments only. While we are optimistic that the results we have obtained carry over to other imaging methods that are sensitive to neural oscillations such as MEG or ECoG, further analysis is necessary to confirm this.

Conclusion

In summary, we analyzed which methods are best suited to reveal Granger causal links from brain oscillations to target variables of interest. Using both simulations and real EEG recordings, we showed that computing Granger causality on channel-wise spectral power suffers from poor signal-to-noise ratio and that multivariate spatial filtering approaches such as ICA or SPoC are suitable tools to alleviate this issue. Moreover, we presented a novel method called GrangerCPA, which directly optimizes for Granger causality. We believe that GrangerCPA will be a helpful tool for understanding the functional role of oscillations in electrophysiological recordings.

Acknowlegments

SH was partially supported by the German Federal Ministry of Education and Research (BMBF), grant no. 01GQ0850. AKP was supported by the German Federal Ministry of Education and Research (BMBF), FKZ 01GQ0850. SD acknowledges funding from the German Research Foundation (DFG) grant no. MU987/19-1. AK and SD were also supported by the DFG Research Training Group Sensory Computation in Neural Systems (GRK 1589/1). KRM acknowledges support by the Brain Korea 21 Plus Program as well as the SGER Grant 2014055911 through the National Research Foundation of Korea funded by the Ministry of Education. This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein.

Appendix A. GrangerCPA Algorithm

Gradient

The gradient of the non-convex cost function Eq. (2.14) is obtained as

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = -4\left(z - z\widetilde{T}_{\mathbf{w}}^{\top}\left(\widetilde{T}_{\mathbf{w}}\widetilde{T}_{\mathbf{w}}^{\top} + \lambda I\right)^{-1}\widetilde{T}_{\mathbf{w}}\right) \cdot \sum_{i=1}^{N}\left\{\left(\left(J^{i}\right)^{\top}\mathfrak{C} + \mathfrak{D}J^{i}z^{\top} - \mathfrak{D}\widetilde{T}_{\mathbf{w}}\left(J^{i}\right)^{\top}\mathfrak{C} - \mathfrak{D}J^{i}\widetilde{T}_{\mathbf{w}}^{\top}\mathfrak{C}\right) \cdot \left(\frac{X_i X_i^{\top}\mathbf{w}}{\mathbf{w}^{\top}X_i X_i^{\top}\mathbf{w}} - \frac{1}{N}\sum_{j=1}^{N}\frac{X_j X_j^{\top}\mathbf{w}}{\mathbf{w}^{\top}X_j X_j^{\top}\mathbf{w}}\right)\right\} \tag{A.1}$$

where $\mathfrak{C} := \left(\widetilde{T}_{\mathbf{w}}\widetilde{T}_{\mathbf{w}}^{\top} + \lambda I\right)^{-1}\widetilde{T}_{\mathbf{w}}z^{\top}$ and $\mathfrak{D} := \widetilde{T}_{\mathbf{w}}^{\top}\left(\widetilde{T}_{\mathbf{w}}\widetilde{T}_{\mathbf{w}}^{\top} + \lambda I\right)^{-1}$. $J^{i} \in \mathbb{R}^{2P \times N-P}$ denotes the vertical concatenation of a zero matrix of size $P \times (N-P)$ with a Toeplitz matrix with ones on the i-th descending diagonal and zero elsewhere, i.e. $J^{i}_{P+1,i-P+1} = J^{i}_{P+2,i-P} = \ldots = J^{i}_{2P,i} = 1$.

Extraction of further sources

Optimizing the objective function stated in Eq. (2.14) yields a single weight vector $\mathbf{w} \in \mathbb{R}^{M}$ that extracts neural source activity. If desired, further weight vectors can be obtained by a so-called *deflation scheme*. In such a scheme, we assume without loss of generality that the input data $X \in \mathbb{R}^{M \times (T \cdot N)}$ has been whitened, i.e. all instantaneous correlations between channels have been removed such that $XX^{\top} \propto I$. Then the optimization of Eq. (2.14) is performed in this 'whitened space'.

If q spatial filters $W = (\mathbf{w}_1, \ldots, \mathbf{w}_q) \in \mathbb{R}^{M \times q}$ have already been obtained, the next filter can be obtained according to the following steps:

1. Project X onto the space that is orthogonal to all previously obtained weight vectors. The required projection matrix $B \in R^{M \times (M-q)}$, $B^{\top}W = 0, B^{\top}B = I$, is the null space of $W^{\top}$ and can be computed by singular value decomposition. Denote with $X' = B^{\top}X \in \mathbb{R}^{(M-q) \times (T \cdot N)}$ the projected data.
2. Optimize Eq. (2.14) on the projected data X' to obtain $\mathbf{w}'_{q+1} \in \mathbb{R}^{M-q}$.
3. Project the resulting weight vector $\mathbf{w}'_{q+1}$ back into the original space, $\mathbf{w}_{q+1} = B\mathbf{w}'_{q+1} \in \mathbb{R}^{M}$.

The resulting filter now extracts the most Granger causal activity that is orthogonal to the previous extracted activity, i.e. it holds $\mathbf{w}_{q+1}^{\top}X = (\mathbf{w}'_{q+1})^{\top}X'$ and

$$\left(W^{\top}X\right) \cdot \left(\mathbf{w}_{q+1}^{\top}X\right)^{\top} = W^{\top}XX^{\top}\mathbf{w}_{q+1} = W^{\top}\mathbf{w}_{q+1} = W^{\top}B\mathbf{w}'_{q+1} = 0.$$

Appendix B. Simulated EEG

We simulated epoched EEG recordings with 26 electrodes according to the following steps. First, we generated the time course of the target variable $z \in \mathbb{R}^{N}$ and epoch-wise log band-power $\phi^{*} \in \mathbb{R}^{N}$ following a stable bivariate AR process of order $P = 5$. Granger-causality from ϕ^{*} to z was modeled by allowing the corresponding AR coefficients to be non-zero. Additionally, we generated the log band-power fluctuations of 150 additional sources according to univariate AR models of order 5.

Second, we constructed the band-pass filtered white noise sources $s^{*} \in \mathbb{R}^{T \cdot N}$ and $s_1, \ldots, s_{150} \in \mathbb{R}^{T \cdot N}$ such that their log power modulation corresponds to the log power fluctuations generated in the first step. We chose the alpha band as the frequency band of interest, i.e. 8 to 13 Hz. We considered epochs of 1 s length at sampling frequency 200 Hz (i.e. $T = 200$ data points per epoch).

Finally, these 151 sources time series were mapped into 26 EEG channels. The sources were randomly placed as dipoles and plausible patterns were generated via a realistic EEG forward model (Fonov et al., 2011; Nolte and Dassios, 2005). The artificial EEG signal was generated according to

$$X = \sqrt{\delta}\left(\sqrt{\gamma}\frac{\mathbf{a}^{*}s^{*}}{\|\mathbf{a}^{*}s^{*}\|_F} + \sqrt{1-\gamma}\frac{\sum_{i=1}^{150}\mathbf{a}_i s_i}{\left\|\sum_{i=1}^{150}\mathbf{a}_i s_i\right\|_F}\right) + \sqrt{1-\delta}\frac{\eta}{\|\eta\|_F} \tag{B.1}$$

where $X \in \mathbb{R}^{26 \times (T \cdot N)}$ is the EEG signal, $s^{*} \in \mathbb{R}^{T \cdot N}$ is the source whose power is driving the target z, $s_1, \ldots, s_{150} \in \mathbb{R}^{T \cdot N}$ are the biological background sources, $\mathbf{a}^{*} \in \mathbb{R}^{26}$ and $\mathbf{a}_1, \ldots, \mathbf{a}_{150} \in \mathbb{R}^{26}$ are the spread patterns of the dipolar sources evaluated at 26 electrodes, $\eta \in {}^{26 \times (T \cdot N)}$ is Gaussian sensor noise and $0 \le \gamma, \delta \le 1$ are parameters that adjust the signal-to-noise level. The normalization terms $\|\mathbf{a}^{*}s^{*}\|_F$, $\|\sum_{i=1}^{150}\mathbf{a}_i s_i\|_F$, and $\|\eta\|_F$ are used to equalize the variance of each of the system components, where $\|\cdot\|_F$ denotes the Frobenius norm.

δ governs the ratio between the biological signals and the sensor noise, while γ governs the ratio between the target source and the biological background sources. γ corresponds to the percentage of variance of all biological sources explained by the target source. We kept the

value of δ fixed in all simulations to 0.99 (i.e. 1% sensor noise), while we varied γ and the number of training epochs N.

References

Anderson, M.J., Robinson, J., 2001. Permutation tests for linear models. Aust. N. Z. J. Stat. 43 (1), 75–88.

Antons, J., Schleicher, R., Arndt, S., Möller, S., Porbadnigk, A.K., Curio, G., 2012. Analyzing speech quality perception using electro-encephalography. IEEE J. Sel. Top. Signal Proc. 6 (6), 721–731.

Ashrafulla, S., Haldar, J.P., Joshi, A.A., Leahy, R.M., 2013. Canonical granger causality between regions of interest. NeuroImage 83, 189–199.

Astolfi, L., Cincotti, F., Mattia, D., Marciani, M.G., Baccala, L.A., de Vico Fallani, F., Salinari, S., Ursino, M., Zavaglia, M., Ding, L., Edgar, J.C., Miller, G.A., He, B., Babiloni, F., 2007. Comparison of different cortical connectivity estimators for high-resolution EEG recordings. Hum. Brain Mapp. 28 (2), 143–157.

Aue, W.R., Arruda, J.E., Kass, S.J., Stanny, C.J., 2009. Cyclic variations in sustained human performance. Brain Cogn. 71 (3), 336–344.

Babiloni, C., Vecchio, F., Bultrini, A., Luca Romani, G., Rossini, P.M., 2006. Pre- and post-stimulus alpha rhythms are related to conscious visual perception: a high-resolution EEG study. Cereb. Cortex 16 (12), 1690–1700.

Baillet, S., Mosher, J., Leahy, R., 2001. Electromagnetic brain mapping. Signal Proc. Mag. IEEE 18 (6), 14–30.

Barnett, L., Seth, A.K., 2011. Behaviour of Granger causality under filtering: theoretical invariance and practical application. J. Neurosci. Methods 201, 404–419.

Başar, E., 2012. A review of alpha activity in integrative brain function: fundamental physiology, sensory coding, cognition and pathology. Int. J. Psychophysiol. 86 (1), 1–24.

Bauer, M., Oostenveld, R., Peeters, M., Fries, P., 2006. Tactile spatial attention enhances gamma-band activity in somatosensory cortex and reduces low-frequency activity in parieto-occipital areas. J. Neurosci. 26 (2), 490–501.

Berka, C., Levendowski, D., Lumicao, M., Yau, A., Davis, G., Zivkovic, V., Olmstead, R., Tremoulet, P., Craven, P., 2008. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat. Space Environ. Med. 78, B231–B244.

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 (1), 41–56.

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 BCIs technology. Front. Neurosci. 4 (198).

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 (2), 814–825.

Bressler, S.L., Seth, A.K., 2011. Wiener–Granger causality: a well established methodology. NeuroImage 58 (2), 323–329.

Buzski, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science 304 (5679), 1926–1929.

Dähne, S., Höhne, J., Schreuder, M., Tangermann, M., 2011. Band power features correlate with performance in auditory brain–computer interface. Front. Hum. Neurosci. Conference Abstract: XI International Conference on Cognitive Neuroscience (ICON XI) vol. 109.

Dähne, S., Bießmann, F., Meinecke, F.C., Mehnert, J., Fazli, S., Müller, K.-R., 2013. Integration of multivariate data streams with bandpower signals. IEEE Trans. Multimed. 15 (5), 1001–1013.

Dähne, S., Meinecke, F.C., Haufe, S., Höhne, J., Tangermann, M., Müller, K.-R., Nikulin, V.V., 2014a. SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. NeuroImage 86, 111–122.

Dähne, S., Nikulin, V.V., Ramírez, D., Schreier, P.J., Müller, K.-R., Haufe, S., 2014b. Finding brain oscillations with power dependencies in neuroimaging data. NeuroImage 96, 334–348.

Debener, S., Herrmann, C.A.C.S., Kranczioch, C., Gembris, D., Engel, A.K., 2003. Top-down attentional processing enhances auditory evoked gamma band activity. NeuroReport 14 (5), 683–686.

Dong, Y., Hu, Z., Uchimura, K., Murayama, N., 2011. Driver inattention monitoring system for intelligent vehicles: a review. IEEE Trans. Intell. Transp. Syst. 12 (2), 596–614.

Doppelmayr, M., Klimesch, W., Pachinger, T., Ripper, B., 1999. Individual differences in brain dynamics: important implications for the calculation of event-related band power. Biol. Cybern. 79, 49–57.

Engel, A.K., Gerloff, C., Hilgetag, C.C., Nolte, G., 2013. Intrinsic coupling modes: multiscale interactions in ongoing brain activity. Neuron 80 (4), 867–886.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54 (1), 313–327.

Freedman, D., Lane, D., 1983. A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. 1 (4), 292–298.

Gevins, A., Leong, H., Du, R., Smith, M.E., Le, J., DuRousseau, D., Zhang, J., Libove, J., 1995. Towards measurement of brain function in operational environments. Biol. Psychol. 40 (1–2), 169–186.

Geweke, J., 1982. Measurement of linear dependence and feedback between multiple time series. J. Am. Stat. Assoc. 77 (378), 304–313.

Gómez-Herrero, G., Atienza, M., Egiazarian, K., Cantero, J.L., 2008. Measuring directional coupling between EEG sources. NeuroImage 43 (3), 497–508.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37 (3), 424–438.

Grosse-Wentrup, M., 2011. Fronto-parietal gamma-oscillations are a cause of performance variation in brain–computer interfacing. Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on. IEEE, pp. 384–387.

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

Haegens, S., Händel, B.F., Jensen, O., 2011. Top-down controlled alpha band activity in somatosensory areas determines behavioral performance in a discrimination task. J. Neurosci. 31 (14), 5197–5204.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press.

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 (8), 1954–1963.

Haufe, S., Treder, M.S., Gugler, M.F., Sagebaum, M., Curio, G., Blankertz, B., 2011. EEG potentials predict upcoming emergency brakings during simulated driving. J. Neural Eng. 8 (5).

Haufe, S., Nikulin, V.V., Nolte, G., 2012. Alleviating the influence of weak data asymmetries on Granger-causal analyses. Latent Variable Analysis and Signal Separation. Springer, pp. 25–33.

Haufe, S., Nikulin, V., Müller, K.-R., Nolte, G., 2013. A critical assessment of connectivity measures for EEG data: a simulation study. NeuroImage 64, 120–133.

Haufe, S., Dähne, S., Nikulin, V.V., 2014a. Dimensionality reduction for the analysis of brain oscillations. NeuroImage 101, 583–597.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014b. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110.

Herrmann, C.S., Rach, S., Neuling, T., Strúber, D., 2013. Transcranial alternating current stimulation: a review of the underlying mechanisms and modulation of cognitive processes. Front. Hum. Neurosci. 7 (279).

Hipp, J.F., Hawellek, D.J., Corbetta, M., Siegel, M., Engel, A.K., 2012. Large-scale cortical correlation structure of spontaneous oscillatory activity. Nat. Neurosci. 15 (6), 884–890.

Hyvärinen, A., Oja, E., 1997. A fixed-point algorithm for independent component analysis. Neural Comput. 7, 1483–1492.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. 13 (4–5), 411–430.

Hyvärinen, A., Ramkumar, P., Parkkonen, L., Hari, R., 2010. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. NeuroImage 49, 257–271.

Jensen, O., Mazaheri, A., 2010. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front. Hum. Neurosci. 4.

Jensen, O., Kaiser, J., Lachaux, J.-P., 2007. Human gamma-frequency oscillations associated with attention and memory. Trends Neurosci. 30 (7), 317–324.

Joundi, R.A., Jenkinson, N., Brittain, J.-S., Aziz, T.Z., Brown, P., 2012. Driving oscillatory activity in the human cortex enhances motor performance. Curr. Biol. 22 (5), 403–407.

Kaiser, J., Hertrich, I., Ackermann, H., Lutzenberger, W., 2006. Gamma-band activity over early sensory areas predicts detection of changes in audiovisual speech stimuli. NeuroImage 30 (4), 1376–1382.

Kilavik, E., Zaepffel, M., Brovelli, A., MacKay, W.A., Riehle, A., 2013. The ups and downs of beta oscillations in sensorimotor cortex. Exp. Neurol. 245, 15–26.

Klimesch, W., 1999. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Res. Rev. 29, 169–195.

Klimesch, W., Vogt, F., Doppelmayr, M., 1999. Interindividual differences in alpha and theta power reflect memory performance. Intelligence 27 (4), 347–362.

Klimesch, W., Sauseng, P., Gerloff, C., 2003. Enhancing cognitive performance with repetitive transcranial magnetic stimulation at human individual alpha frequency. Eur. J. Neurosci. 17 (5), 1129–1133.

Klimesch, W., Sauseng, P., Hanslmayr, S., 2007. EEG alpha oscillations: the inhibition-timing hypothesis. Brain Res. Rev. 53 (1), 63–88.

Lal, S.K., Craig, A., 2001. A critical review of the psychophysiology of driver fatigue. Biol. Psychol. 55 (3), 173–194.

Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. NeuroImage 56 (2), 387–399.

Linkenkaer-Hansen, K., Nikouline, V.V., Palva, J.M., Ilmoniemi, R.J., 2001. Long-range temporal correlations and scaling behavior in human brain oscillations. J. Neurosci. 21 (4), 1370–1377.

Linkenkaer-Hansen, K., Nikulin, V.V., Palva, S., Ilmoniemi, R.J., Palva, J.M., 2004. Prestimulus oscillations enhance psychophysical performance in humans. J. Neurosci. 24 (45), 10186–10190.

Ljung, G.M., Box, G.E., 1978. On a measure of lack of fit in time series models. Biometrika 65 (2), 297–303.

Macdonald, J., Mathan, S., Yeung, N., 2011. Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus EEG alpha oscillations. Front. Psychol. 2.

Maeder, C., Sannelli, C., Haufe, S., Blankertz, B., 2012. Pre-stimulus sensorimotor rhythms influence brain–computer interface classification performance. IEEE Trans. Neural Syst. Rehabil. Eng. 20 (5), 653–662.

Martel, A., Dähne, S., Blankertz, B., 2014. EEG predictors of covert vigilant attention. J. Neural Eng. 11 (3), 035009.

Massimini, M., Ferrarelli, F., Esser, S.K., Riedner, B.A., Huber, R., Murphy, M., Peterson, M.J., Tononi, G., 2007. Triggering sleep slow waves by transcranial magnetic stimulation. Proc. Natl. Acad. Sci. 104 (20), 8496–8501.

Michalareas, G., Schoffelen, J.-M., Paterson, G., Gross, J., 2013. Investigating causality between interacting brain areas with multivariate autoregressive models of MEG sensor data. Hum. Brain Mapp. 34 (4), 890–913.

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 (1), 82–90.

Nalatore, H., Ding, M., Rangarajan, G., 2007. Mitigating the effects of measurement noise on Granger causality. Phys. Rev. E. 75, 031123.

Neuper, C., Pfurtscheller, G., 2001. Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. Int. J. Psychophysiol. 43, 41–58.

Nikulin, V., Brismar, T., 2005. Long-range temporal correlations in electroencephalographic oscillations: relation to topography, frequency band, age and gender. Neuroscience 130 (2), 549–558.
Nikulin, V.V., Linkenkaer-Hansen, K., Nolte, G., Lemm, S., Müller, K.R., Ilmoniemi, R.J., Curio, G., 2007. A novel mechanism for evoked responses in the human brain. Eur. J. Neurosci. 25 (10), 3146–3154.
Nikulin, V.V., Nolte, G., Curio, G., 2011. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. NeuroImage 55, 1528–1535.
Nolte, G., Dassios, G., 2005. Analytic expansion of the EEG lead field for realistic volume conductors. Phys. Med. Biol. 50 (16), 3807–3823.
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., 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 (5), 499–515.
O'Connell, R.G., Dockree, P.M., Robertson, I.H., Bellgrove, M.A., Foxe, J.J., Kelly, S.P., 2009. Uncovering the neural signature of lapsing attention: electrophysiological signals predict errors up to 20 s before they occur. J. Neurosci. 29 (26), 8604–8611.
Oken, B., Salinsky, M., Elsas, S., 2006. Vigilance, alertness, or sustained attention: physiological basis and measurement. Clin. Neurophysiol. 117, 1885–1901.
Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E., Jensen, O., 2006. Theta and gamma oscillations predict encoding and retrieval of declarative memory. J. Neurosci. 26 (28), 7523–7531.
Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. NeuroImage 28 (2), 326–341.
Pfurtscheller, G., 1981. Central beta rhythm during sensorimotor activities in man. Electroencephalogr. Clin. Neurophysiol. 51, 253–264.
Porbadnigk, A.K., Treder, M., Blankertz, B., Antons, J., Schleicher, R., Möller, S., Curio, G., Müller, K., 2013. Single-trial analysis of the neural correlates of speech quality perception. J. Neural Eng. 10 (5), 056003.
Rieder, M.K., Rahm, B., Williams, J.D., Kaiser, J., 2011. Human γ-band activity and behavior. Int. J. Psychophysiol. 79, 39–48.
Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using granger causality and fMRI. NeuroImage 25 (1), 230–242.
Romei, V., Gross, J., Thut, G., 2010. On the role of prestimulus alpha rhythms over occipito-parietal areas in visual input regulation: correlation or causation? J. Neurosci. 30 (25), 8692–8697.
Schmidt, M. MinFunc optimization functionwww.di.ens.fr/mschmidt/Software/minFunc.html (Accessed: 03/2012).
Schneider, T., Neumaier, A., 2001. Algorithm 808: ARfit — a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Trans. Math. Softw. 27 (1), 58–65.
Schoffelen, J.-M., Gross, J., 2009. Source connectivity analysis with MEG and EEG. Hum. Brain Mapp. 30 (6), 1857–1865.
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. Cogn. Neurosci. 21 (12), 2407–2419.
Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Stouffer, S., Suchman, E., DeVinney, L., Star, S., Williams, R.J., 1949. The American Soldier, Vol. 1: Adjustment During Army Life. Princeton University Press, Princeton.
Thut, G., Miniussi, C., 2009. New insights into rhythmic brain activity from TMS–EEG studies. Trends Cogn. Sci. 13 (4), 182–189.
Thut, G., Nietzel, A., Brandt, S.A., Pascual-Leone, A., 2006. Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. J. Neurosci. 26 (37), 9494–9502.
Thut, G., Miniussi, C., Gross, J., 2012. The functional importance of rhythmic activity in the brain. Curr. Biol. 22 (16), R658–R663.
Tsay, R.S., 2005. Analysis of Financial Time Series vol. 543. John Wiley and Sons.
van Dijk, H., Schoffelen, J.-M., Oostenveld, R., Jensen, O., 2008. Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. J. Neurosci. 28 (8), 1816–1823.
Vicente, R., Wibral, M., Lindner, M., Pipa, G., 2011. Transfer entropy — a model-free measure of effective connectivity for the neurosciences. J. Comput. Neurosci. 30, 45–67.
Whitlock, M.C., 2005. Combining probability from independent tests: the weighted z-method is superior to Fisher's approach. J. Evol. Biol. 18 (5), 1368–1373.
Womelsdorf, T., Fries, P., 2007. The role of neuronal synchronization in selective attention. Curr. Opin. Neurobiol. 17 (2), 154–160.
Zaehle, T., Sandmann, P., Thorne, J., Jancke, L., Herrmann, C., 2011. Transcranial direct current stimulation of the prefrontal cortex modulates working memory performance: combined behavioural and electrophysiological evidence. BMC Neurosci. 12 (1), 2.

4.2 Conclusion

Summary

In this paper, we have analyzed how to identify Granger causal links from brain oscillations to behavioral variables of interest. We considered the case in which we simultaneously measure EEG band power ϕ and a target variable z over time. Then ϕ is said to Granger cause z if ϕ helps to predict the future z above what is predicted by the past of z alone (cf. Section 2.3.3). Here, z can be any signal of interest, such as a behavioral output (e.g., reaction time, sensory detection, task rating, evoked potentials), a physiological measure (e.g., muscular activity, heart rate variability) or a second power time course.

The simplest approach to testing for Granger causality is to consider each channel separately. However, the physics of EEG implies that the activity measured at a given channel is a mixture of contributions from several neuronal sources (cf. Section 2.1.2). This leads to a low signal-to-noise (SNR) ratio and may hinder physiological interpretation of the results.

It is therefore beneficial to attempt to recover the underlying neuronal source signals from scalp recordings prior to the computation of band-power dynamics and the test for Granger causality. Here, we compared channel-wise Granger causality testing with three multivariate spatial filtering approaches:

- Independent Component Analysis (ICA), which seeks maximally statistically independent sources, and is one of the most popular method for extracting interesting neuronal sources from EEG data (cf. Section 2.2.2). Here we used FastICA (Hyvärinen and Oja, 1997; Hyvärinen, 1999).

- The recently proposed SPoC method (Dähne et al., 2014a) which extracts neuronal sources whose band-power is maximally correlated with the target variable.

- A novel method which extracts a source whose band-power maximally Granger causes the target variable, referred to as Granger Causal Power Analysis (GrangerCPA). Because we are interested in recovering sources whose power dynamics Granger cause an external variable, we might benefit from basing the reconstruction of source activity exactly on this assumed dependency.

  To state GrangerCPA's objective function, let us introduce the following notation. We consider time-windowed and band-pass filtered EEG measurements $X = (X_1, \ldots, X_{N_e}) \in \mathbb{R}^{M \times (T_e \cdot N_e)}$ where M is the number of channels, N_e is the number of epochs and T_e the number of data points per epoch. Given a target variable $z = (z_1, \ldots, z_{N_e}) \in \mathbb{R}^{1 \times N_e}$, we seek a spatial filter $\mathbf{w} \in \mathbb{R}^M$ such that

the epoch-wise log power of the estimated source $\mathbf{w}^\top X$ maximally Granger causes the target variable z.

The log band-power of the source extracted by the spatial filter $\mathbf{w}$ in epoch i can be computed as

$$\phi_{\mathbf{w}i} := \log\left(\mathrm{Var}(\mathbf{w}^\top X_i)\right) = \log\left(\mathbf{w}^\top \mathrm{Cov}(X_i)\mathbf{w}\right) ,$$

and we further transform z and $\phi_{\mathbf{w}}$ to have zero mean. Granger causality from $\phi_{\mathbf{w}}$ to z is then defined as

$$\log \frac{\mathrm{Var}(\text{Residuals from regressing } z_i \text{ on } z_{i-1}, \ldots, z_{i-p})}{\mathrm{Var}(\text{Residuals from regressing } z_i \text{ on } z_{i-1}, \ldots, z_{i-p} \text{ and } \phi_{\mathbf{w}i-1}, \ldots, \phi_{\mathbf{w}i-p})}$$

where p is the number of considered time lags.

Fortunately the numerator does not depend on $\mathbf{w}$. Thus, maximizing Granger causality with respect to $\mathbf{w}$ reduces to minimizing the following cost function:

$$\mathcal{L}(\mathbf{w}) := \mathrm{Var}(\text{Residuals from regressing } z_i \text{ on } z_{i-1}, \ldots, z_{i-p} \text{ and } \phi_{\mathbf{w}i-1}, \ldots, \phi_{\mathbf{w}i-p})$$

This objective is a non-convex, higher order nonlinear function of the spatial filter $\mathbf{w}$. A local minimum can be obtained by means of standard nonlinear optimization techniques. We used an L-BFGS algorithm implemented by (Schmidt, 2005).

A comparison of these methods was carried out both in simulations and on two real EEG data sets:

- *Self-paced braking.* This data set stems from 18 subjects which performed self-paced foot movements (Haufe et al., 2011), and it is the same data set we analyzed in Section 3.4. It served as a proof of concept, since brief, self-paced movements are preceded by a well-documented event-related desynchronization (ERD) of 15-30 Hz (beta band) rhythms over corresponding sensorimotor areas (Neuper and Pfurtscheller, 2001).

  We applied all considered methods in order to extract oscillatory beta components which predict subsequent muscle activity. Indeed, identified neuronal sources whose power significantly Granger causes muscular activity showed neurophysiologically plausible motor-related patterns. In contrast, components with patterns that are ambiguous or artifactual did not significantly Granger cause the motor activity.

- *Reaction Times.* In the second real EEG data set (Antons et al., 2012; Porbadnigk et al., 2013), we applied Granger causal analysis to find oscillatory

alpha components that are predictive of reaction times. As reaction times are autocorrelated (cf. (Aue et al., 2009)), we can hope to identify brain activity which reflect this variation. It is known that ongoing pre-stimulus alpha power modulates perception performance (Jensen and Mazaheri, 2010), and is negatively correlated with subjective attentional state (Macdonald et al., 2011).

The results show that computing Granger causality in sensor space suffers from poor signal-to-noise ratio. The multivariate spatial filtering approaches we considered alleviated this issue. Their comparison must be undertaken in more detail:

- SPoC extracts a source whose band-power maximally correlates with the target variable. Simulations showed that, if a source is both Granger causal and highly correlated with the target variable, SPoC will identify a causal source. However, if the Granger causal source is only weakly correlated with the target variable or if a second merely correlated source is present, SPoC fails to recover the causal source. Importantly, such cases can occur in real data. For example, SPoC gets stuck on artifactual activity in one subject of the selfpaced braking data set.

- FastICA was slightly outperformed by GrangerCPA both in our simulated EEG data and on the reaction time EEG data set. However, both methods performed equally well on the self-paced braking EEG data set. In general, both algorithms may arrive at relatively similar solutions using entirely different assumptions.

Limitations and Future Work

While the causal role of oscillatory neuronal activity is a very interesting research question (Buzsáki and Draguhn, 2004; Thut and Miniussi, 2009), the main limitation of this work is that Granger causality in fact offers only weak evidence for causality (cf. Section 2.3.3). In the context of the paper, the following are the most serious limitations:

- *Hidden common drivers.* Already Granger pointed out that standard Granger causality can lead to spurious results if not all relevant variables are incorporated in the model (Granger, 1969). This poses serious issues: Suppose GrangerCPA identified an oscillatory alpha component and we assess that it Granger causes reaction times. We might argue that we found a causal source because GrangerCPA optimizes over all possible alpha sources. However, we cannot exclude the possibility that the alpha component and reaction times are only correlated, because they may both be driven by an unobserved theta rhythm or another unobserved confounder.

- *Measurement noise.* Spurious Granger causality may arise due to measurement noise: If two variables measure the same signal but are superposed with noise, they mutually help predicting each other's future (Nalatore et al., 2007; Nolte et al., 2008).

  Fortunately, spurious Granger causality cannot arise from noise on the neuronal source in the way we applied Granger causality here. This is because noise on the neuronal band power ϕ will never help to predict the future of the target variable z.

  However, spurious causality may arise if the target variable is strongly affected by measurement noise. Consider a case in which ϕ and z are correlated, but there is no causal influence from ϕ onto z. Suppose further that z is affected by additive measurement noise ϵ', and we only measure $z' := z + \epsilon'$. Now ϕ may contain cleaner information about z' than z' itself and may therefore help to predict future z'. We would infer Granger causality in the absence of causality.

  It is important to note here that the target variable z itself is allowed to be noisy. That is, additive measurement noise ϵ' on z and innovation noise in an AR model of z have very different effects. It is only measurement noise that poses a problem for Granger causal analysis. This is because measurement noise implies that the AR model for z' does not describe the underlying process z that we are actually interested in. This issue refers to the interpretation of a Granger causal finding: Suppose we found an alpha component which Granger causes reaction times, and we somehow knew that there are no hidden confounders. It might be valid to infer that the alpha component is causally involved in determining the reaction times. However, we cannot infer that the alpha component causes motivational shifts which we measure using reaction times as a noisy proxy.

- *Downsampling.* Spurious Granger causality has been reported to arise due to downsampling (McCrorie and Chambers, 2006; Zhou et al., 2014). This may also be problematic in our scenario, since we estimate band-power as the variance over some pre-defined, possibly long, time window.

- *Non-Stationarity.* Autoregressive modeling and Granger causality assumes stationarity. However, EEG signals are intrinsically non-stationary (Dornhege et al., 2007; von Bünau et al., 2009; Samek et al., 2012). When we record data over a longer time interval, the autoregressive model may therefore be inadequate, which might also induce spurious causality.

Stronger evidence for causal links may be gathered from intervention techniques such as repetitive Transcranial Magnetic Stimulation (rTMS) and Transcranial Al-

ternating Current Stimulation (TACS) (Thut et al., 2012; Herrmann et al., 2013). Indeed, accumulating evidence suggests that rhythmic stimulation induces behavioral consequences, for instance on visual perception (Romei et al., 2010), motor performance (Joundi et al., 2012), working memory (Zaehle et al., 2011), and sleep stages (Massimini et al., 2007).

Nevertheless, statistical methods offer the advantage of non-invasiveness and should therefore be explored. Many other causal inference problems also suffer from similar issues. Therefore, the development and validation of more robust causality scores is an important research direction, and causal inference algorithms which are more robust with respect to unobserved hidden variables (Hoyer et al., 2008; Entner and Hoyer, 2011; Chen and Chan, 2013; Tashiro et al., 2014; Geiger et al., 2015), or measurement noise (Nalatore et al., 2007; Nolte et al., 2008; Vicente et al., 2011; Haufe et al., 2013; Vinck et al., 2015) are being developed. We explore a remedy to the problem of measurement noise proposed in (Haufe et al., 2013) in the next chapter.

5 Time-reversal for Granger causality

5.1 Validity of time reversal for testing Granger causality

Irene Winkler, Danny Panknin, Daniel Bartz ,Klaus-Robert Müller, and Stefan Haufe. Validity of time reversal for testing Granger causality. Submitted.
The manuscript is available online as arXiv preprint *arXiv:1509.07636*.
`http://arxiv.org/abs/1509.07636`

Short summary. To alleviate the problem of spurious causality due to measurement noise, (Haufe et al., 2013) proposed to contrast causality scores on the original against those from the time-reversed time series. The intuitive idea is that, if temporal order is crucial to tell a driver from a recipient, causality results should change if the temporal order is reversed. In a recent independent simulation study, (Vinck et al., 2015) confirmed that this approach, time-reversed Granger causality (TRGC), leads to a much smaller fraction of false positives as compared to original Granger causality, and also compares favorably against another more noise-robust causality metric, the Phase Slope Index (PSI) (Nolte et al., 2008).

While time-reversed Granger causality thus yields encouraging results in simulation studies, it was not well understood why and how computing Granger causality on the time-reversed signals links to the causal interactions on the original time-series. In this manuscript, we prove that TRGC will indeed indicate the correct directionality in finite order autoregressive processes with unidirectional information flow.

Contributions. I wrote the majority of the paper and carried out all the simulations. The proof in Appendix A-C stems from Danny Panknin (my previous version relied on an additional assumption that A_p is invertible).

Validity of time reversal for testing Granger causality

Irene Winkler, Danny Panknin, Daniel Bartz, Klaus-Robert Müller, *Member, IEEE,* and Stefan Haufe

***Abstract*—Inferring causal interactions from observed data is a challenging problem, especially in the presence of measurement noise. To alleviate the problem of spurious causality, Haufe et al. (2013) proposed to contrast measures of information flow obtained on the original data against the same measures obtained on time-reversed data. They show that this procedure, time-reversed Granger causality (TRGC), robustly rejects causal interpretations on mixtures of independent signals. While promising results have been achieved in simulations, it was so far unknown whether time reversal leads to valid measures of information flow in the presence of true interaction. Here we prove that, for linear finite-order autoregressive processes with unidirectional information flow, the application of time reversal for testing Granger causality indeed leads to correct estimates of information flow and its directionality. Using simulations, we further show that TRGC is able to infer correct directionality with similar statistical power as the net Granger causality between two variables, while being much more robust to the presence of measurement noise.**

***Index Terms*—Granger causality, time reversal, noise, TRGC**

I. Introduction

THE estimation of causal relations between time series is a signal processing topic promising to enhance our understanding of dynamical systems in numerous application domains. For data with time structure, the concept of Granger causality (GC) has gained popularity as a simple testable definition of causality based on temporal precedence. Signal processing techniques based on Granger-causality have been studied in a variety of fields such as econometrics [1], neuroscience [2], [3], [4], [5], and climate science [6], [7].

In its original formulation, a time series x_t is said to Granger-cause a time series y_t, if the past of x_t helps to predict y_t above what can be predicted by using 'all other information in the universe' besides the past of x_t [8]. In practice, it is common to consider only the information contained in the past of x_t and y_t (cf. [9]).

A serious problem for the estimation of information flow using Granger causality is that spurious Granger causality can occur due to measurement noise. On one hand, if two sensors measuring the same signal are superimposed with noise, they mutually help predicting each other's future [10], [11]. This is a problem especially in the study of brain connectivity using non-invasive electrophysiology, where the activity at a given sensor is typically a mixture of contributions from several neuronal sources due to the volume conduction of electric currents in the head [12], [13], [14], [15], [16]. On the other hand, noise that is correlated across sensors has a similar adverse effect on estimates of directed interaction even if the actual signals-of-interest are not mixed into different sensors [17], [18]. Such spurious causality can occur in any measure based on the concept of Granger causality, including multivariate [19], [20] and non-linear [21], [22], [23] variants.

Recently, a number of ways to make causality estimates more robust to the presence of mixed signals and noise have been proposed. These include novel measures of directed information flow [12], [10], [11] as well as novel ways of assessing their statistical significance [24], [23], [17], [16], [18]. Recently, Haufe et. al [17], [16] suggested to contrast causality scores obtained on the original time series to those obtained on time-reversed signals. The intuitive idea behind this approach is that, if temporal order is crucial to tell a driver from a recipient, directed information flow should be reduced (if not reversed) if the temporal order is reversed. In fact, Haufe et al. showed that for correlated, but non-interacting signals, the use of time reversal for testing Granger causality scores (here referred to as time-reversed Granger causality, TRGC) and other metrics based on cross-spectral estimates or linear autoregressive modeling correctly leads to rejection of causal interpretations. This was confirmed for Granger causality in an independent simulation study [18] showing that TRGC leads to a much smaller fraction of false positive detections compared to the original Granger causality index, and also compares favorably against the Phase Slope Index (PSI) [11].

While time-reversed Granger causality thus displays an intriguing noise robustness property, and yields very encouraging results in simulations, its behavior *in the presence of causal interactions* is still poorly understood. In particular, it is currently unclear how Granger causality scores computed on time-reversed signals link to the causal interactions on the original time-series, and therefore whether TRGC correctly indicates the direction of causality. Theoretical guarantees have only been derived for special cases in which either the signal's auto- and cross-covariances are very small in magnitude, or in which both signals have very similar autocorrelations [18].

The aim of this paper is two-fold. In the theory section, we provide new theoretical insights on time-reversal for testing Granger causality. After introducing the concepts of linear

This work was supported by a Marie Curie International Outgoing Fellowship (grant No. 625991) within the 7th European Community Framework Program, the BMBF project ALICE II, Autonomous Learning in Complex Environments (01IB15001B), and the Brain Korea 21 Plus Program as well as the SGER Grant 2014055911 through the National Research Foundation of Korea funded by the Ministry of Education.

I. Winkler, D. Panknin, D. Bartz, K.-R. Müller and S. Haufe are with the Machine Learning Group, Technische Universität Berlin, Germany. K.-R. Müller is also with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea. S. Haufe is also with the Laboratory for Intelligent Imaging and Neural Computing, Columbia University, New York, USA. Correspondence to: {i.winkler,stefan.haufe}@tu-berlin.de.

autoregressive modeling, Granger causality, and time-reversed Granger causality (Section II-A and II-B), we elaborate on the existing result of Haufe et al. [16] showing that, for mixtures of independent signals, causality measures based on cross-covariances are invariant to the reversal of the temporal order (Section II-C). This is the theoretical basis for the noise-robustness property of time-reversal testing of causality scores. We then investigate the time-reversal of a process fulfilling the assumptions typically made by Granger causality estimators: a finite-order vector autoregressive (VAR) process that is unaffected by measurement noise. We review what is known about the time-reversal of a VAR process (Section II-D), based on what we provide an analytic description of Granger causality scores of time-reversed signals in terms of their autoregressive coefficients (Section II-E) and a minimal example (Section II-F). Using these insights, we prove our main result stating that, in the case of unambiguous unidirectional information flow from x_t to y_t, time reversal leads to a decrease of the Granger-causal net information flow relative to the original time series. The difference of net Granger causality scores obtained on original and time-reversed data thus indicates the correct direction of interaction (Section II-G).

In the second part of the paper (Section III), we revisit scenarios known to cause problems for conventional Granger causality. Using simulations, we illustrate when and how the theoretical guarantees of TRGC lead to measurable performance increases in practice. We point out the implications of our theoretical and empirical results in Section IV, along with a discussion of ambiguities in causal interpretation caused by the presence of correlated residuals in VAR models.

II. Theory

Vectors are considered to be column vectors (unless otherwise stated), and are generally typed in bold. The symbol $\cdot^\top$ denotes the transpose operator, I the identity matrix, and $[\cdot,\cdot]$ concatenation. The symbol $\otimes$ refers to the Kronecker product, and $\mathrm{vec}(\cdot)$ to the vectorization operator, which converts a matrix into a column vector. The symbol $\langle\cdot\rangle$ denotes expectation. The cross-covariance matrices of a stationary process $\mathbf{z}_t$ are denoted by

$$C_{\mathbf{z}}(h) := \left\langle (\mathbf{z}_t - \langle\mathbf{z}\rangle)(\mathbf{z}_{t-h} - \langle\mathbf{z}\rangle)^\top \right\rangle \qquad \forall h \in \mathbb{Z}\ .$$

We use the notation $\mathbf{z}_t$ both for an observed time series and its underlying data generating process. We denote all quantities related to the time-reversed process $\tilde{\mathbf{z}}_t := \mathbf{z}_{-t}$ with a tilde.

A process $\boldsymbol{\epsilon}_t$ is said to be *white noise* if it is stationary with mean zero, finite covariance and zero autocorrelation; that is, if $C_\epsilon(h) = 0\ \forall h \in \mathbb{Z}\backslash\{0\}$. Note that the covariance matrix $C_\epsilon(0)$ is not necessarily diagonal, and that neither independence nor joint Gaussianity is required.

A. Granger causality and the linear VAR model

Consider a stable bivariate vector autoregressive process of lag order p (VAR(p) process), $\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \mathbb{R}^2$,

$$\mathbf{z}_t = A_1\mathbf{z}_{t-1} + A_2\mathbf{z}_{t-2} + \ldots + A_p\mathbf{z}_{t-p} + \epsilon_t\ , \qquad (1)$$

where $\epsilon_t \in \mathbb{R}^2$ is a 2-dimensional white noise process (that is, $\langle\epsilon_t\rangle = 0$, $\langle\epsilon_t\epsilon_{t-h}^\top\rangle = 0$ for $h \in \mathbb{Z}\setminus\{0\}$, and $\langle\epsilon_t\mathbf{z}_{t-h}^\top\rangle = 0$ for $h \in \mathbb{N}\setminus\{0\}$) with residual covariance matrix

$$\Sigma = \langle\epsilon_t\epsilon_t^\top\rangle = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{bmatrix}\ . \qquad (2)$$

The noise variables ϵ_t are also called *innovations* or *residuals*. Stability requires that $\det(I - A_1\lambda - \ldots - A_p\lambda^p) \neq 0$ for all $\lambda \in \mathbb{C}$ with $|\lambda| \leq 1$.

Following [25], x_t and y_t possess themselves autoregressive (AR) representations, which we denote by

$$x_t = \sum_{k=1}^{\infty} a_k x_{t-k} + \xi_t^x, \quad \mathrm{Var}(\xi_t^x) =: \Sigma_x \quad \text{and} \qquad (3)$$

$$y_t = \sum_{k=1}^{\infty} b_k y_{t-k} + \xi_t^y, \quad \mathrm{Var}(\xi_t^y) =: \Sigma_y\ . \qquad (4)$$

The residuals ξ_t^x and ξ_t^y of these two univariate processes are each serially uncorrelated, but may be correlated with each other at various lags. Importantly, even though the bivariate autoregressive process (1) is of finite order, the univariate processes (3) and (4) are in general of infinite order. We refer to (1) as the *unrestricted* or *full* model, while (3) and (4) contain the *restricted* models.

Directed Granger-causal information flow is defined based on the so-called *Granger-scores* [25]

$$F_{y\to x} := \log\left(\frac{\Sigma_x}{\Sigma_{xx}}\right) \quad \text{and} \quad F_{x\to y} := \log\left(\frac{\Sigma_y}{\Sigma_{yy}}\right)\ . \qquad (5)$$

Granger causality from x_t to y_t implies that information from the past of x_t improve the prediction of the present of y_t compared to what can be predicted by the past of y_t alone. That is, the residual variance Σ_{yy} of the unrestricted model is required to be smaller than the residual variance Σ_y of the restricted model. Under the assumption of Gaussian-distributed residuals, $F_{y\to x}$ and $F_{x\to y}$ are asymptotically χ^2 distributed, giving rise to an analytical test of their significance [25]. An asymptotically equivalent test is given by an F-test of the goodness-of-fit of the two models (cf. [9], [4]). We refer to this approach as *standard Granger causality* (standard GC).

As variables in physical systems often mutually influence each other, it is also of interest to determine the *net* driver of the interaction by assessing whether more information is flowing from x_t to y_t then from y_t to x_t or vice versa. Following [11], [26], *net Granger causality* (Net-GC) is defined as the difference of the Granger causality scores, that is

$$F_{x\to y}^{(net)} := F_{x\to y} - F_{y\to x} \quad \text{and} \quad F_{y\to x}^{(net)} := -F_{x\to y}^{(net)}\ . \qquad (6)$$

As the analytical distributions of these differences are unkown, statistical significance of Net-GC scores needs to be assessed using resampling methods as outlined in Section III-A.

B. Time-reversed Granger causality (TRGC)

To avoid false detections of causal interactions, Haufe et al. proposed to contrast causality measures applied to the original time series with the same measures obtained from time-reversed signals $\tilde{\mathbf{z}}_t := \mathbf{z}_{-t}$ [17], [16]. Here, we formalize this idea in the context of Granger causality.

Given a bivariate VAR(p) process, its time-reversed process $\tilde{\mathbf{z}}_t$ also possesses a VAR(p) representation, which we derive in Section II-D. We denote the residual covariance matrix of the time-reversed process by

$$\tilde{\Sigma} = \begin{bmatrix} \tilde{\Sigma}_{xx} & \tilde{\Sigma}_{xy} \\ \tilde{\Sigma}_{xy} & \tilde{\Sigma}_{yy} \end{bmatrix} . \tag{7}$$

The restricted AR models of the time-reversed data have a simple structure, as they are concerned with univariate time series. The autocovariance function of a univariate time series is symmetric, i.e., we have $C_x(h) = C_x(-h)$ and $C_y(h) = C_y(-h)$ for all $h \in \mathbb{Z}$. As a result of this and (19) (Section II-C), the time-reversed signals will have the same autocovariances as the original series. Because the AR representation is uniquely determined by the autocovariance function (cf. Section II-D1), they also share the same AR representation. The restricted models of the time-reversed univariate processes are thus given by

$$x_t = \sum_{k=1}^{\infty} a_k x_{t+k} + \tilde{\xi}_t^x, \quad \mathrm{Var}(\tilde{\xi}_t^x) =: \tilde{\Sigma}_x \quad \text{and} \tag{8}$$

$$y_t = \sum_{k=1}^{\infty} b_k y_{t+k} + \tilde{\xi}_t^y, \quad \mathrm{Var}(\tilde{\xi}_t^y) =: \tilde{\Sigma}_y \tag{9}$$

with

$$\tilde{\Sigma}_x = \Sigma_x \qquad \text{and} \qquad \tilde{\Sigma}_y = \Sigma_y . \tag{10}$$

In analogy to the original time series, we define the time-reversed Granger scores as

$$\tilde{F}_{\tilde{y}\to\tilde{x}} := \log\left(\frac{\tilde{\Sigma}_x}{\tilde{\Sigma}_{xx}}\right) \quad \text{and} \quad \tilde{F}_{\tilde{x}\to\tilde{y}} := \log\left(\frac{\tilde{\Sigma}_y}{\tilde{\Sigma}_{yy}}\right) , \tag{11}$$

and the net Granger causality scores as

$$\tilde{F}^{(net)}_{\tilde{x}\to\tilde{y}} := \tilde{F}_{\tilde{x}\to\tilde{y}} - \tilde{F}_{\tilde{y}\to\tilde{x}} \quad \text{and} \quad \tilde{F}^{(net)}_{\tilde{y}\to\tilde{x}} := -\tilde{F}^{(net)}_{\tilde{x}\to\tilde{y}} . \tag{12}$$

Finally, the differences of the Granger scores obtained on original and time-reversed signals are given by

$$\tilde{D}_{y\to x} := F_{y\to x} - \tilde{F}_{\tilde{y}\to\tilde{x}} , \tag{13}$$

$$\tilde{D}_{x\to y} := F_{x\to y} - \tilde{F}_{\tilde{x}\to\tilde{y}} , \quad \text{and} \tag{14}$$

$$\tilde{D}^{(net)}_{x\to y} := F^{(net)}_{x\to y} - \tilde{F}^{(net)}_{\tilde{x}\to\tilde{y}} . \tag{15}$$

Time-reversed Granger causality can be applied in the following variants.

a) Conjunction-based time-reversed Granger causality (Conj-TRGC): Here, net information flow from x_t to y_t is inferred if

$$F^{(net)}_{x\to y} > 0 \quad \text{and} \quad \tilde{F}^{(net)}_{\tilde{x}\to\tilde{y}} < 0 , \tag{16}$$

that is, if the directionality of net Granger causality reverses for time-reversed signals. This variant has been investigated in [18].

b) Difference-based time-reversed Granger causality (Diff-TRGC): Here, net information flow from x_t to y_t is inferred if

$$\tilde{D}^{(net)}_{x\to y} > 0 , \tag{17}$$

that is, we require that net Granger causality from x_t to y_t is reduced on the time-reversed signals. Note that this is a weaker requirement than conjunction-based TRGC, as all signals for which (16) holds also fulfill (17).

c) Conjunction of Net-GC and Diff-TRGC: Finally, we can require both the time-reversed net difference and the net Granger score to be significantly larger than zero in order to infer net information flow from x_t to y_t, that is

$$\tilde{D}^{(net)}_{x\to y} > 0 \quad \text{and} \quad F^{(net)}_{x\to y} < 0 . \tag{18}$$

Just as for Net-GC, statistical significance of Conj-TRGC and Diff-TRGC, as well as the combination of Net-GC and Diff-TRGC can be assessed using resampling techniques (see Section III-A).

C. Robustness of time-reversed Granger causality (TRGC)

In [16] it is pointed out that time-reversed Granger causality robustly rejects causal interpretations for mixtures of non-interacting signals such as correlated noise sources. The mathematical basis for this noise robustness property is the fact that the cross-covariance matrices $\tilde{C}_{\tilde{\mathbf{z}}}(\cdot)$ of the time-reversed signals are equal to the transposed cross-covariance matrices of the original signals, that is

$$\tilde{C}_{\tilde{\mathbf{z}}}(h) = \langle \tilde{\mathbf{z}}_t \tilde{\mathbf{z}}_{t-h}^\top \rangle = \langle \mathbf{z}_t \mathbf{z}_{t+h}^\top \rangle = C_{\mathbf{z}}(-h) = \big(C_{\mathbf{z}}(h)\big)^\top \tag{19}$$

for all $h \in \mathbb{Z}$. If a series η_t only contains a mixture of independent signals, all its cross-covariance matrices are symmetric [27]: consider $\eta_t = M s_t$ where s_t contains a number of independent sources. Then, for all $h \in \mathbb{Z}$, $C_s(h) = diag$ and thus $C_\eta(h) = M C_s(h) M^\top$ is symmetric. For mixtures of independent noise sources, any causality measure that is solely based on a series' cross-covariance matrices therefore yields the same result on the original and the time-reversed signals. This includes Granger causality, but also other popular variants such as directed transfer function (DTF) [19] and partial directed coherence (PDC) [20]. Given sufficient amounts of data, the conditions for Conj-TRGC and Diff-TRGC cannot be fulfilled for mixtures of independent sources using these measures, preventing the detection of spurious interaction.

D. The VAR representation of a time-reversed process

There is so far no theoretical argument guaranteeing that time-reversed Granger causality correctly indicates the presence of information flow as well as its direction *in the presence of actual interaction*. In order to provide such a guarantee, we here study the time-reversal of (linear) finite-order VAR processes. Note that studying this case is sufficient since, as a results of Wold's decomposition theorem, every stationary, purely nondeterministic, process can be approximated well by a finite order VAR process [28], [1].

We start by briefly revisiting the link between cross-covariance matrices and VAR representation, which we use throughout the paper, in Section II-D1. In Sections II-D2 and II-D3, we then review the theoretical result of Andel [29] stating that the time-reversed signal of any VAR(p) process has again a VAR(p) representation that can be expressed analytically in terms of the original process. As the description for $p > 1$ is mathematically involved, we only treat the case $p = 1$ in the main paper, while the proof for arbitrary p is presented in Appendix A.

We use these results to provide an analytic description of difference-based TRGC scores in terms of their autoregressive coefficients (Section II-E), give a minimal example (Section II-F), and prove our main result stating that, in the case of unambiguous unidirectional information flow, difference-based time-reversed Granger causality indeed yields the correct result (Section II-G).

1) The cross-covariance function of a VAR process: Most of the insights in this paper are based on the direct link between autoregressive coefficient matrices $A_1, \ldots, A_p$ and residual covariance matrices Σ on one hand, and cross-covariance matrix $C_{\mathbf{z}}(\cdot)$ on the other hand. This link is established by the Yule-Walker equations as follows (see e.g. [1]). For a VAR(1) process

$$\mathbf{z}_t = A_1 \mathbf{z}_{t-1} + \epsilon_t \ , \tag{20}$$

the Yule-Walker equations read

$$C_{\mathbf{z}}(0) = A_1 \cdot C_{\mathbf{z}}(0) \cdot A_1^\top + \Sigma \quad \text{and} \tag{21}$$

$$C_{\mathbf{z}}(h) = A_1 \cdot C_{\mathbf{z}}(h-1) \quad (\forall h \in \mathbb{N} \setminus \{0\}) \ . \tag{22}$$

Given A_1 and Σ, the cross-covariances are uniquely determined from (21) through

$$\mathrm{vec}(C_{\mathbf{z}}(0)) = (I - A_1 \otimes A_1)^{-1} \, \mathrm{vec}\, \Sigma \ , \tag{23}$$

while higher-order cross-covariances $C_{\mathbf{z}}(h)$ can be recursively computed using (22). Conversely, A_1 and Σ are uniquely determined by the cross-covariances through

$$A_1 = C_{\mathbf{z}}(1) C_{\mathbf{z}}(0)^{-1} \quad \text{and} \tag{24}$$

$$\Sigma = C_{\mathbf{z}}(0) - A_1 C_{\mathbf{z}}(0) A_1^\top \ . \tag{25}$$

Results on VAR(1) processes can typically be extended to higher-order VAR(p) processes by reducing VAR(p) processes to their VAR(1) form. The VAR(1) representation of a VAR(p) process as well as the Yule-Walker equations for general VAR(p) processes are provided in Appendix A-A.

2) The VAR representation of a time-reversed VAR(1) process: The time-reversed autoregressive representation of a VAR(1) process $\mathbf{z}_t$ has been derived by Bartlett in 1955 [30]. Suppose we generate an infinite sequence of $\mathbf{z}_t$ according to the VAR(1) process (20). The VAR representation of the *time-reversed* or *backward* process is given by

$$\mathbf{z}_t = \tilde{A}_1 \mathbf{z}_{t+1} + \tilde{\epsilon}_t \ , \tag{26}$$

where

$$\tilde{A}_1 = C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1} \ , \tag{27}$$

and where the reversed residuals $\tilde{\epsilon}_t$ are calculated from $\mathbf{z}_t$ as

$$\tilde{\epsilon}_t := \mathbf{z}_t - \tilde{A}_1 \mathbf{z}_{t+1} \tag{28}$$

with residual covariance matrix

$$\tilde{\Sigma} = \langle \tilde{\epsilon}_t \tilde{\epsilon}_t^\top \rangle = C_{\mathbf{z}}(0) - C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1} \cdot A_1 \cdot C_{\mathbf{z}}(0) \ . \tag{29}$$

It is easy to show that the sequence $\tilde{\epsilon}_t$ is indeed white noise, that is for all $h \in \mathbb{Z} \setminus \{0\}$: $\langle \tilde{\epsilon}_t \cdot \tilde{\epsilon}_{t-h}^\top \rangle = 0$ and for all $h \in \mathbb{N} \setminus \{0\}$: $\langle \tilde{\epsilon}_t \cdot \mathbf{z}_{t+h}^\top \rangle = 0$.

From (27), we see that the time-reversed coefficient matrix $\tilde{A}_1$ is similar to A_1, and thus shares some of its properties, notably its eigenvalues, determinant, trace and rank. However, in the context of Granger causality, it is important to note that many properties of A_1 do not transfer to $\tilde{A}_1$. In particular, if A_1 is triangular, diagonal, or symmetric, this is not generally the case for $\tilde{A}_1$.

3) The VAR representation of a time-reversed VAR(p) process: The result of Bartlett on the time-reversed VAR(1) process has been generalized to VAR(p) processes by Andel in 1972 [29], in a paper that received, so far, little attention. Andel showed that any stable VAR(p) process (1) has a time-reversed representation

$$\mathbf{z}_t = \tilde{A}_1 \mathbf{z}_{t+1} + \tilde{A}_2 \mathbf{z}_{t+2} + \ldots + \tilde{A}_p \mathbf{z}_{t+p} + \tilde{\epsilon}_t \tag{30}$$

that is again of order p with uniquely defined autoregressive coefficients $\tilde{A}_1, \ldots \tilde{A}_p$ and residual covariance matrix $\tilde{\Sigma}$. We reproduce this result in Appendix A-B. Note that, while we only treat bivariate VAR processes in this paper, the analytic description of the time-reversed VAR process holds for processes of arbitrary dimensionality.

E. Analytic description of Diff-TRGC

Contrasting Granger scores obtained on original with those obtained on time-reversed signals is simplified by the fact that the AR representation of a univariate time series does not depend on the direction of time. It follows immediately from (10), that the differences of the Granger scores related to original and time-reversed data do not depend on the restricted models:

$$\begin{aligned} \tilde{D}_{y \to x} &= F_{y \to x} - \tilde{F}_{\tilde{y} \to \tilde{x}} = \log \tilde{\Sigma}_{xx} - \log \Sigma_{xx} \\ \tilde{D}_{x \to y} &= F_{x \to y} - \tilde{F}_{\tilde{x} \to \tilde{y}} = \log \tilde{\Sigma}_{yy} - \log \Sigma_{yy} \\ \tilde{D}^{(net)}_{x \to y} &= F^{(net)}_{x \to y} - \tilde{F}^{(net)}_{\tilde{x} \to \tilde{y}} \\ &= (F_{x \to y} - F_{y \to x}) - (\tilde{F}_{\tilde{x} \to \tilde{y}} - \tilde{F}_{\tilde{y} \to \tilde{x}}) \\ &= \log \tilde{\Sigma}_{yy} - \log \tilde{\Sigma}_{xx} - \log \Sigma_{yy} + \log \Sigma_{xx} \ . \end{aligned} \tag{31}$$

The Granger score differences $\tilde{D}_{y \to x}, \tilde{D}_{x \to y}$, and $\tilde{D}^{(net)}_{x \to y}$ thus only depend on the residual covariance matrices of the full models of the original and time-reversed data. For the VAR(1) process, these are given in (25) and (29). For VAR(p) processes, the residual covariance matrices can be obtained through (55) and (58) as described in Appendix A-A and A-B.

Please note that while (31) implies that the unrestricted models can be neglected when computing Granger scores differences, we might gain from including them in finite sample settings. We investigate this issue through simulations in Section III.

F. A minimal example

It is not intuitive to see how the residual variance of the time-reversed process, and thus Granger causality, depends on the autoregressive coefficients of the model. Interpretation is made difficult by the occurrence of $C_{\mathbf{z}}(0)^{-1}$ in (29).

Let us therefore consider the following minimal case: a VAR(1) process $\mathbf{z}_t$ with $C_{\mathbf{z}}(0) = I$. In that case, $C_{\mathbf{z}}(h) = A_1^h$ and $C_{\mathbf{z}}^\top(h) = (A_1^\top)^h$ for all $h \in \mathbb{Z} \setminus \{0\}$ (from (22)). All

asymmetries in the cross-covariance matrices $C_{\mathbf{z}}(h)$ are thus due to asymmetries in A_1.

Furthermore, time-reversing the signal leads to transposition of the autoregressive coefficient matrix $\tilde{A}_1 = A_1^\top$ as a result of (27). The residual covariance matrices (25) and (29) are now given by

$$\Sigma = I - A_1 \cdot A_1^\top \quad \text{and} \quad \tilde{\Sigma} = I - A_1^\top \cdot A_1 \ .$$

Denote with $A_1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ the autoregressive coefficients. We then have

$$\Sigma_{xx} = 1 - a_{11}^2 - a_{12}^2 \quad , \quad \tilde{\Sigma}_{xx} = I - a_{11}^2 - a_{22}^2 \ ,$$

and

$$\begin{aligned} \tilde{D}_{y\to x} = \log \tilde{\Sigma}_{xx} - \log \Sigma_{xx} > 0 &\Leftrightarrow \Sigma_{xx} < \tilde{\Sigma}_{xx} \\ &\Leftrightarrow a_{12}^2 > a_{21}^2 \ . \end{aligned}$$

The difference of the Granger scores computed on the original and time-reversed time series thus indicates the correct *net* direction of information flow. We will in general not be able to infer whether x_t has a Granger-causal influence on y_t. However, we will be able to tell whether x_t Granger-causes y_t more than y_t Granger-causes x_t, or vice versa.

While this simple case will almost never occur in practice, we give theoretical guarantees for more general cases in the next section.

G. Validity of TRGC for unidirectional information flow

We now prove our main result, the validity of difference-based time-reversed Granger causality in the presence of unidirectional information flow. Consider a bivariate VAR(p) process with unambiguous unidirectional information flow. This is the case when all coefficient matrices are triangular and the residual covariance matrix Σ is diagonal. Then the following theorem holds.

Theorem 1. *Let* $\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \mathbb{R}^2$ *be a stable bivariate VAR(p) process* (1) *with the time-reversed representation* (30)*. Under the assumptions*

(A1) $A_1, \ldots, A_p$ *are lower triangular matrices (i.e., x_t may Granger-cause y_t, but y_t does not Granger-cause x_t), and*

(A2) Σ *is a diagonal matrix, i.e.* $\Sigma_{xy} = 0$ *(the residuals are uncorrelated), and*

(A3) $C_{\mathbf{z}}(0)$ *is invertible* ,

it holds that

$$\tilde{\Sigma}_{xx} \leq \Sigma_{xx} \ , \tag{32}$$

and that

$$\tilde{\Sigma}_{yy} \geq \Sigma_{yy} \ . \tag{33}$$

Corollary 1. *Under assumptions* (A1)-(A3)*, Theorem 1 and* (31) *immediately imply the following inequalities for the differences of Granger scores:*

$$\tilde{D}_{y\to x} = F_{y\to x} - \tilde{F}_{\tilde{y}\to\tilde{x}} \leq 0 \tag{34}$$

$$\tilde{D}_{x\to y} = F_{x\to y} - \tilde{F}_{\tilde{x}\to\tilde{y}} \geq 0 \tag{35}$$

$$\tilde{D}^{(net)}_{x\to y} = (F_{x\to y} - F_{y\to x}) - (\tilde{F}_{\tilde{x}\to\tilde{y}} - \tilde{F}_{\tilde{y}\to\tilde{x}}) \geq 0 \ . \tag{36}$$

As a result of Corollary 1, net Granger-causal information flow from x_t to y_t is reduced or remains the same when the signal is time-reversed. Thus, in the case of unambiguous unidirectional information flow, difference-based time-reversed Granger causality yields the correct result. Note that it is not true in general that the net flow between the time-reversed signals $\tilde{x}_t$ and $\tilde{y}_t$, $\tilde{F}^{(net)}_{\tilde{x}\to\tilde{y}}$, is negative (reverses compared to the original series). That is, conjunction-based TRGC might in some cases incorrectly reject the presence of true causal interaction.

Corollary 1 states that each of the three difference scores, $\tilde{D}_{y\to x}$, $\tilde{D}_{x\to y}$, and $\tilde{D}^{(net)}_{x\to y}$ alone is sufficient to infer the correct directionality under assumptions (A1)–(A3). As (A1) requires information flow to be unidirectional, the individual scores $\tilde{D}_{y\to x}$ and $\tilde{D}_{x\to y}$ only indicate *net* information flow, which is what is also observed in Section II-F.

The three scores will behave differently if the assumption of uncorrelated residuals (A2) is violated. Then, $\tilde{\Sigma}_{xx} \leq \Sigma_{xx}$ and $\tilde{D}_{y\to x} \leq 0$ still hold, but the inequalities $\tilde{\Sigma}_{yy} \geq \Sigma_{yy}$, $\tilde{D}_{x\to y} \geq 0$ and $\tilde{D}^{(net)}_{x\to y} \geq 0$ do not. On average, the net difference $\tilde{D}^{(net)}_{x\to y}$ (which equals $\tilde{D}_{x\to y} - \tilde{D}_{y\to x}$) is less affected by the presence of correlations in the residuals than any of the individual scores, which is why we defined difference-based TRGC based on $\tilde{D}^{(net)}_{x\to y}$ in (17). Nevertheless, all three scores are valid measures for net information flow, as residuals should be uncorrelated if the VAR model accurately describes a physical process.

Sketch of the proof. The first inequality (32) is relatively easy to prove. The intuition is the following: Since y_t does not Granger-cause x_t, the prediction of x_t is only based on past x_t. In contrast, the coefficient matrices $\tilde{A}_1, \ldots, \tilde{A}_p$ of the time-reversed representation are in general not triangular. This means that prediction of the time-reversed signals $\tilde{x}_t$ is not only based on past $\tilde{x}_t$, but can also use information from past $\tilde{y}_t$. We would thus expect that $\tilde{x}_t$ can be better predicted than x_t, and that the corresponding residuals are smaller.

The proof of the second inequality (33) is more involved. The intuition is the following: we would expect that the 'amount' of unexplainable variance is the same for both the original and the time-reversed process. Thus, since the residual variance of x_t decreases, the residual variance of y_t should increase. Mathematically, we prove that

$$\det(\Sigma) = \det(\tilde{\Sigma}) \ . \tag{37}$$

The proof of (37) is the only part that requires the analytic description of $\tilde{\Sigma}$, and is the main difficulty of the overall proof. It is not straightforward, because $\tilde{\Sigma}$ depends on the inverse of the covariance matrix $C_{\mathbf{z}}(0)$, while we only have an analytic description of $\operatorname{vec} C_{\mathbf{z}}(0)$. From (37), it is easy to infer $\tilde{\Sigma}_{yy} \leq \Sigma_{yy}$, which completes the proof. It is only in this final step that we need assumption (A2) that Σ is diagonal.

Proof (Part 1: Proof that $\tilde{\Sigma}_{xx} \leq \Sigma_{xx}$):

As $A_1, \ldots, A_p$ are lower triangular matrices (assumption (A1)), x_t is an autoregressive process of order p,

$$x_t = a_1 x_{t-1} + \ldots + a_p x_{t-p} + \xi_t^x \quad \text{with} \quad \operatorname{Var}(\xi_t^x) = \Sigma_{xx} \tag{38}$$

Its time-reversed representation (cf. Section II-B) is

$$x_t = a_1 x_{t+1} + \ldots + a_p x_{t+p} + \tilde{\xi}_t^x, \quad \text{with } \mathrm{Var}(\tilde{\xi}_t^x) = \Sigma_{xx} \quad (39)$$

Because the unrestricted (or full) model (30) extends the restricted model by including y_t, (32) follows:

$$\tilde{\Sigma}_{xx} \leq \mathrm{Var}(\tilde{\xi}_t^x) = \Sigma_{xx} \ . \quad (40)$$

□

Proof (Part 2: Proof that $\tilde{\Sigma}_{yy} \leq \Sigma_{yy}$):

As mentioned in the proof sketch, we need to derive (37), the equality of the determinants $\det \Sigma$ and $\det \tilde{\Sigma}$. To improve readability, we here treat only the case $p = 1$, and derive (37) for general $p \in \mathbb{N} \setminus \{0\}$ in Appendix A-C.

The proof relies on Sylvester's determinant theorem [31], which states that for any matrices $K \in \mathbb{R}^{n \times m}$, $L \in \mathbb{R}^{m \times n}$:

$$\det(I + KL) = \det(I + LK) \ . \quad (41)$$

We then have:

$$\begin{aligned}
\det \Sigma &\overset{(21)}{=} \det(C_{\mathbf{z}}(0) - A_1 \cdot C_{\mathbf{z}}(0) \cdot A_1^\top) \\
&= \det(C_{\mathbf{z}}(0)) \cdot \det(I - A_1 \cdot C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1}) \\
&\overset{(41)}{=} \det(C_{\mathbf{z}}(0)) \cdot \det(I - C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1} A_1) \\
&= \det(C_{\mathbf{z}}(0) - C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1} A_1 \cdot C_{\mathbf{z}}(0)) \\
&\overset{(29)}{=} \det \tilde{\Sigma} \ .
\end{aligned}$$

From the result of *Part 1* (32), the equality of residual covariance determinants (37) (derived for general p in Appendix A-C), and assumption (A2) of uncorrelated residuals in Σ, we then obtain:

$$\begin{aligned}
\Sigma_{xx}\Sigma_{yy} \overset{(A2)}{=} \det \Sigma \overset{(37)}{=} \det \tilde{\Sigma} &= \tilde{\Sigma}_{xx}\tilde{\Sigma}_{yy} - \tilde{\Sigma}_{xy}\tilde{\Sigma}_{xy} \\
&\leq \tilde{\Sigma}_{xx}\tilde{\Sigma}_{yy} \\
&\overset{(32)}{\leq} \Sigma_{xx}\tilde{\Sigma}_{yy} \\
\Leftrightarrow \Sigma_{yy} &\leq \tilde{\Sigma}_{yy} \ .
\end{aligned}$$

□

III. Experiments

In this section, we provide an empirical investigation of model violations and other factors influencing the performance of Granger causal measures using numerical simulations. After describing the tested methods and performance measures (Section III-A), we compare several variants of TRGC in either the presence or absence of noise (Section III-B). We then investigate the influence of common drivers, various types of noise (Section III-C and III-D) and downsampling (Section III-E) on standard Granger causality and Diff-TRGC.

A. Experimental setup

We consider bivariate time series in the presence of unidirectional information flow ($x_t \to y_t$) as well as in the absence of causal interaction. Unless otherwise stated, time series of length $T = 2000$ are generated from stationary VAR(5) processes, whose autoregressive coefficients are drawn from a normal distribution with mean 0 and standard deviation $\sigma_A = 0.2$. The absence of causal interaction is modeled by setting respective AR coefficients to zero. Residuals are generated from a normal distribution with diagonal covariance matrix, whose entries are drawn from the standard uniform distribution.

We compare standard GC as well as Net-GC to Diff-TRGC (see (17)). In Section III-B, we also include Conj-TRGC (see (16)), the conjunction of Net-GC and Diff-TRGC (see (18)), and a variation of Diff-TRGC, in which $\tilde{D}_{x \to y}^{(net)}$ is computed using only the full bivariate models according to (31). This variant is denoted by *Diff-TRGC (full)*.

All statistical tests are performed at significance level $\alpha = 0.05$. For standard GC, we perform two separate F-tests, one to assess whether x_t Granger-causes y_t, and one to assess whether y_t Granger-causes x_t. It is possible that both variables are estimated to Granger-cause each other. In contrast, all other metrics indicate net directionality. We assess their statistical significance by bootstrapping residuals from the regression model: We regress $\mathbf{z}_t$ on its past and future values $\mathbf{z}_{t-p}, \ldots, \mathbf{z}_{t-1}, \mathbf{z}_{t+1}, \ldots \mathbf{z}_{t+p}$, and retain the fitted values $\hat{\mathbf{z}}_t$ and residuals $\hat{\epsilon}_t := \mathbf{z}_t - \hat{\mathbf{z}}_t$. In each bootstrap repetition, causality metrics are computed on synthetic variables $\mathbf{z}_t^* := \hat{\mathbf{z}}_t + \hat{\epsilon}_s$, where s is selected randomly for each t. Percentile confidence intervals are then constructed from the bootstrap sampling distribution. Significance is determined by evaluating if the confidence interval does not contain 0. We use 500 bootstrap samples and select the number of lags p as the optimizer of Schwarz's Bayesian Information Criterion (BIC) [32].

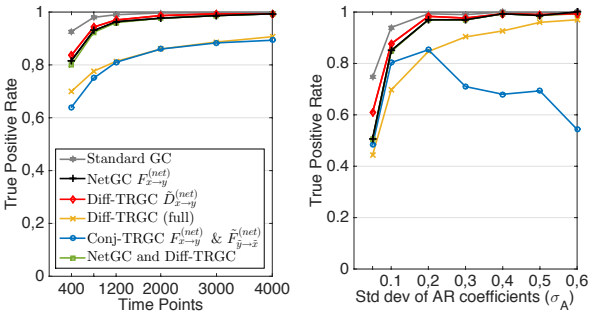
All experiments are repeated 300 times. In each run, a true positive (TP) is defined as a significant detection of the true direction of interaction. The *true positive rate* (TPR) is the fraction of true positives among all runs. It is here also referred to as the *sensitivity* or *power*. A false positive (FP) is defined as a significant detection of the wrong direction of interaction, or a significant detection of causal interaction in the absence of any causal interaction. The *false positive rate* (FPR) is the fraction of false positives among all tested runs.

B. Comparison of TRGC variants under interaction

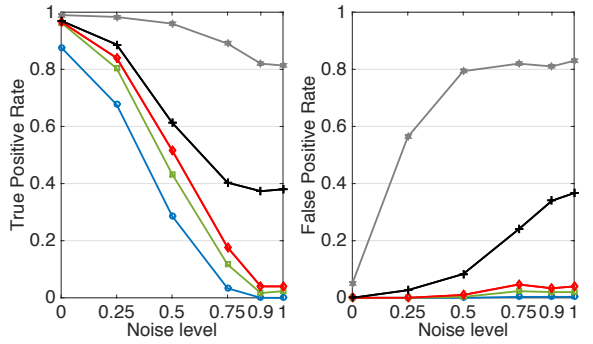
We assess Granger causality and time-reversed Granger causality in the presence of unidirectional interaction considering differing sample sizes, standard deviations of the AR parameters, noise types and signal-to-noise ratios (SNR).

In a first experiment, we consider the noiseless case, and vary the sample size from 400 to 4000 for a fixed standard deviation $\sigma_A = 0.2$ of the AR coefficients. In a second experiment, we vary the standard deviation σ_A at a constant sample size of $T = 2000$. This experiment thus tests the impact of the strength of the causal connections relative to the innovation noise. The standard deviations tested are 0.05, 0.1, 0.2, ..., and 0.6. Finally, for a fixed standard deviation $\sigma_A = 0.2$, and a fixed sample size $T = 2000$, we add linearly mixed, autocorrelated measurement noise $\eta_t \in \mathbb{R}^2$ to each system according to

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = (1 - \gamma) \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} + \gamma \cdot \eta_t \ , \quad (42)$$



(a) No measurement noise



(b) Linearly mixed, auto-correlated noise

Fig. 1. Performance of Granger causality and different variants of time-reversed Granger causality (TRGC). (a) True positive rate in the noiseless case as a function of the number of samples T for fixed standard deviation $\sigma_A = 0.2$ of the AR coefficients, and as a function of σ_A for fixed $T = 2000$. (b) True and false positive rates as a function of the SNR for additive mixed autocorrelated noise (according to (42)) for $T = 2000$ and $\sigma_A = 0.2$.

where the subscript $^{(l)}$ denotes the underlying *latent* variables and γ defines the signal-to-noise ratio (SNR). Noise η_t is generated by multiplying two independent AR(5) time-series with a random matrix B, with $\det(B) = 1$. We consider the signal-to-noise ratios 0, 0.25, 0.5, 0.75, 0.9 and 1.

The TP and FP rates attained in the three experiments are depicted in Figure 1. From Figure 1(a), we see that Diff-TRGC (full), which computes the difference score $\tilde{D}_{x\to y}^{(net)}$ only using the full model according to (31), seems to be suboptimal for finite samples. While we have demonstrated the equivalence of (31) to the original definition (15) for infinite samples in Section II-E, this equivalence does not hold for the finite samples studied here. Estimating residuals from the restricted models increased the power of the test for all investigated parameter settings.

Conj-TRGC has lower power relative to Diff-TRGC. This is particularly so for high σ_A, which corresponds to a dominance of the dynamical and causal aspects of the model comprised in the AR coefficients relative to the innovation noise. This result is not unexpected, as time-reversing the signals does not necessarily reverse the direction of information flow. Note that, on the other hand, Conj-TRGC is the more conservative measure compared to Diff-TRGC and could be expected to produce fewer spurious results in the presence of noise. However, as we see in Figure 1(b), both variants yield almost no spurious results in the presence of measurement noise. We will therefore use Diff-TRGC in the remaining experiments.

C. Impact of latent variables and measurement noise in the absence of causal interaction

Already Granger pointed out that standard Granger causality can lead to spurious results if not all relevant variables are incorporated in the model [8]. In a bivariate system, GC cannot determine whether the observed variables x_t and y_t are both driven by a third common cause. This argument extends to multivariate systems, if a relevant confounding variable is not part of the measurement. Furthermore, standard GC is susceptible to *measurement noise* [33], [10], [11], [26], [34], [18] and to *instantaneous linear mixing* of activity, which is a major problem for example in the analysis of electroencephalographic (EEG) recordings [13], [14], [16]. We demonstrate these effects here in additional simulations, in all of which no actual interaction occurs. We consider three different scenarios.

(A) Linear mixing. The observed time series x_t and y_t are a linear mixture of two independent signals $x_t^{(l)}$, $y_t^{(l)}$, that is

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = M \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} , \tag{43}$$

where $M \in \mathbb{R}^{2\times 2}$ denotes the mixing matrix. $x_t^{(l)}$ and $y_t^{(l)}$ were generated as two independent univariate AR(5) processes.

(B) Common hidden cause. The observed time series x_t and y_t are driven by a common unobserved cause g_t. Time series x_t, y_t, and g_t are generated from a three-dimensional VAR(5) model with $\sigma_A = 0.3$, in which g_t Granger-causes x_t and y_t, with no causal interaction between x_t and y_t as modeled by the respective AR coefficients being set to zero.

(C) Additive noise. The observed time series x_t and y_t are a superposition of two independent univariate AR(5) processes $x_t^{(l)}$, $y_t^{(l)}$ and additive noise η_t as in (42), with $\gamma \in \{0, 0.25, 0.5, 0.75, 0.9, 1\}$ adjusting the SNR. We consider three different types of noise. *Independent white noise* is generated from a normal distribution with diagonal covariance matrix, whose entries are drawn from the standard uniform distribution. *Mixed white noise* is created by multiplying independent noise with a random matrix B with $\det(B) = 1$. *Mixed autocorrelated noise* is created by multiplying two independent AR(5) time-series with B.

Figure 2 illustrates the behavior of standard Granger causality, Net-GC and Diff-TRGC in the various simulation settings. Values on the y-axis indicate the FP rate at significance level $\alpha = 0.05$. As all experiments are characterized by the absence of any interaction between x_t and y_t, any significant detection of information flow either from x_t to y_t or y_t to x_t is counted as a false positive.

It is apparent from Figure 2 that standard GC and Net-GC lead to spurious detection of causality in all tested scenarios. Their behavior in the presence of noise (panel C) depends on the properties of that noise. Mixed noise (left and center plots of panel C) is generally very problematic, especially if

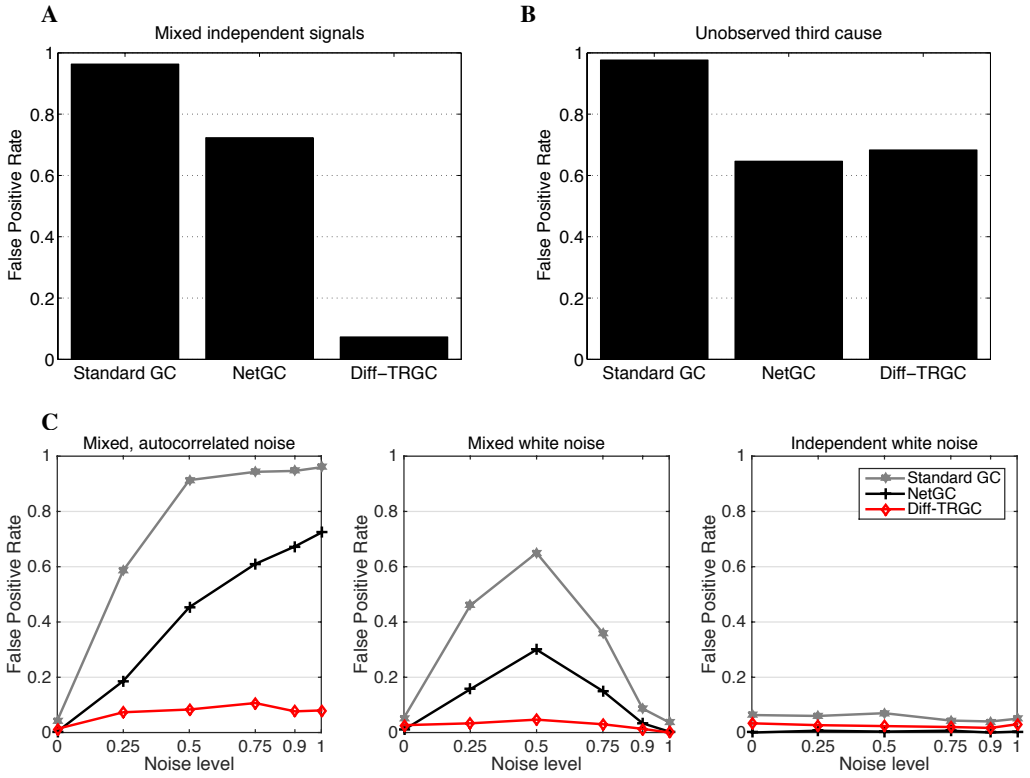


Fig. 2. False positive rates of Granger causality (standard GC and Net-GC) and difference-based time-reversed Granger causality (Diff-TRGC) as a function of the SNR for two signals lacking any causal connection. **(A)** Instantaneous linear mixture of two independent univariate AR(5) processes. **(B)** Common unobserved cause. x_t and y_t. **(C)** Superposition of two independent univariate AR(5) processes with additive Gaussian noise.

it is also autocorrelated (left part). As x_t and y_t are already independent, adding independent noise (obviously) does not pose a problem here (right part of panel C).

In contrast to standard GC and Net-GC, time-reversed Granger causality implemented through Diff-TRGC is insensitive to mixtures of independent sources regardless of their spatial and temporal correlation structure (see panels A and C). This behavior thus reflects its known theoretical properties discussed in Section II-C. The presence of a hidden common confounder, however, cannot be ruled out by using time-reversed Granger causality (panel B).

D. Impact of noise in the presence of causal interaction

We further study the behavior of standard GC, Net-GC and Diff-TRGC in the presence of unidirectional causal interactions superimposed with noise. Four different scenarios are considered. In all cases, data are generated according to (42) with $x_t^{(l)}$ Granger-causing $y_t^{(l)}$. In the first three scenarios, (A-C), interacting signals from bivariate VAR(5) models are superimposed with noise. As in Section III-C, we use mixed autocorrelated noise (scenario A), mixed white noise (B), and independent white noise (C). The same signal to noise ratios as in Section III-C are used.

In the fourth scenario, (D), we simulate the following VAR(1) process with long memory:

$$\begin{aligned} \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} &= \begin{bmatrix} 0.95 & 0 \\ 1 & 0.5 \end{bmatrix} \begin{bmatrix} x_{t-1}^{(l)} \\ y_{t-1}^{(l)} \end{bmatrix} + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, I) \\ x_t &= (1-\gamma) \cdot x_t^{(l)} + \gamma \cdot \eta_t \qquad\quad \eta_t \sim \mathcal{N}(0,1) \\ y_t &= y_t^{(l)} \ , \end{aligned} \tag{44}$$

adopted from [26], where $\mathcal{N}$ denotes the normal distribution.

True positive and false positive rates as estimated from 300 simulation runs are reported in Figure 3 (A-D). Just as in the absence of causality (cf. Section III-C), we observe that linearly mixed, autocorrelated noise leads to the highest numbers of false detections for standard GC, while independent white noise leads to lowest FP rates. Diff-TRGC is characterized by negligible amounts of false positives in all cases at the cost of slightly decreased sensitivity as compared to standard GC in scenarios (A-C). Interestingly, Net-GC behaves very similar to Diff-TRGC in the presence of non-autocorrelated noise both in terms of sensitivity and specificity (B-C). In these settings, spurious causality could already be almost entirely eliminated by testing for net Granger causality. This result, however, does not imply that Net-GC cannot be affected by non-autocorrelated noise in general. A counterexample is the system with long memory studied in scenario (D). Here, Net-GC (as well as standard GC) fails, because y_t contains delayed

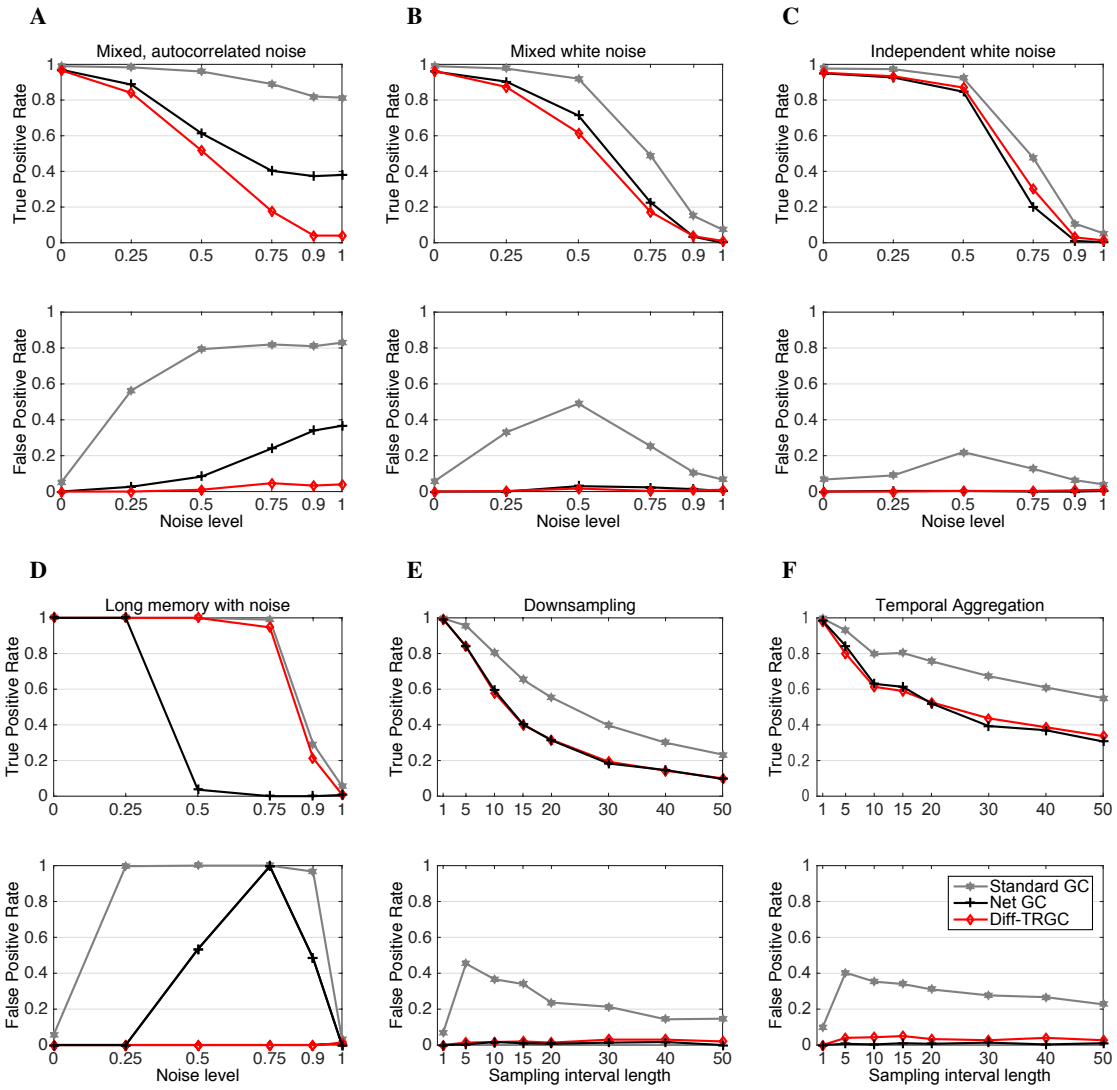


Fig. 3. Performance of Granger causality (standard GC and Net-GC) and difference-based time-reversed Granger causality (Diff-TRGC) for two signals with unidirectional information flow from x_t to y_t. Shown are the fractions of true positives ($x_t \to y$ detected) and false positives ($y_t \to x_t$ detected), when x_t and y_t are corrupted by noise (A-D), downsampling (E), and temporal aggregation (F). The underlying latent signals $x^{(l)}$ and $y^{(l)}$ were generated from VAR(5) processes with random AR coefficients, except for D, in which signals follow a VAR(1) process with long memory according to (44).

but cleaner information about $x_t^{(l)}$ than x_t itself and thus may help to predict future x_t. Diff-TRGC, however, robustly identifies x_t as the driver.

Our examples show time-reversed Granger causality almost completely eliminates spurious causalities arising from any kind of additive noise. At the same time, it exhibits similar statistical power as net Granger causality. We also observe that net Granger causality is typically more robust with respect to additive noise than standard Granger causality.

E. Impact of downsampling and temporal aggregation

Spurious Granger causality has also been reported to arise due to downsampling and temporal aggregation [35], [36], [37], posing serious problems, for example, in functional magnetic resonance imaging (fMRI) [38], [39].

We generate data using a VAR(5) model with random coefficients with $\sigma_A = 0.3$, in which x_t Granger-causes y_t. These data are decimated at different factors τ in two ways. In the downsampling scenario (E), causal measures are applied to time series of length $T = 2000$ constructed from the original time series by skipping $\tau - 1$ time points in between sampled data points. In the temporal aggregation scenario (F), time series of length $T = 2000$ are constructed from the original time series by averaging over τ data points. No noise was added.

Figure 3 (E-F) depicts TP and FP rates attained in the two scenarios as a function of τ. We see that Net-GC and Diff-

TRGC are more robust then standard GC. Both Net-GC and Diff-TRGC did not result in spurious causality.

IV. Discussion

We established the theoretical guarantee that difference-based time-reversed Granger causality (Diff-TRGC) indicates the correct direction of causality in bivariate autoregressive processes characterized by unambiguous unidirectional information flow. Our results complement previous work by [16], [17] showing that TRGC in general correctly rejects causal interpretations for mixtures of non-interacting sources (thus, in the absence of any causality). While further compelling intuitive ideas for robust causality measures have been presented [11], [23], [18], our result provides, to the best of our knowledge, the first proof of the correctness of one of such techniques (Diff-TRGC) for a relatively general class of time-series models.

Our theory is accompanied by simulations, in which we confirmed that time-reversed Granger causality robustly detects the presence of true causal interactions in various realistic scenarios including mixed noise and downsampling. We showed that Diff-TRGC is able to infer correct directionality with similar power as net Granger causality, while at the same time producing fewer (in most cases, negligible amounts of) false alarms than Net-GC and standard GC. We therefore suggest to use Diff-TRGC whenever the data under study are likely to be corrupted by noise.

A. Correlated residuals

To define an unambiguous uni-directional information flow, our theory assumes uncorrelated residuals, as is common in the literature. Correlated residuals indicate instantaneous effects that the variables exert on each other. While we would not expect correlated residuals if the VAR model accurately describes the data generating process, such effects are likely to occur in practice (e.g., if the sampling rate of the acquired data falls below the time scale of the causal interactions). They pose severe problems for causal estimation, because they can be explained by several possible data generating models, the coefficients of which cannot be uniquely identified using second order information only.

Data generating models. Instantaneous interactions can be modeled implicitly through correlated residuals in classical VAR processes, or explicitly, for example using so-called 'structural' VAR (SVAR) processes [1], [40], [41]. By augmenting the VAR model with an instantaneous mixing matrix Γ_0, the SVAR model

$$\mathbf{z}_t = \sum_{h=0}^{p} \Gamma_h \mathbf{z}_{t-h} + \bar{\epsilon}_t \ , \tag{45}$$

achieves that the residuals $\bar{\epsilon}_t$ are uncorrelated. Here, the diagonal of Γ_0 is assumed to be zero.

Correlated residuals emerge naturally in electrophysiological neuroimaging data, where the signals observable at the sensors (e.g., EEG electrodes) are a linear mixture of the latent activity of possibly interacting neuronal populations within the brain. A model for such mixtures of potentially interacting sources is given by

$$\mathbf{z}_t = M\mathbf{z}_t^{(l)} \ , \quad \mathbf{z}_t^{(l)} \ = \sum_{h=1}^{p} B_h \mathbf{z}_{t-h}^{(l)} + \acute{\epsilon}_t \ , \tag{46}$$

where $\mathbf{z}_t \in \mathbb{R}^d$ denotes the observed data, $\mathbf{z}_t^{(l)} \in \mathbb{R}^d$ denotes the activity of underlying latent variables (e.g., brain sources) following a VAR(p) process with uncorrelated residuals $\acute{\epsilon}_t$, and $M \in \mathbb{R}^{d\times d}$ is an unknown mixing matrix (representing, e.g., the volume conduction effect of the human head). We call (46) the 'mixture of interacting sources' model.

Note that VAR models with correlated residuals, SVAR models, and mixture of interacting sources models can be used interchangeably to represent the same statistical process. For example, an interacting sources model (46) can be equivalently written as a VAR(p) process (1) with coefficients

$$A_h = MB_hM^{-1} \quad , \quad h \in \{1,\dots,p\} \tag{47}$$

and correlated residuals $\epsilon_t = M\tilde{\epsilon}_t$. Likewise, an SVAR($p$) process (45) can be converted into a VAR(p) process with correlated residuals $\epsilon_t = (I-\Gamma_0)^{-1}\bar{\epsilon}_t$ and coefficients

$$A_h = (I-\Gamma_0)^{-1}\Gamma_h \quad , \quad h \in \{1,\dots,p\} \ . \tag{48}$$

The reverse transformations from VAR models to SVAR or interacting source models, as well as the transformations between SVAR and interacting source models, are, however, not unique (see *Model identifiability*).

Ambiguous causal interpretations can emerge in cases where one of the three models indicates time-delayed causal interactions through non-zero off-diagonal coefficients in the A_h, B_h or Γ_h, while another one does not. This ambiguity can in general only be resolved if the model generating the data is known a-priori. In case of EEG data, for example, (46) reflects the true data-generating process. Therefore, only the parameters B_h of the source VAR process (46) permit meaningful causal interpretation (wrt. to the source variables $\mathbf{z}_t^{(l)}$), while, for example, the VAR parameters in (47) are distorted by the mixing matrix M.

Model identifiability. A further complication in the presence of instantaneous effects in the data is that for mixture of interacting sources as well as SVAR models, the parameters are not uniquely defined from second order information only. This can be best seen for the latter model (46). Identifying the model parameters requires the estimation of a full factorization of the data into a mixing matrix M and source time series $\mathbf{z}_t^{(l)}$. This means that the estimation problem falls into the blind source separation (BSS) setting, in which Gaussianity of the factors is not sufficient for their identification. The classical approach to BSS, independent component analysis (ICA) assumes statistical independence and non-Gaussianity of the sources $\mathbf{z}_t^{(l)}$ to ensure identifiability. This concept can be adopted in the context of source AR models by enforcing independence/non-Gaussianity of the residuals of the source AR process in (46) [13], [15], [42]. In a similar way, independence of residuals has been used in the identification of SVAR models [40], [43].

Example. Consider the following VAR(1) process with correlated residuals:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ -0.12 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \epsilon_t, \quad \langle \epsilon_t \epsilon_t^\top \rangle = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} .$$

This process can also be represented by the SVAR(1) model

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0.6 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0.7 & 0 \\ -0.54 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \bar{\epsilon}_t$$

as well as the mixtures of interacting sources model

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.6 & 0.8 \end{bmatrix} \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix}, \quad \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1}^{(l)} \\ y_{t-1}^{(l)} \end{bmatrix} + \acute{\epsilon}_t,$$

with uncorrelated residuals $\langle \bar{\epsilon}_t \bar{\epsilon}_t^\top \rangle = \begin{bmatrix} 1 & 0 \\ 0 & 0.64 \end{bmatrix}$, $\langle \acute{\epsilon}_t \acute{\epsilon}_t^\top \rangle = I$. Note that both the VAR(1) and the SVAR(1) representation indicate unidirectional causal interaction between the observed variables x_t and y_t, whereas the mixture model suggests that the observed data can also arise from a mixture of two independent latent sources $x_t^{(l)}$ and $y_t^{(l)}$. However, another equivalent mixture model

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} -\sqrt{0.2} & \sqrt{0.8} \\ \sqrt{0.2} & \sqrt{0.8} \end{bmatrix} \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix}, \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} = \begin{bmatrix} 0.86 & 0.08 \\ 0.08 & 0.74 \end{bmatrix} \begin{bmatrix} x_{t-1}^{(l)} \\ y_{t-1}^{(l)} \end{bmatrix} + \tilde{\tilde{\epsilon}}_t$$

with $\langle \tilde{\tilde{\epsilon}}_t \tilde{\tilde{\epsilon}}_t^\top \rangle = I$ suggests bidirectional informational flow on the source level. Similarly, the following SVAR(1) model indicates bidirectional flow

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0 & 0.6 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0.772 & -0.54 \\ -0.12 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \bar{\bar{\epsilon}}_t \quad ,$$
$$\langle \bar{\bar{\epsilon}}_t \bar{\bar{\epsilon}}_t^\top \rangle = \begin{bmatrix} 0.64 & 0 \\ 0 & 1 \end{bmatrix}.$$

B. Future work

Further effort is required to investigate the behavior of TRGC in the presence of bidirectional information flow, and to extend the theoretical analysis of time-reversal to general multivariate signals. Furthermore, it would be desirable to obtain theoretical guarantees for the performance of TRGC in the presence of true interaction superimposed by noise in the form of bounds on the false positive rate. A major difficulty here is to obtain the residual covariance of the superposition of a VAR process and additive noise. Analytically computing Granger causality in the presence of noise is mathematically involved even for special cases [44].

Finally, [16] showed that for any causality measure based on cross-covariances, differences of the scores obtained on the original and time-reversal signals correctly indicate the absence of causality on mixtures of independent sources. While we focused here on Granger causality, it remains to be shown whether validity of time-reversal in the presence of causal interaction can also be demonstrated for other causality measures.

APPENDIX A
PROOFS FOR VAR(p)

A. The VAR(p) process and its cross-covariance function

Consider a stable bivariate VAR(p) process, $\mathbf{z}_t \in \mathbb{R}^d$, as defined in (1),

$$\mathbf{z}_t = A_1 \mathbf{z}_{t-1} + A_2 \mathbf{z}_{t-2} + \ldots + A_p \mathbf{z}_{t-p} + \epsilon_t \ ,$$

where $\epsilon_t \in \mathbb{R}^2$ is a 2-dimensional white noise process (i.e. $\langle \epsilon_t \rangle = 0$, $\langle \epsilon_t \epsilon_{t-h}^\top \rangle = 0$ for $h \in \mathbb{Z} \setminus \{0\}$, and $\langle \epsilon_t \mathbf{z}_{t-h}^\top \rangle = 0$ for $h \in \mathbb{N} \setminus \{0\}$) with residual covariance matrix $\Sigma = \langle \epsilon_t \epsilon_t^\top \rangle$.

Many results on VAR(1) processes can be extended to higher order VAR(p) processes by considering their VAR(1) form. Given the 2-dimensional VAR(p) process $\mathbf{z}_t$, the corresponding $2p$-dimensional VAR(1) representation is defined as

$$Z_t = \mathbf{A} Z_{t-1} + E_t \ , \tag{49}$$

with

$$Z_t = \begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}_{t-1} \\ \vdots \\ \mathbf{z}_{t-p+1} \end{bmatrix}, \mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ & & \ddots & & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}, E_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} ,$$

and residual covariance matrix

$$\Sigma_E = \langle E_t E_t^\top \rangle = \begin{bmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} . \tag{50}$$

The cross-covariances of Z_t are linked to the cross-covariances of $\mathbf{z}_t$ through

$$C_Z(h) = \begin{bmatrix} C_{\mathbf{z}}(h) & C_{\mathbf{z}}(h+1) & \cdots & C_{\mathbf{z}}(h+p-1) \\ C_{\mathbf{z}}(h-1) & C_{\mathbf{z}}(h) & \cdots & C_{\mathbf{z}}(h+p-2) \\ \vdots & \vdots & \ddots & \vdots \\ C_{\mathbf{z}}(h-p+1) & C_{\mathbf{z}}(h-p+2) & \cdots & C_{\mathbf{z}}(h) \end{bmatrix} \tag{51}$$

for all $h \in \mathbb{Z}$. The Yule-Walker equations can then be expressed as

$$C_Z(0) = \mathbf{A} \cdot C_Z(0) \cdot \mathbf{A}^\top + \Sigma_E \quad \text{and} \tag{52}$$
$$C_Z(h) = \mathbf{A} \cdot C_Z(h-1) \qquad \forall h \in \mathbb{N} \setminus \{0\} \ . \tag{53}$$

Given $A_1, \ldots, A_p$, and Σ, the cross-covariances are uniquely determined: Equation (52) implies that $\mathrm{vec}(C_Z(0)) = (I - \mathbf{A} \otimes \mathbf{A})^{-1} \mathrm{vec}\, \Sigma_E$, while $C_Z(h)$ for $h > 1$ can be recursively computed using (53). Conversely, A_1, …, A_p and Σ are uniquely determined by the cross-covariances through

$$[A_1, A_2, \cdots, A_p] = [C_{\mathbf{z}}(1), C_{\mathbf{z}}(2), \cdots, C_{\mathbf{z}}(p)] \cdot C_Z(0)^{-1} \tag{54}$$

and

$$\Sigma = C_{\mathbf{z}}(0) - [A_1, A_2, \cdots, A_p] \cdot C_Z(0) \cdot [A_1, A_2, \ldots, A_p]^\top \ . \tag{55}$$

B. The time-reversed VAR(p) process

The results of Bartlett on the analytical description of time-reversed VAR(1) processes have been generalized to VAR(p) processes by Andel in 1972 [29]. Given a 2-dimensional VAR(p) process $\mathbf{z}_t$ as in (1), Andel considers a second VAR(p) process

$$\mathfrak{z}_t = \tilde{A}_1 \mathfrak{z}_{t-1} + \tilde{A}_2 \mathfrak{z}_{t-2} + \ldots + \tilde{A}_p \mathfrak{z}_{t-p} + \mathfrak{e}_t \ , \tag{56}$$

where $\mathfrak{e}_t$ is white noise with covariance matrix $\tilde{\Sigma} = \langle \mathfrak{e}_t \mathfrak{e}_t^\top \rangle$.

Now, denote with $Q := C_Z(0)^{-1}$ the inverse of the covariance of Z_t, with block matrix notation

$$Q =: (Q_{lk})_{l,k=1}^{p} = \begin{bmatrix} Q_{1,1} & Q_{1,2} & \cdots & Q_{1,p} \\ Q_{2,1} & Q_{2,2} & \cdots & Q_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{p,1} & Q_{p,p-1} & \cdots & Q_{p,p} \end{bmatrix} \in \mathbb{R}^{2p\times 2p} ,$$

where Q_{lk} are 2×2 blocks.

Andel proves that $C_{\mathfrak{z}}(h) = C_{\mathbf{z}}(-h)$ for all $h \in \mathbb{Z}$ (that is, $\mathfrak{z}_t$ has the same cross-covariance matrices as $\mathbf{z}_t$ reversed in time), if and only if $\tilde{A}_1, \ldots, \tilde{A}_p$ and $\tilde{\Sigma}$ are defined as follows: for $1 \le j \le p$,

$$\tilde{A}_j = -(Q_{pp} + A_p^\top \Sigma^{-1} A_p)^{-1}(Q_{p,p-j} + A_p^\top \Sigma^{-1} A_{p-j}) \quad (57)$$

and

$$\tilde{\Sigma} = (Q_{pp} + A_p^\top \Sigma^{-1} A_p)^{-1} , \quad (58)$$

where $Q_{p,0} := 0$ and $A_0 := -I$. Andel further proves that $\tilde{A}_p \neq 0$, if and only if $A_p \neq 0$, and that, if $\mathbf{z}_t$ is stable, so is $\mathfrak{z}_t$.

Note that, while we only treat bivariate VAR processes in this paper, the analytic description reviewed above holds for arbitrary dimensionality.

C. Proof that $\det(\Sigma) = \det(\tilde{\Sigma})$ for general p – this completes the proof of Theorem 1

Given Andel's result, we can complete the proof for Theorem 1. The only missing part of the proof (cf. Section II-G) is the proof of (37), $\det(\Sigma) = \det(\tilde{\Sigma})$, for arbritrary $p \in \mathbb{N} \setminus \{0\}$.

Preliminaries. We use that the following statements generally hold for block matrices: Let K be a positive definite matrix with $L = K^{-1}$, and let

$$K = \begin{bmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{bmatrix}, \qquad L = \begin{bmatrix} L_{1,1} & L_{1,2} \\ L_{2,1} & L_{2,2} \end{bmatrix}$$

the block matrix notations of L and K, where $K_{1,1}$ is a square matrix of the same size as $L_{1,1}$. Then (see e.g. [45], p. 22),

$$L_{2,2} = \left[K_{2,2} - K_{2,1}K_{1,1}^{-1}K_{1,2}\right]^{-1} . \quad (59)$$

Let T and W be invertible matrices, then for all matrices U and V of fitting size

$$\det(T + UWV) = \det(W^{-1} + VT^{-1}U)\det(T)\det(W) . \quad (60)$$

This relation is known as the generalized *matrix determinant lemma* and a straightforward extension of Sylvester's determinant theorem (41).

Let K be a matrix with block notation as above, and $K_{1,1}$ be invertible, then, (see e.g. [45], p. 22),

$$\det(K) = \det(K_{1,1})\det(K_{2,2} - K_{2,1}K_{1,1}^{-1}K_{1,2}) . \quad (61)$$

Let us also introduce the following notation for the blocks of $C_Z(0)$:

$$C_Z(0) = \begin{bmatrix} C_{Z\setminus p} & R^\top \\ \bar{R} & C_{\mathbf{z}}(0) \end{bmatrix} = \begin{bmatrix} C_{\mathbf{z}}(0) & \bar{R} \\ \bar{R}^\top & C_{Z\setminus p} \end{bmatrix} ,$$

where we define

$$R := \left[C_{\mathbf{z}}(p-1)^\top \; C_{\mathbf{z}}(p-2)^\top \; \cdots \; C_{\mathbf{z}}(1)^\top\right] \in \mathbb{R}^{2\times 2(p-1)}$$
$$\bar{R} := \left[C_{\mathbf{z}}(1) \; \ldots \; C_{\mathbf{z}}(p-1)\right] \in \mathbb{R}^{2\times 2(p-1)} ,$$

and

$$C_{Z\setminus p} := \begin{bmatrix} C_{\mathbf{z}}(0) & \cdots & C_{\mathbf{z}}(p-2) \\ \vdots & \ddots & \vdots \\ C_{\mathbf{z}}(2-p) & \cdots & C_{\mathbf{z}}(0) \end{bmatrix} \in \mathbb{R}^{2(p-1)\times 2(p-1)} .$$

Step 1: Analytic expression for $C_{\mathbf{z}}(0) - \bar{R}C_{Z\setminus p}^{-1}\bar{R}^\top$.

We first prove that

$$C_{\mathbf{z}}(0) - \bar{R}C_{Z\setminus p}^{-1}\bar{R}^\top = \Sigma + A_p Q_{pp}^{-1} A_p^\top , \quad (62)$$

that is, the residual variance when regressing z_t on $z_{t-1}, \ldots, z_{t-(p-1)}$ given by $C_{\mathbf{z}}(0) - \bar{R}C_{Z\setminus p}^{-1}\bar{R}^\top$ can be expressed as the sum of $A_p Q_{pp}^{-1} A_p$ and the residual variance when regressing z_t on $z_{t-1}, \ldots, z_{t-(p-1)}, z_{t-p}$, given by Σ.

Recall the Yule-Walker equation (52)

$$C_Z(0) = \mathbf{A} \cdot C_Z(0) \cdot \mathbf{A}^\top + \Sigma_E .$$

and let us rewrite

$$\Sigma_E = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_{\setminus p} & A_p \\ I & 0 \end{bmatrix},$$

where we define

$$\mathbf{A}_{\setminus p} := \begin{bmatrix} A_1 \; \ldots \; A_{p-1} \end{bmatrix} \in \mathbb{R}^{2\times 2(p-1)} .$$

The Yule-Walker equation can then be written in blocks as

$$\begin{bmatrix} C_{\mathbf{z}}(0) & \bar{R} \\ \bar{R}^\top & C_{Z\setminus p} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{\setminus p} & A_p \\ I & 0 \end{bmatrix} \begin{bmatrix} C_{Z\setminus p} & R^\top \\ R & C_{\mathbf{z}}(0) \end{bmatrix} \begin{bmatrix} \mathbf{A}_{\setminus p}^\top & I \\ A_p^\top & 0 \end{bmatrix} + \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} .$$

We see from the top line that

$$\begin{aligned} & \bar{R} = \mathbf{A}_{\setminus p} C_{Z\setminus p} + A_p R \\ \Leftrightarrow \quad & A_p R = \bar{R} - \mathbf{A}_{\setminus p} C_{Z\setminus p} , \end{aligned} \quad (63)$$

and that

$$\begin{aligned} & C_{\mathbf{z}}(0) \\ & = \mathbf{A}_{\setminus p} C_{Z\setminus p} \mathbf{A}_{\setminus p}^\top + \mathbf{A}_{\setminus p} R^\top A_p^\top + A_p R \mathbf{A}_{\setminus p}^\top + A_p C_{\mathbf{z}}(0) A_p^\top + \Sigma \\ & \overset{(62)}{=} -\mathbf{A}_{\setminus p} C_{Z\setminus p} \mathbf{A}_{\setminus p}^\top + \mathbf{A}_{\setminus p} \bar{R}^\top + \bar{R} \mathbf{A}_{\setminus p}^\top + A_p C_{\mathbf{z}}(0) A_p^\top + \Sigma \end{aligned} \quad (64)$$

from which we conclude that

$$\begin{aligned} \Sigma + A_p Q_{pp}^{-1} A_p^\top & \overset{(59)}{=} \Sigma + A_p \left[C_{\mathbf{z}}(0) - R C_{Z\setminus p}^{-1} R^\top\right] A_p^\top \\ & = \Sigma + A_p C_{\mathbf{z}}(0) A_p^\top - A_p R C_{Z\setminus p}^{-1} R^\top A_p^\top \\ & \overset{(62)}{=} \Sigma + A_p C_{\mathbf{z}}(0) A_p^\top - \left[\bar{R} - \mathbf{A}_{\setminus p} C_{Z\setminus p}\right] C_{Z\setminus p}^{-1} \left[\bar{R} - \mathbf{A}_{\setminus p} C_{Z\setminus p}\right]^\top \\ & = \Sigma + A_p C_{\mathbf{z}}(0) A_p^\top - \bar{R} C_{Z\setminus p}^{-1} \bar{R}^\top + \mathbf{A}_{\setminus p} \bar{R}^\top + \bar{R} \mathbf{A}_{\setminus p}^\top - \mathbf{A}_{\setminus p} C_{Z\setminus p} \mathbf{A}_{\setminus p}^\top \\ & \overset{(63)}{=} C_{\mathbf{z}}(0) - \bar{R} C_{Z\setminus p}^{-1} \bar{R}^\top . \end{aligned}$$

Step 2. Derive $\det\Sigma = \det\tilde{\Sigma}$ from Andel and (**??**).

From Andel (58) we know that $\tilde{\Sigma} = (Q_{pp} + A_p^\top \Sigma^{-1} A_p)^{-1}$. As Q is positive definite, Q_{pp} is invertible such that

$$\frac{1}{\det(\tilde{\Sigma})} = \det(Q_{pp} + A_p^\top \Sigma^{-1} A_p) \overset{(60)}{=} \det(\Sigma + A_p Q_{pp}^{-1} A_p^\top) \frac{\det(Q_{pp})}{\det(\Sigma)} .$$

It therefore suffices to show that

$$\det(\Sigma + A_p Q_{pp}^{-1} A_p^\top) = \det(Q_{pp}^{-1}) . \quad (65)$$

Drawing on Step 1, Equation (64) can be proven as follows:

$$\begin{aligned} & \det(\Sigma + A_p Q_{pp}^{-1} A_p^\top) = \det(Q_{pp}^{-1}) \\ \overset{(??),(59)}{\Leftrightarrow} \quad & \det(C_{\mathbf{z}}(0) - \bar{R} C_{Z\setminus p}^{-1} \bar{R}^\top) = \det(C_{\mathbf{z}}(0) - R C_{Z\setminus p}^{-1} R^\top) \\ \overset{(61)}{\Leftrightarrow} \quad & \det\left(\begin{bmatrix} C_{Z\setminus p} & \bar{R}^\top \\ \bar{R} & C_{\mathbf{z}}(0) \end{bmatrix}\right) \det(C_{Z\setminus p}^{-1}) = \det(C_Z(0)) \det(C_{Z\setminus p}^{-1}) \\ \Leftrightarrow \quad & \det\left(\begin{bmatrix} C_{Z\setminus p} & \bar{R}^\top \\ \bar{R} & C_{\mathbf{z}}(0) \end{bmatrix}\right) = \det(C_Z(0)) . \end{aligned}$$

Switching rows or columns of a matrix leaves its determinant invariant up to a factor of $(-1)^{i+j}$, where i and j are corresponding row or column indices. In the following, we perform block-wise rotation of a matrix block to the bottom, and to right, respectively. This is a concatenation of several row and column switches giving

us a factor $(-1)^r$ for a given r. Note that r is the same for both operations due to their symmetric behavior. Therefore we have

$$\begin{aligned}\det\left(\begin{bmatrix} C_{Z\setminus p} & \bar{R}^\top \\ \bar{R} & C_{\mathbf{z}}(0)\end{bmatrix}\right) &= (-1)^r \det\left(\begin{bmatrix} \bar{R} & C_{\mathbf{z}}(0) \\ C_{Z\setminus p} & \bar{R}^\top\end{bmatrix}\right) \\ &= (-1)^{2r} \det\left(\begin{bmatrix} C_{\mathbf{z}}(0) & \tilde{R} \\ \tilde{R}^\top & C_{Z\setminus p}\end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} C_{\mathbf{z}}(0) & \bar{R} \\ \bar{R}^\top & C_{Z\setminus p}\end{bmatrix}\right) = \det(C_Z(0)) \ ,\end{aligned}$$

which completes the proof.

REFERENCES

[1] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer, 2007.

[2] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, D. Klan, R. Bauer, J. Timmer, and H. Witte, "Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems," *Signal processing*, vol. 85, no. 11, pp. 2137–2160, 2005.

[3] W. Mader, D. Feess, R. Lange, D. Saur, V. Glauche, C. Weiller, J. Timmer, and B. Schelter, "On the detection of direct directed information flow in fMRI," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 965–974, 2008.

[4] S. L. Bressler and A. K. Seth, "Wiener-Granger Causality: A well established methodology," *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.

[5] A. Bolstad, B. D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[6] R. K. Kaufmann and D. I. Stern, "Evidence for human influence on climate from hemispheric temperature relations," *Nature*, vol. 388, no. 6637, pp. 39–44, 1997.

[7] U. Triacca, "On the use of Granger causality to investigate the human influence on climate," *Theoretical and Applied Climatology*, vol. 69, no. 3-4, pp. 137–138, 2001.

[8] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[9] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, 1994.

[10] H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on Granger causality," *Phys. Rev. E*, vol. 75, p. 031123, 2007.

[11] G. Nolte, A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller, "Robustly estimating the flow direction of information in complex physical systems," *Phys. Rev. Lett.*, vol. 100, p. 234101, 2008.

[12] G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett, "Identifying true brain interaction from EEG data using the imaginary part of coherency," *Clinical neurophysiology*, vol. 115, no. 10, pp. 2292–2307, 2004.

[13] G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J. L. Cantero, "Measuring directional coupling between EEG sources," *NeuroImage*, vol. 43, no. 3, pp. 497 – 508, 2008.

[14] J.-M. Schoffelen and J. Gross, "Source connectivity analysis with MEG and EEG," *Human Brain Mapping*, vol. 30, no. 6, pp. 1857–1865, 2009.

[15] S. Haufe, R. Tomioka, G. Nolte, K.-R. Müller, and M. Kawanabe, "Modeling sparse connectivity between underlying brain sources for EEG/MEG," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 8, pp. 1954–1963, 2010.

[16] S. Haufe, V. Nikulin, K.-R. Müller, and G. Nolte, "A critical assessment of connectivity measures for EEG data: A simulation study," *Neuroimage*, vol. 64, pp. 120–133, 2013.

[17] S. Haufe, V. V. Nikulin, and G. Nolte, "Alleviating the influence of weak data asymmetries on granger-causal analyses," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 25–33.

[18] M. Vinck, L. Huurdeman, C. A. Bosman, P. Fries, F. P. Battaglia, C. M. Pennartz, and P. H. Tiesinga, "How to detect the Granger-causal flow direction in the presence of additive noise?" *NeuroImage*, vol. 108, pp. 301–318, 2015.

[19] M. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological cybernetics*, vol. 65, no. 3, pp. 203–210, 1991.

[20] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biological cybernetics*, vol. 84, no. 6, pp. 463–474, 2001.

[21] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality," *Phys. Rev. Lett.*, vol. 100, p. 144103, 2008.

[22] M. Grosse-Wentrup, "Understanding brain connectivity patterns during motor imagery for brain-computer interfacing," in *Advances in neural information processing systems*, 2009, pp. 561–568.

[23] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy - a model-free measure of effective connectivity for the neurosciences," *Journal of Computational Neuroscience*, vol. 30, pp. 45–67, 2011.

[24] M. Breakspear, M. J. Brammer, E. T. Bullmore, P. Das, and L. M. Williams, "Spatiotemporal wavelet resampling for functional neuroimaging data," *Hum Brain Mapp*, vol. 23, no. 1, pp. 1–25, Sep 2004.

[25] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 304–313, 1982.

[26] G. Nolte, A. Ziehe, N. Krämer, F. Popescu, and K.-R. Müller, "Comparison of Granger causality and Phase Slope Index." in *NIPS Causality: Objectives and Assessment*, 2010, pp. 267–276.

[27] G. Nolte, F. C. Meinecke, A. Ziehe, and K.-R. Müller, "Identifying interactions in mixed and noisy complex systems," *Phys. Rev. E*, vol. 73, p. 051913, 2006.

[28] H. Wold, *A study in the analysis of stationary time series*. Almqvist & Wiksell, 1938.

[29] J. Andel, "Symmetric and reversed multiple stationary autoregressive series," *The Annals of Mathematical Statistics*, vol. 43, no. 4, pp. 1197–1203, 1972.

[30] M. S. Bartlett, *An Introduction to Stochastic Processes*. Cambridge University Press, 1955.

[31] A. G. Akritas, E. K. Akritas, and G. I. Malaschonok, "Various proofs of Sylvester's (determinant) identity," *Mathematics and Computers in Simulation*, vol. 42, no. 4, pp. 585–593, 1996.

[32] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, p. 461464, 1978.

[33] P. Newbold, "Feedback induced by measurement errors," *International Economic Review*, pp. 787–791, 1978.

[34] L. Sommerlade, M. Thiel, B. Platt, A. Plano, G. Riedel, C. Grebogi, J. Timmer, and B. Schelter, "Inference of Granger causal time-dependent influences in noisy multivariate time series," *Journal of Neuroscience Methods*, vol. 203, no. 1, pp. 173 – 185, 2012.

[35] G. C. Tiao and W. S. Wei, "Effect of temporal aggregation on the dynamic relationship of two time series variables," *Biometrika*, vol. 63, no. 3, pp. 513–523, 1976.

[36] J. R. McCrorie and M. J. Chambers, "Granger causality and the sampling of economic processes," *Journal of Econometrics*, vol. 132, no. 2, pp. 311 – 336, 2006.

[37] D. Zhou, Y. Zhang, Y. Xiao, and D. Cai, "Reliability of the Granger causality inference," *New Journal of Physics*, vol. 16, no. 4, p. 043016, 2014.

[38] A. K. Seth, P. Chorley, and L. C. Barnett, "Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling," *NeuroImage*, vol. 65, pp. 540 – 555, 2013.

[39] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for fMRI," *Neuroimage*, vol. 54, no. 2, pp. 875–891, 2011.

[40] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity," *The Journal of Machine Learning Research*, vol. 11, pp. 1709–1731, 2010.

[41] A. Moneta, N. Chla, D. Entner, and P. Hoyer, "Causal search in structural vector autoregressive models," in *Causality in Time Series Challenges in Machine Learning*, 2011, vol. 5, pp. 95–118.

[42] J. Chiang, Z. J. Wang, and M. J. McKeown, "A generalized multivariate autoregressive (GmAR)-based approach for EEG source connectivity analysis," *IEEE Transactions on Signal Processing*, vol. 60, pp. 453–465, 2012.

[43] J. Peters, D. Janzing, and B. Schölkopf, "Causal inference on time series using restricted structural equation models," in *Advances in Neural Information Processing Systems*, 2013, pp. 154–162.

[44] H. Nalatore, N. Sasikumar, and G. Rangarajan, "Effect of measurement noise on Granger causality," *Physical Review E*, vol. 90, no. 6, p. 062127, 2014.

[45] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, p. 15, 2008.

5.2 Conclusion

Summary

In this manuscript, we established a theoretical guarantee for time-reversed Granger causality (TRGC). Previous work by (Haufe et al., 2013) showed that TRGC correctly rejects causal interpretations for mixtures of non-interacting sources. Furthermore, very promising results have been achieved in simulations (Haufe et al., 2013; Vinck et al., 2015). However, it was unknown whether time reversal leads to valid measures of information flow in the presence of true interaction.

We first reviewed the theoretical result of (Andel, 1972) stating that the time-reversed signal of any VAR(p) process has again a VAR(p) representation that can be expressed analytically in terms of the original process. We used these results to prove our main result stating that, in the case of unambiguous unidirectional information flow, TRGC indeed yields the correct result:

Corollary 1. *Let* $\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \mathbb{R}^2$ *be a stable bivariate VAR(p) process* (2.9),

$$\mathbf{z}_t = A_1 \mathbf{z}_{t-1} + \ldots + A_p \mathbf{z}_{t-p} + \epsilon_t \ ,$$

where $\epsilon_t \in \mathbb{R}^2$ *is a 2-dimensional white noise process with residual covariance matrix* $\Sigma = \langle \epsilon_t \epsilon_t^\top \rangle$. *Under the assumptions*

(A1) $A_1, \ldots, A_p$ *are lower triangular matrices (i.e.,* x_t *may Granger-cause* y_t*, but* y_t *does not Granger-cause* x_t*), and*

(A2) Σ *is a diagonal matrix (i.e. the residuals are uncorrelated), and*

(A3) *the covariance matrix of* $\mathbf{z}_t$ *is invertible ,*

it holds that

$$\tilde{D}^{(net)}_{x \to y} \geq 0 \ ,$$

where $\tilde{D}^{(net)}_{x \to y}$ *denotes the differences of the Granger scores obtained on original and time-reversed signals as defined in* (2.25).

While further compelling intuitive ideas for robust causality measures have been presented (Nolte et al., 2008; Vicente et al., 2011; Vinck et al., 2015), our result provides, to the best of our knowledge, the first proof of the correctness of one of such techniques for a relatively general class of time-series models.

Our theoretical results were complemented by simulations, in which we confirm that time-reversal testing yields robust causality estimates also in the case of true causal interaction superposed with mixed measurement noise or under downsampling. We showed that TRGC is able to infer correct directionality with similar statistical power as net Granger causality (defined in (2.14)), while being more robust with respect to measurement noise.

Limitations and Future Work

We would like to comment on the following limiting or open aspects in our theory which should be addressed in future work:

- *Definiteness of TRGC?* An open question, which we plan to address in the future, is whether under the assumptions (A1), (A2), (A3) difference-based TRGC is definite. That is, does equality $\tilde{D}^{(net)}_{x\to y} = 0$ only hold if the $A_1, \ldots, A_p$ are diagonal? It is relatively easy to prove that definiteness holds for $p = 1$, but the general p case is more involved.

  Definiteness does not hold if the residuals are correlated, that is, if (A2) is violated. Consider the following process where y Granger causes x:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.4 & -0.2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \epsilon_t \quad \langle \epsilon_t \epsilon_t^\top \rangle = \begin{bmatrix} 0.88 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

  The cross-covariance matrices $C(\cdot)$ of the process are completely symmetric

$$C(0) = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \; C(1) = \begin{bmatrix} 0.3 & 0 \\ 0 & 0 \end{bmatrix}, \; C(2) = \begin{bmatrix} 0.12 & 0 \\ 0 & 0 \end{bmatrix}, \; \ldots$$

  In this case, the VAR representation of the time-reversed process will be exactly the same as the original one, and the Granger causal flow will remain the same. Time-reversed testing will thus infer no causality. This is because mixed noise also exhibits symmetric covariance structures, and both scenarios cannot be distinguished.

  The example also illustrates that directed information flow as measured by Granger causality can arise when all cross-covariance matrices are symmetric. In fact, here there are no lagged cross-covariances at all, but still there is information flow from y to x. y_{t-1} is not correlated with x_t, but it is still predictive of x_t, because y_{t-1} correlates with $(x_t - x_{t-1})$.

- *Instantaneous effects.* Correlated residuals violate assumption (A2) of Corollary 1. They encode instantaneous effects that the variables of interest have

on each other. In this case, classical autoregressive modeling suffer from lack of identifiability (cf. (Lütkepohl, 2007; Hyvärinen et al., 2010b; Moneta et al., 2011)).

A VAR model can be seen as a 'reduced form' model, from which the 'structural' VAR model cannot uniquely be inferred using covariance information only. A Structural VAR model (SVAR) is a VAR model 'augmented' by the present variables,

$$\mathbf{z}_t = \Gamma_0 \mathbf{z}_t + \Gamma_1 \mathbf{z}_{t-1} + \Gamma_2 \mathbf{z}_{t-2} + \ldots + \Gamma_p \mathbf{z}_{t-p} + \bar{\epsilon}_t \tag{5.1}$$

where Γ_0 has zeros on its diagonal, and the residuals are expected to be uncorrelated. Notice that there are more unobserved parameters in the SVAR model than in the VAR model. Deriving the SVAR from the VAR model thus requires further assumptions regarding the relationships between the variables (Lütkepohl, 2007; Hyvärinen et al., 2010b; Peters et al., 2013).

Consider, for example, the following VAR(1) process:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \epsilon_t, \quad \langle \epsilon_t \epsilon_t^\top \rangle = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Pre-multiplying each side with a matrix $L = \begin{bmatrix} 1 & 0 \\ -0.5 & 1 \end{bmatrix}$ and subtracting $(I - L) \begin{bmatrix} x_t \\ y_t \end{bmatrix}$ from each side yields

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0.5 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0.4 & 0 \\ -0.2 & 0.4 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \bar{\epsilon}_t, \quad \langle \bar{\epsilon}_t \bar{\epsilon}_t^\top \rangle = \begin{bmatrix} 1 & 0 \\ 0 & 0.75 \end{bmatrix}.$$

Both processes have the same means and autocovariances, and thus characterize the same joint distribution of x and y if the residuals are Gaussian. However, the two representations have quite different causal interpretations.

Our theory is not applicable to instantaneous effects. However, we can argue that residuals should be independent if the VAR model accurately describes the data generating process (and is not effected by typical problems such as common drivers, noise or temporal aggregation). According to Reichenbach's common cause principle (Reichenbach, 1956), a dependency between the two residual variables would imply that one is causing the other or that both are driven by a third unobserved cause. There should also be no instantaneous interaction between x_t and y_t since physical interaction takes time which the VAR process should capture if it correctly models the latency of the data generating process.

Nevertheless, correlated residuals do occur in practice, due to downsampling or measurement noise. It might be interesting to try to incorporate instantaneous effects in the modeling, and to apply time-reversed testing to an estimated SVAR model.

- *Directed interaction superimposed with noise.* Our simulations showed that TRGC is relatively robust with respect to measurement noise and downsampling. Furthermore, we now have theoretical results both for the case of uncorrupted unidirectional information flow, and for the case of no causal interaction. Now it would be desirable to obtain theoretical guarantees for the performance of TRGC in the presence of true interaction *and* noise. Unfortunately, this might not be possible because false positives do in fact occur. Nevertheless, bounds on the false positive rate would be interesting.

- *Bidirectional information flow.* Time-reversed Granger causality indicates the *net* information flow between two variables. This is because the information flow from $\tilde{x}_t$ to $\tilde{y}_t$ in the reversed signals depends on the flow from y_t to x_t in the original signals.

  However, our theory only considers unidirectional flow of information. A result analogous to Corollary 1 for underlying bidirectional flows would be desirable. However, in this case already the definition of a 'true' dominant (net) flow direction is non-trivial, as whether x_t causes y_t more than y_t causes x_t cannot be intuitively defined from the autoregressive coefficients. Vinck et al. use the Net-GC score $F_{x\to y}^{(net)}$ of the 'raw' simulated system before adding any noise to define the ground truth about net information flow (Vinck et al., 2015). This definition has theoretical appeal, as GC is equivalent to the transfer entropy, an information-theoretic measure of time-directed information transfer, for Gaussian variables (Barnett et al., 2009). However, this definition naturally introduces a bias towards Net-GC, as any measure deviating from it will be penalized.

  Conjecture for bidirectional information flow. Let us specify very restrictive conditions, under which net information flow in bi-directional VAR systems may be defined 'unambiguously' from the AR coefficients. We define such 'unambiguous' bidirectional flow from x_t to y_t through fulfillment of the following conditions:

  (B1) For the autoregressive coefficients $A_1, \dots A_p$ it holds that:

  (B1.1) All autoregressive coefficients are non negative.

  (B1.2) Diagonal autoregressive coefficients are the same. That is, for all $h \in \{1, \dots, p\}$ it holds that $A_h(1,1) = A_h(2,2)$. Here $A_h(i,j)$ denotes the (i,j)th entry of A_p $(i,j \in \{1,2\})$.

(B1.3) Upper off-diagonal elements are smaller than lower off-diagonal elements, i.e. for all $i \in \{1, \ldots, p\}$ it holds that $A_i(1,2) \leq A_i(2,1)$.

(B2) $\Sigma = I$, i.e. the residuals are uncorrelated and have variance 1.

(B3) The covariance matrix of $\mathbf{z}_t$ is invertible.

We conjecture that Corollary 1 also holds under these conditions. Preliminary simulations indicated that this is the case.

- *Extension to the multivariate case.* We showed through simulations that the presence of a hidden common confounder cannot be ruled out by using time-reversed Granger causality. In many cases, information about the confounder may be available, and we would like to include it in the modeling. For example, in the paper in Chapter 4 we presented a preliminary simulation, in which three oscillatory neuronal sources A, B, C were causally connected to a target variable z in two different schemes, [A→z, B→z, C→z] and [C→B→A→z]. The question arises whether it is possible to distinguish both scenarios in case we are able to extract all three sources. While it is straightforward to apply post-hoc multivariate AR modeling, A, B, and C will be estimated with noise. The remaining volume artifact may therefore still generate spurious Granger causality (Schoffelen and Gross, 2009; Haufe et al., 2013).

  A challenging future research direction is therefore the development of robust measures which are able to analyze directed information flow in more than two variables. So far TRGC is only able to infer the causality structure two variables. Other proposed noise-robust measures such as PSI (Nolte et al., 2008) also focus on the two variables case. The extension of the theoretical analysis of time-reversal to general multivariate signals would therefore be very interesting.

6 Conclusion

EEG signals as measured from electrodes placed on the scalp contain useful, but noisy and spatially smeared information about the brain's electrical activity. Therefore, advanced signal processing techniques are needed to extract useful features and to recover the underlying neuronal signals of interest. In this cumulative thesis, we have contributed several methodological advances for EEG signal analysis.

Summary. First, we addressed necessary pre-processing steps. We developed the open-source EEGLAB toolbox MARA (Multiple Artifact Rejection Algorithm), which automatically identifies artifactual ICA components. MARA is a supervised machine learning algorithm that learns from 1290 labeled ICA components by extracting six features which were optimized to solve the binary classification problem 'reject vs. accept'. Thus, MARA is not limited to a specific type of artifact, and should be able to handle eye artifacts, muscular artifacts and loose electrodes equally well. It has been thoroughly evaluated on several data sets, its graphical user interface is relatively easy to use, and it is used in our group (Höhne and Tangermann, 2014; Hwang et al., 2015) and by others (Gomez Rojas, 2012; Nagya et al., 2014; Ho et al., 2015; Alday, 2015; Tóth, 2015; Wang et al., 2015). We have also used MARA to evaluate ICA-based artifact reduction in BCI systems, and to evaluate pre-processing options that are necessary to obtain a good ICA decomposition.

Second, we have analyzed how to extract oscillatory neuronal sources which Granger causally link to experimental variables of interest. The computation of Granger causality in sensor space suffers from a poor signal-to-noise ratio, however multivariate spatial filtering approaches such as ICA alleviate this issue. We presented a novel method called GrangerCPA which optimizes for Granger causality. Its ability to reliably extract oscillations that Granger cause a given external target variable was demonstrated both on simulated and real data.

Third, we contributed mathematical theory on possible solutions to problems that measurement noise poses for causal inference. More specifically, we proved that time-reversed Granger causality (TRGC) scores indicate the correct directionality in finite order autoregressive processes with unidirectional information flow. While further compelling intuitive ideas for robust causality measures have been presented, our result provides, to the best of our knowledge, the first proof of the correctness of one such technique for a relatively general class of time-series models. Furthermore,

simulations confirmed that TRGC is able to infer correct directionality with high statistical power, while being relatively robust with respect to measurement noise. We hope that these insights provide a justification and/or incentive to use TRGC to study directed interactions if the available data is corrupted by noise. This is the case when studying brain connectivity using EEG.

Outlook. Inference of cause-effect relationships using only observational data is a challenging task, for which the scientific methodology is subject to intense research. In recent years many interesting causal inference techniques have been proposed, which solve the causal inference problem under a variety of assumptions. The question is then to what degree these assumptions hold on real-world data. For TRGC, we have promising results on simulated data only. It would be very interesting to evaluate its usefulness on real data in which the ground truth is known. For example, many hobby cyclists post their GPS trails online. From GPS, noisy measurements of slope and speed can be inferred, and we know that the slope (as well as other factors) causally influences the speed, but the cyclist's speed does not change the slope. It would also be very interesting to compare and evaluate many causal inference algorithms on data from real-world problems. Steps in this direction have been carried out for instance in (Guyon et al., 2011; Peters et al., 2013).

More data is also needed in order to improve artifactual independent component classification. While MARA is a useful tool in many cases, we have encountered data sets in which its performance could be better. MARA, as well as any supervised artifactual component classifier, could be made more reliable if more training components were available. Because artifact removal with ICA is a very widespread tool, many EEG research groups do in fact have manually labeled ICA components. However, to the best of our knowledge, almost none of this data is currently publicly available. This is a shame, since, in our opinion, the lack of data is the limiting factor of artifact classification performance.

Most neuroscientific data sets are not openly available, which may slow down advances in many subfields of neuroscience. Many researchers therefore advocate open data neuroscience solutions, for example (Milham, 2012):

> Countless data sets comprising both phenotypic and neuroimaging data remain stored in laboratory archives long after publication and are often lost to the scientific community forever. Such a loss commonly reflects a lack of appreciation of the potential value of one's data to others beyond the primary study focus. Additionally, such a loss can arise from concerns about losing a competitive advantage. Regardless of motive, the end result is a missed opportunity to advance our understanding of brain-behavior relationships and the methodologies required to successfully characterize them.

For the artifact classification problem, creating or collecting a useful large data set might not be too difficult (although it requires a lot of work). This is because most class-discriminative information is contained in the scalp pattern and the spectrum of a component. These require a lot less disk space and contain less sensitive information than the component's time courses. In fact, even though features from the time series contain some non-redundant information (Winkler et al., 2011; Frølich et al., 2015a), many existing methods achieve good classification even thought they ignore the component's time series (Halder et al., 2007; Viola et al., 2009; Bigdely-Shamlo et al., 2013). A worthwhile research project would therefore be to collect or create a large data set of labeled independent components with scalp maps and spectrum and make these publicly available.

Bibliography

Acqualagna, L. and Blankertz, B. (2013). Gaze-independent BCI-spelling using rapid serial visual presentation (RSVP). *Clinical Neurophysiology*, 124(5):901–908.

Alday, P. M. (2015). *Quantity and Quality: Not a Zero-Sum Game*. PhD thesis, Philipps-Universität Marburg.

Amari, S.-i., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In *Advances in neural information processing systems (NIPS)*, pages 757–763.

Andel, J. (1972). Symmetric and reversed multiple stationary autoregressive series. *The Annals of Mathematical Statistics*, 43(4):1197–1203.

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Aust. N. Z. J. Stat.*, 43(1):75–88.

Antons, J., Schleicher, R., Arndt, S., Möller, S., Porbadnigk, A. K., and Curio, G. (2012). Analyzing speech quality perception using electro-encephalography. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):721–731.

Aue, W. R., Arruda, J. E., Kass, S. J., and Stanny, C. J. (2009). Cyclic variations in sustained human performance. *Brain and Cognition*, 71(3):336–344.

Baccalá, L. A. and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474.

Baillet, S., Mosher, J. C., and Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, 18(6):14–30.

Barnett, L., Barrett, A. B., and Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.*, 103:238701.

Barrett, A. B., Barnett, L., and Seth, A. K. (2010). Multivariate Granger causality and generalized variance. *Physical Review E*, 81.

Bartz, D. and Müller, K.-R. (2014). Covariance shrinkage for autocorrelated data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1592–1600.

Bear, M. F., Paradiso, M. A., and Connors, B. W. (2015). *Neuroscience: Exploring the Brain*. Lippincott Williams & Wilkins, 4th edition.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444.

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *European Archives of Psychiatry and Clinical Neuroscience*, 87(1):527–570.

Berka, C., Levendowski, D., Lumicao, M., Yau, A., Davis, G., Zivkovic, V., Olmstead,

R., Tremoulet, P., and Craven, P. (2008). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, 78:B231–B244.

Bigdely-Shamlo, N., Kreutz-Delgado, K., Kothe, C., and Makeig, S. (2013). Eyecatch: Data-mining over half a mizllion EEG independent components to construct a fully-automated eye-component detector. In *IEEE Conference of the Engineering in Medicine and Biology Society (EMBC)*, pages 5845–5848.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Blankertz, B., Curio, G., and Müller, K.-R. (2002). Classifying single trial EEG: Towards brain computer interfacing. In *Advances in neural information processing systems (NIPS)*, pages 157–164.

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

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

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

Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L., Sturm, I., Curio, G., and Müller, K.-R. (2010b). The Berlin Brain–Computer Interface: non-medical uses of BCI technology. *Frontiers in neuroscience*, 4.

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

Blythe, D. A., Haufe, S., Müller, K.-R., and Nikulin, V. V. (2014). The effect of linear mixing in the EEG on hurst exponent estimation. *NeuroImage*, 99:377–387.

Brandl, S., Höhne, J., Müller, K.-R., and Samek, W. (2015). Bringing BCI into everyday life: Motor imagery in a pseudo realistic environment. In *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*, pages 224–227.

Bressler, S. L. and Seth, A. K. (2011). Wiener-Granger Causality: A well established methodology. *NeuroImage*, 58(2):323–329.

Buzsáki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, 304(5679):1926–1929.

Cardoso, J.-F. and Souloumiac, A. (1993). Blind beamforming for non-gaussian signals. *IEE Proceedings F (Radar and Signal Processing)*, 140(6):362–370.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Duxbury Pacific Grove, CA.

Chaumon, M., Bishop, D. V., and Busch, N. A. (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of Neuroscience Methods*, 250:47–63.

Chen, Z. and Chan, L. (2013). Causality in linear nongaussian acyclic models in the presence of latent gaussian confounders. *Neural computation*, 25(6):1605–1641.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Crespo-Garcia, M., Atienza, M., and L.Cantero, J. (2008). Muscle artifact removal from human sleep EEG by using independent component analysis. *Ann. Biomed. Eng.*, 36:467–475.

Croft, R. J. and Barry, R. J. (2000). Removal of ocular artifact from the EEG: a review. *Clinical Neurophysiology*, 30:5–19.

Dähne, S., Meinecke, F. C., Haufe, S., Höhne, J., Tangermann, M., Müller, K.-R., and Nikulin, V. V. (2014a). SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *NeuroImage*, 86:111–122.

Dähne, S., Nikulin, V. V., Ramírez, D., Schreier, P. J., Müller, K.-R., and Haufe, S. (2014b). Finding brain oscillations with power dependencies in neuroimaging data. *NeuroImage*, 96:334–348.

del R. Millán, J., Rupp, R., Mueller-Putz, G., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kübler, A., Leeb, R., Neuper, C., Müller, K.-R., and Mattia, D. (2010). Combining brain-computer interfaces and assistive technologies: State-of-the-art and challenges. *Frontiers in Neuroprosthetics*, 4.

Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21.

Delorme, A., Palmer, J., Onton, J., Oostenveld, R., and Makeig, S. (2012). Independent EEG sources are dipolar. *PloS One*, 7(2):e30135.

Dmochowski, J. P., Sajda, P., Dias, J., and Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention–a possible marker of engagement? *Frontiers in human neuroscience*, 6:112.

Dornhege, G., del R. Milán, J., Hinterberg, T., McFarland, D., and Müller, K.-R., editors (2007). *Toward brain-computer interfacing*. MIT press.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, pages 251–276.

Entner, D. and Hoyer, P. O. (2011). Discovering unconfounded causal relationships using linear non-gaussian models. In *New frontiers in artificial intelligence*, pages 181–195.

Ewald, A., Marzetti, L., Zappasodi, F., Meinecke, F. C., and Nolte, G. (2012). Estimating true brain connectivity from EEG/MEG data invariant to linear and static transformations in sensor space. *NeuroImage*, 60(1):476–488.

Fabiani, M., Gratton, G., and Coles, M. (2000). Event-related brain potentials: Methods, theory. *Handbook of psychophysiology*, pages 53–84.

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

Fatourechi, M., Bashashati, A., Ward, R. K., and Birch, G. E. (2007). EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, 118:480–494.

Fazli, S., Dähne, S., Samek, W., Bießmann, F., and Müller, K.-R. (2015). Learning from more than one data source: Data fusion techniques for sensorimotor rhythm-based brain–computer interfaces. *Proceedings of the IEEE*, 103(6):891–906.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Fitzgibbon, S. P., Powers, D. M. W., Pope, K. J., and Clark, C. R. (2007). Removal of EEG noise and artifact using blind source separation. *Clinical Neurophysiology*, 24(3):232–243.

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302.

Frølich, L., Andersen, T. S., and Mørup, M. (2015a). Classification of independent components of EEG into multiple artifact classes. *Psychophysiology*, 52(1):32–45.

Frølich, L., Winkler, I., Müller, K.-R., and Samek, W. (2015b). Investigating effects of different artefact types on motor imagery BCI. In *IEEE Engineering in Medicine and Biology Society (EMBC)*. In press.

Geiger, P., Zhang, K., Schölkopf, B., Gong, M., and Janzing, D. (2015). Causal inference by identification of vector autoregressive processes with hidden components. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1917–1925.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313.

Ghaderi, F., Kim, S. K., and Kirchner, E. A. (2014). Effects of eye artifact removal methods on single trial P300 detection, a comparative study. *Journal of Neuroscience Methods*, 221:41–47.

Gómez-Herrero, G., Atienza, M., Egiazarian, K., and Cantero, J. L. (2008). Measuring directional coupling between EEG sources. *NeuroImage*, 43(3):497–508.

Gomez Rojas, D. M. (2012). *Language and its multi-level organization. How the brain puts order in the speech signal*. PhD thesis, Scuola Internazionale Superiore di Studi Avanzati (SISSA).

Goncharova, I. I., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2003). EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, 114:1580–1593.

Görnitz, N., Porbadnigk, A. K., Binder, A., Sannelli, C., Braun, M., Müller, K.-R., and Kloft, M. (2014). Learning and evaluation in presence of non-iid label noise. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 293–302.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Groppe, D. M., Makeig, S., and Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage*, 45(4):1199–1211.

Grosse-Wentrup, M., Schölkopf, B., and Hill, J. (2011). Causal influence of gamma oscil-

lations on the sensorimotor rhythm. *NeuroImage*, 56(2):837–842.

Guyon, I., Statnikov, A. R., and Aliferis, C. F. (2011). Time series analysis with the causality workbench. In *NIPS Mini-Symposium on Causality in Time Series*, pages 115–139.

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

Halder, S., Bensch, M., Mellinger, J., Bogdan, M., Kübler, A., Birbaumer, N., and Rosenstiel, W. (2007). Online artifact removal for brain-computer interfaces using support vector machines and blind source separation. *Computational Intelligence and Neuroscience*, 7(3):1–10.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Haufe, S., Dähne, S., and Nikulin, V. V. (2014a). Dimensionality reduction for the analysis of brain oscillations. *NeuroImage*, 101:583–597.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014b). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110.

Haufe, S., Nikulin, V., Müller, K.-R., and Nolte, G. (2013). A critical assessment of connectivity measures for EEG data: A simulation study. *NeuroImage*, 64:120–133.

Haufe, S., Nikulin, V. V., and Nolte, G. (2012). Alleviating the influence of weak data asymmetries on Granger-causal analyses. In *Latent Variable Analysis and Signal Separation*, pages 25–33.

Haufe, S., Nikulin, V. V., Ziehe, A., Müller, K.-R., and Nolte, G. (2008). Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *NeuroImage*, 42(2):726–738.

Haufe, S., Tomioka, R., Nolte, G., Müller, K.-R., and Kawanabe, M. (2010). Modeling sparse connectivity between underlying brain sources for EEG/MEG. *Biomedical Engineering, IEEE Transactions on*, 57(8):1954–1963.

Haufe, S., Treder, M. S., Gugler, M. F., Sagebaum, M., Curio, G., and Blankertz, B. (2011). EEG potentials predict upcoming emergency brakings during simulated driving. *Journal of Neural Engineering*, 8(5):056001.

Herrmann, C. S., Rach, S., Neuling, T., and Strúber, D. (2013). Transcranial alternating current stimulation: A review of the underlying mechanisms and modulation of cognitive processes. *Frontiers in Human Neuroscience*, 7(279).

Ho, H. T., Schröger, E., and Kotz, S. A. (2015). Selective attention modulates early human evoked potentials during emotional face–voice processing. *Journal of cognitive neuroscience*, 27(24):798–818.

Hoffmann, S. and Falkenstein, M. (2008). The correction of eye blink artefacts in the EEG: a comparison of two prominent methods. *PLoS One*, 3(8):e3004.

Höhne, J., Holz, E., Staiger-Sälzer, P., Müller, K.-R., Kübler, A., and Tangermann, M. (2014). Motor imagery for severely motor-impaired patients: evidence for brain-computer interfacing as superior control solution. *PLoS One*, 9(8):e104854.

Höhne, J., Schreuder, M., Blankertz, B., and Tangermann, M. (2011). A novel 9-class auditory ERP paradigm driving a predictive text entry system. *Frontiers in neuroscience*, 5:99.

Höhne, J. and Tangermann, M. (2014). Towards user-friendly spelling with an auditory brain-computer interface: the charstreamer paradigm. *PLoS One*, 9(6):e98322.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008). Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378.

Hwang, H.-J., Ferreria, V. Y., Ulrich, D., Kilic, T., Chatziliadis, X., Blankertz, B., and Treder, M. (2015). A gaze independent brain-computer interface based on visual stimulation through closed eyelids. *Scientific Reports*. In press.

Hwang, H.-J., Hahne, J. M., and Müller, K.-R. (2014). Channel selection for simultaneous and proportional myoelectric prosthesis control of multiple degrees-of-freedom. *Journal of neural engineering*, 11(5):056008.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, New York.

Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Hyvärinen, A., Ramkumar, P., Parkkonen, L., and Hari, R. (2010a). Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. *NeuroImage*, 49:257–271.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *The Journal of Machine Learning Research*, 14(1):111–152.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010b). Estimation of a structural vector autoregression model using non-gaussianity. *The Journal of Machine Learning Research*, 11:1709–1731.

Iwasaki, M., Kellinghaus, C., Alexopoulos, A. V., Burgess, R. C., Kumar, A. N., Han, Y. H., Lüders, H. O., and Leigh, R. J. (2005). Effects of eyelid closure, blinks, and eye movements on the electroencephalogram. *Clinical Neurophysiology*, 116(4):878–885.

Jensen, O. and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Frontiers in Human Neuroscience*, 4.

Joundi, R. A., Jenkinson, N., Brittain, J.-S., Aziz, T. Z., and Brown, P. (2012). Driving oscillatory activity in the human cortex enhances motor performance. *Current Biology*, 22(5):403–407.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., and Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37:163–178.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10.

Kaminski, M. and Blinowska, K. J. (1991). A new method of the description of the information flow in the brain structures. *Biological cybernetics*, 65(3):203–210.

Kauppi, J.-P., Hahne, J., Müller, K.-R., and Hyvärinen, A. (2015). Three-way analysis of spectrospatial electromyography data: Classification and interpretation. *PloS One*, 10(6):e0127231.

Kawanabe, M., Vidaurre, C., Scholler, S., Blankertz, B., and Müller, K.-R. (2009). Robust common spatial filters with a maxmin approach. In *IEEE EMBS-Conference*, pages 2470–2473.

Kierkels, J., van Boxtel, G., and Vogten, L. (2006). A model-based objective evaluation of eye movement correction in EEG recordings. *Biomedical Engineering, IEEE Transactions on*, 53(2):246–253.

Klekowicz, H., Malinowska, U., Piotrowska, A., Wołyńczyk-Gmaj, D., Niemcewicz, S., and Durka, P. J. (2009). On the robust parametric detection of EEG artifacts in polysomnographic recordings. *Neuroinformatics*, 7(2):147–160.

Klem, G. H., Lüders, H. O., Jasper, H., and Elger, C. (1999). The ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol Suppl.*, 52.

Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition-timing hypothesis. *Brain Research Reviews*, 53(1):63–88.

Kuhlen, A. K., Allefeld, C., and Haynes, J.-D. (2012). Content-specific coordination of listeners' to speakers' EEG during communication. *Frontiers in Human Neuroscience*, 6(266).

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387–399.

Lemm, S., Curio, G., Hlushchuk, Y., and Müller, K.-R. (2006). Enhancing the signal-to-noise ratio of ICA-based extracted ERPs. *Biomedical Engineering, IEEE Transactions on*, 53(4):601–607.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer.

Macdonald, J., Mathan, S., and Yeung, N. (2011). Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus eeg alpha oscillations. *Frontiers in Psychology*, 2.

Maeder, C., Sannelli, C., Haufe, S., and Blankertz, B. (2012). Pre-stimulus sensorimotor rhythms influence brain computer interface classification performance. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 20(5):653–662.

Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. *Advances in neural information processing systems (NIPS)*, 8:145–151.

Manly, B. F. (2007). *Randomization, Bootstrap and Monte Carlo methods in Biology*. CRC

Press.

Marinazzo, D., Pellicoro, M., and Stramaglia, S. (2008). Kernel method for nonlinear Granger causality. *Phys. Rev. Lett.*, 100:144103.

Massimini, M., Ferrarelli, F., Esser, S. K., Riedner, B. A., Huber, R., Murphy, M., Peterson, M. J., and Tononi, G. (2007). Triggering sleep slow waves by transcranial magnetic stimulation. *Proceedings of the National Academy of Sciences*, 104(20):8496–8501.

McCrorie, J. R. and Chambers, M. J. (2006). Granger causality and the sampling of economic processes. *Journal of Econometrics*, 132(2):311–336.

McMenamin, B. W., Shackman, A. J., Maxwell, J. S., Bachhuber, D. R. W., Koppenhaver, A. M., Greischar, L. L., and Davidson, R. J. (2010). Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG. *NeuroImage*, 49:2416–2432.

Meinecke, F., Ziehe, A., Kawanabe, M., and Müller, K.-R. (2002). Resampling approach to estimate the stability of one- or multidimensional independent components. *IEEE Transactions on Biomedical Engineering*, 49(12):1514–1425.

Milham, M. P. (2012). Open neuroscience solutions for the connectome-wide association era. *Neuron*, 73(2):214–218.

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 48(2):229–240.

Molgedey, L. and Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical review letters*, 72(23):3634.

Moneta, A., Chlağ, N., Entner, D., and Hoyer, P. (2011). Causal search in structural vector autoregressive models. In *Causality in Time Series Challenges in Machine Learning*, volume 5, pages 95–118.

Müller, K.-R., Anderson, C. W., and Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 11(2):165–169.

Müller, K.-R., Philips, P., and Ziehe, A. (1999). JADE TD: Combining higher-order statistics and temporal information for blind source separation (with noise). In *ICA'99: 1st Int. Workshop on Independent Component Analysis and Signal Separation*, pages 87–92.

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, 167(1):82–90.

Mumford, J. A. and Ramsey, J. D. (2014). Bayesian networks for fMRI: a primer. *NeuroImage*, 86:573–582.

Nagya, T., Telleza, D., Divak, A., Logob, E., Kolesb, M., and Hamornikb, B. (2014). Predicting arousal with machine learning of EEG signals. In *IEEE Conference on Cognitive Infocommunications (CogInfoCom)*, pages 137–140.

Nalatore, H., Ding, M., and Rangarajan, G. (2007). Mitigating the effects of measurement noise on Granger causality. *Phys. Rev. E*, 75:031123.

Neuper, C. and Pfurtscheller, G. (2001). Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. *International Journal of Psychophysiology*, 43:41–58.

Nikulin, V. V., Nolte, G., and Curio, G. (2011). A novel method for reliable and fast extraction of neuronal EEG/MEG osciallations on the basis of spatio-spectral decomposition. *NeuroImage*, 55:1528–1535.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., and Hallett, M. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clinical neurophysiology*, 115(10):2292–2307.

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

Nolte, G., Ziehe, A., Krämer, N., Popescu, F., and Müller, K.-R. (2010). Comparison of Granger causality and Phase Slope Index. In *NIPS Causality: Objectives and Assessment*, pages 267–276.

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

Nunez, P. and Srinivasan, R. (2006). *Electric Fields of the Brain: The neurphysics of EEG*. Oxford University Press.

Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. (2005). Recipes for the linear analysis of EEG. *NeuroImage*, 28(2):326–341.

Patel, R. S., Bowman, F. D., and Rilling, J. K. (2006). A bayesian approach to determining connectivity of the human brain. *Human brain mapping*, 27(3):267–276.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053.

Pfurtscheller, G. and Aranibar, A. (1979). Evaluation of event-related desynchronization preceding and following voluntary self-paced movement. *Electroencephalography and Clinical Neurophysiology*, 46:138–146.

Pham, D. T. and Garat, P. (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *Signal Processing, IEEE Transactions on*, 45(7):1712–1725.

Pignat, J. M., Koval, O., Ville, D. V. D., Voloshynovskiy, S., Michel, C., and Pun, T. (2013). The impact of denoising on independent component analysis of functional magnetic resonance imaging data. *Journal of Neuroscience Methods*, 213(1):105–122.

Porbadnigk, A. K., Görnitz, N., Sannelli, C., Binder, A., Braun, M., Kloft, M., and Müller, K.-R. (2015). Extracting latent brain states - towards true labels in cognitive neuroscience experiments. *NeuroImage*, 120:225–253.

Porbadnigk, A. K., Treder, M., Blankertz, B., Antons, J., Schleicher, R., Möller, S., Curio,

G., and Müller, K. (2013). Single-trial analysis of the neural correlates of speech quality perception. *Journal of Neural Engineering*, 10(5):056003.

Reichenbach, H. (1956). *The direction of time*. University of Los Angeles Press.

Romei, V., Gross, J., and Thut, G. (2010). On the role of prestimulus alpha rhythms over occipito-parietal areas in visual input regulation: Correlation or causation? *The Journal of Neuroscience*, 30(25):8692–8697.

Romero, S., Mañanas, M. A., and Barbanoj, M. J. (2008). A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: A simulation case. *Computers in Biology and Medicine*, 38:348–360.

Rosenzweig, M. R., Breedlove, S. M., and Leiman, A. L. (2002). *Biological Psychology*. Sinauer, Sunderland, 3rd edition.

Samek, W., Kawanabe, M., and Müller, K.-R. (2014). Divergence-based framework for common spatial patterns algorithms. *Biomedical Engineering, IEEE Reviews in*, 7:50–72.

Samek, W., Vidaurre, C., Müller, K.-R., and Kawanabe, M. (2012). Stationary common spatial patterns for brain–computer interfacing. *Journal of neural engineering*, 9(2):026013.

Scheeringa, R., Fries, P., Petersson, K.-M., Oostenveld, R., Grothe, I., Norris, D. G., Hagoort, P., and Bastiaansen, M. C. (2011). Neuronal dynamics underlying high- and low-frequency EEG oscillations contribute independently to the human BOLD signal. *Neuron*, 69(3):572–583.

Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., and Pfurtscheller, G. (2007). A fully automated correction method of EOG artifacts in EEG recordings. *Clinical Neurophysiology*, 118:98–104.

Schlögl, A., Ziehe, A., and Müller, K.-R. (2009). Automated ocular artifact removal: comparing regression and component-based methods. *available from Nature Precedings, http://precedings. nature. com/documents/3446/version/1/files/npre20093446-1.pdf*.

Schmidt, M. (2005). minFunc: unconstrained differentiable multivariate optimization in Matlab. www.cs.ubc.ca/ schmidtm/Software/minFunc.html. Accessed: 03/2012.

Schoffelen, J.-M. and Gross, J. (2009). Source connectivity analysis with MEG and EEG. *Human Brain Mapping*, 30(6):1857–1865.

Schreuder, M., Rost, T., and Tangermann, M. (2011). Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI. *Frontiers in Neuroprosthetics*, 5(112).

Seth, A. K., Chorley, P., and Barnett, L. C. (2013). Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage*, 65:540–555.

Shahbazi, F., Ewald, A., and Nolte, G. (2015). Self-consistent MUSIC: An approach to the localization of true brain interactions from EEG/MEG data. *NeuroImage*, 112:299–309.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). Network modelling methods for fMRI. *Neuroimage*, 54(2):875–891.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Tashiro, T., Shimizu, S., Hyvärinen, A., and Washio, T. (2014). ParceLiNGAM: a causal ordering method robust against latent confounders. *Neural computation*, 26(1):57–83.

Thut, G. and Miniussi, C. (2009). New insights into rhythmic brain activity from TMS-EEG studies. *Trends in Cognitive Sciences*, 13(4):182–189.

Thut, G., Miniussi, C., and Gross, J. (2012). The functional importance of rhythmic activity in the brain. *Current Biology*, 22(16):R658–R663.

Tiao, G. C. and Wei, W. S. (1976). Effect of temporal aggregation on the dynamic relationship of two time series variables. *Biometrika*, 63(3):513–523.

Tóth, V. (2015). Measurement of stress intensity using EEG. Master's thesis, Budapest University of Technology and Economics.

Treder, M. S. and Blankertz, B. (2010). Research (c)overt attention and visual speller design in an ERP-based brain-computer interface. *Behavioral and Brain Functions*, 6:28.

Uhlhaas, P. J., Haenschel, C., Nikolić, D., and Singer, W. (2008). The role of oscillations and synchrony in cortical networks and their putative relevance for the pathophysiology of schizophrenia. *Schizophrenia bulletin*, 34(5):927–943.

Urigüen, J. A. and Garcia-Zapirain, B. (2015). EEG artifact removal – state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001.

Venthur, B., Blankertz, B., Gugler, M., and Curio, G. (2010). Novel applications of BCI technology: Psychophysiological optimization of working conditions in industry. In *Systems Man and Cybernetics (SMC)*, pages 417–421.

Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011). Transfer entropy - a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30:45–67.

Vidovic, M. M.-C., Paredes, L. P., Hwang, H.-J., Amsuuss, S., Pahl, J., Hahne, J. M., Graimann, B., Farina, D., and Müller, K.-R. (2014). Covariate shift adaptation in EMG pattern recognition for prosthetic device control. In *IEEE Engineering in Medicine and Biology Society (EMBC) Conference*, pages 4370–4373.

Vigário, R. and Oja, E. (2008). BSS and ICA in neuroinformatics: From current practices to open challenges. *Biomedical Engineering, IEEE Reviews in*, 1:50–61.

Vigário, R., Särelä, J., Jousmiki, V., Hämäläinen, M., and Oja, E. (2000). Independent component approach to the analysis of EEG and MEG recordings. *Biomedical Engineering, IEEE Transactions on*, 47(5):589–593.

Vigário, R. N. (1997). Extraction of ocular artefacts from EEG using independent compo-

nent analysis. *Electroencephalography and clinical neurophysiology*, 103(3):395–404.

Vinck, M., Huurdeman, L., Bosman, C. A., Fries, P., Battaglia, F. P., Pennartz, C. M., and Tiesinga, P. H. (2015). How to detect the Granger-causal flow direction in the presence of additive noise? *NeuroImage*, 108:301–318.

Viola, F. C., Thorne, J., Edmonds, B., Schneider, T., Eichele, T., and Debener, S. (2009). Semi-automatic identification of independent components representing EEG artifact. *Clinical Neurophysiology*, 120:868–877.

von Bünau, P., Meinecke, F. C., Király, F. C., and Müller, K.-R. (2009). Finding stationary subspaces in multivariate time series. *Physical Review Letters*, 103(21):214101.

Wallstrom, G. L., Kass, R. E., Miller, A., Cohn, J. F., and Fox, N. A. (2004). Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods. *Psychophysiology*, 53:105–19.

Wang, D., Mo, F., Zhang, Y., Yang, C., Liu, J., Chen, Z., and Zhao, J. (2015). Auditory evoked potentials in patients with major depressive disorder measured by emotiv system. *Bio-Medical Materials and Engineering*, 26(s1):917–923.

Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological reviews*, 90(3):1195–1268.

Wikipedia (2015). Neuron. `en.wikipedia.org/wiki/Neuron#/media/File:Neuron_Hand-tuned.svg`. Accessed July 2015.

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., and Tangermann, M. (2014). Robust artifactual independent component classification for BCI practitioners. *J. Neural. Eng.*, 11(3):035013.

Winkler, I., Debener, S., Müller, K.-R., and Tangermann, M. (2015a). On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In *IEEE Engineering in Medicine and Biology Society (EMBC)*. In press.

Winkler, I., Haufe, S., Porbadnigk, A. K., Müller, K.-R., and Dähne, S. (2015b). Identifying Granger causal relationships between neural power dynamics and variables of interest. *NeuroImage*, 111:489–504.

Winkler, I., Haufe, S., and Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, 7:30.

Winkler, I., Jäger, M., Mihajlovic, V., and Tsoneva, T. (2010). Frontal EEG asymmetry based classification of emotional valence using common spatial patterns. In *World Academy of Science, Engineering and Technology*, volume 45, pages 373–378.

Winkler, I., Panknin, D., Bartz, D., Müller, K.-R., and Haufe, S. (2015c). Validity of time reversal for testing Granger causality. *arXiv preprint arXiv:1509.07636*.

Wold, H. (1938). *A study in the analysis of stationary time series*. Almqvist & Wiksell.

Wolpaw, J. and Wolpaw, E. W. (2012). *Brain-computer interfaces: principles and practice*. Oxford University Press.

Zaehle, T., Sandmann, P., Thorne, J., Jancke, L., and Herrmann, C. (2011). Transcranial direct current stimulation of the prefrontal cortex modulates working memory performance: combined behavioural and electrophysiological evidence. *BMC Neuroscience*, 12(1):2.

Zakeri, Z., Assecondi, S., Bagshaw, A., and Arvanitis, T. (2014). Influence of signal preprocessing on ICA-based EEG decomposition. In *XIII MEDICON 2013*, pages 734–737.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655.

Zhou, D., Zhang, Y., Xiao, Y., and Cai, D. (2014). Reliability of the Granger causality inference. *New Journal of Physics*, 16(4):043016.

Ziehe, A., Laskov, P., Nolte, G., and Müller, K.-R. (2004). A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, 5:801–818.

Ziehe, A., Müller, K.-R., Nolte, G., Mackert, B.-M., and Curio, G. (2000). Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Transactions on Biomedical Engineering*, 47(1):75–87.

Ziehe, A. and Müller, K.-R. (1998). TDSEP - an efficient algorithm for blind source separation using time structure. In *ICANN 98*, pages 675–680.