Decomposition and classification of electroencephalography data

Frølich, Laura

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Decomposition and classification of electroencephalography data

Laura Frølich

Kongens Lyngby 2016
PHD-2016-408

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

PHD: ISSN 0909-3192

Summary (English)

This thesis is about linear and multi-linear analyses of electroencephalography (EEG) data and classification of estimated EEG sources. One contribution consists of an automatic classification method for independent components (ICs) of EEG data and a freely available implementation as an EEGLab plug-in, "IC Classification into Multiple Artefact Classes" (IC_MARC). Four artefact classes (blinks, heart beats, lateral eye movements, and muscle contractions), a neural class, and a mixed class (representing none or a mix of the other classes) were considered. We showed that classification is possible between subjects within studies over all classes. When generalising across studies a high classification rate of neural vs. non-neural ICs was retained but the multi-class performance dropped. In another study, we used IC_MARC to compare the ability to separate artefactual from neural sources of six linear decomposition methods. This study showed that high-pass filtering data at high cut-off frequencies improved artefact removal performances in an Event-Related Desynchronisation setting, providing similar performances of the three included Independent Component Analysis variants. IC_MARC was also used to inspect effects of artefacts on motor imagery based Brain-Computer Interfaces (BCIs) in two studies, where removing artefactual ICs had little performance impact. Finally, we investigated multi-linear classification on single trials of EEG data, proposing a rigorous optimisation approach. To enforce orthonormality of projection matrices, objective functions quantifying class discrimination were optimised on a cross-product of Stiefel (orthonormal matrix) manifolds. Supervised feature extraction outperformed unsupervised methods, but the choice of supervised method mattered less. We suggested completions of methods to include both PARAFAC and Tucker structures. The two structures provided similar performances, making the more interpretable PARAFAC models appealing.

Summary (Danish)

Denne afhandling omhandler lineære og multi-lineære analyser af elektroencefalografi (EEG) data og klassifikation af estimerede EEG kilder. Et bidrag består af en metode til automatisk klassifikation of independent components i EEG data og en frit tilgængelig implementation af denne som et EEGLab plug-in, "IC Classification into Multiple Artefact Classes" (IC_MARC). Vi betragede fire artifakt klasser (blink, hjerteslag, sidelæns øjenbevægelser, og muskelsammentrækninger), en neural klasse, og en mixet klasse (denne repræsenterer andre klasser end de nævnte og kombinationer af dem). Vi viste at klassifikation imellem individer indenfor studier var muligt over alle klasser. Ved generalisering imellem studier forblev klassifikationsraten høj ved skelnen mellem neurale og ikke-neurale komponenter, men multi-klasse præstationen faldt. I et andet studie brugte vi IC_MARC til at sammenligne evnen til at separere artifakt og neurale komponenter for seks lineære dekompositionsmetoder. Dette studie viste at højpas filtrering af data ved høje skæringsfrekvenser forbedrede evnen til fjernelse af artifakter i et Event-Related Desynchronisation paradigme og resulterede i sammenlignelige præstationer af de tre inkluderede Independent Component Analysis varianter. Yderligere brugte vi IC_MARC til at undersøge artifakters effekter på motor imagery baserede Brain-Computer Interfaces (BCIs), hvor det at fjerne artifkakt ICs kun havde lille effekt. Endelig undersøgte vi multi-lineær klassifikation på single-trial niveau i EEG data, hvor vi foreslog en stringent optimeringstilgang. For at sikre ortonormale projektionsmatricer optimerede vi objektivfunktioner der kvantificerer klassediskrimination på et krydsprodukt af Stiefel (ortonomal matrix) mangfoldigheder. Vejledt feature udtrækning afstedkom bedre klassifikationpræstationer end ikke-vejledte metoder, men det specifikke valg af vejledt metode var mindre betydningsfuldt. Vi introducerede PARAFAC og Tucker strukturer i metoderne og fandt at deres

præstationer var sammenlignelige. Dette gør de lettere fortolkelige PARAFAC modeller attraktive.

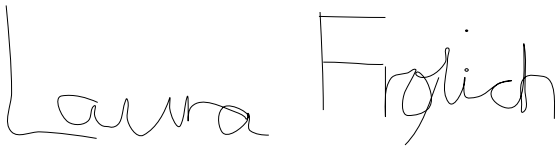
Preface

This thesis was prepared at the Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with linear and multi-linear modelling of electroencephalography (EEG) data. The thesis presents a method for automatic classification of independent components of EEG data and three studies employing this classifier in other analyses. Additionally, the intrinsic multi-linear structure of EEG data is exploited in a study of single-trial classification using multi-linear methods.

The thesis consists of a summary report, two published articles, and three submitted articles, written between 2011-2016.

Kgs. Lyngby, 03-March-2016



Laura Frølich

Papers included in the thesis

[A] Laura Frølich, Tobias S. Andersen, Morten Mørup. Classification of independent components of EEG into multiple artifact classes. Psychophysiology, 2015. Published.

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

[C] Stephanie Brandl, Laura Frølich, Johannes Höhne, Klaus-Robert Müller Wojciech Samek. Brain-Computer Interfacing under Distraction: An Evaluation Study, 2016. Submitted in a slightly different version from the pre-print included in this thesis. Specifically, Figure 8 in the submitted manuscript has a part concerning artefacts which might need to be revised, and hence was taken out in the pre-print.

[D] Laura Frølich, Irene Winkler. Removal of muscular artifacts in EEG signals: A comparison of ICA and other linear decomposition methods, 2016. Submitted.

[E] Laura Frølich, Tobias S. Andersen, Morten Mørup. Multi-linear Discriminant Analysis with Tucker and PARAFAC Structures optimised on the Stiefel Manifold, 2016. Submitted.

Acknowledgements

First, I would like to express my gratitude to my principal supervisor, Associate Professor Tobias S. Andersen to whom I am grateful for unfaltering support, encouragement, and guidance. Likewise, my co-supervisor Associate Professor Morten Mørup has been a constant source of inspiration and educative technical discussions.

I am also thankful to Klaus-Robert Müller for hosting me during my 3-month visit to his group, the Machine Learning/Intelligent Data Analysis group at Technische Universität Berlin. I am grateful to Irene Winkler and Wojciech Samek for making time for many technical discussions during this visit.

Furthermore, I am grateful to Professor Lars Kai Hansen for the colloquial yet instructive environment that he inspires and sustains at the Section for Cognitive Systems. Likewise, I would like to thank everyone at the section who collectively create a pleasant atmosphere. In particular, I would like to thank my past and present office mates Ditte Hald Høvenhoff, Søren Føns Vind Nielsen, Camilla Birgitte Falk Jensen, Kit Melissa Larsen, Carsten Stahlhut, and Dan Tito Svenstrup for making daily PhD life in the office enjoyable.

I would like to express my gratitude to my husband Morten Frølich and children, Amanda and Anton, for putting up with my absences. In particular, Morten has been helpful throughout the process by offering moral support and managing practicalities of daily life when I needed longer hours in the last several months.

Contents

Notes

The following symbols and conventions retain their meaning throughout this work (with possible exceptions in the papers attached as appendices). Lowercase letters not in bold denote scalars (a) while lowercase letters in bold denote column vectors ($\boldsymbol{a}$). Row vectors are written as column vectors transposed: $\boldsymbol{a}^\top$. Matrices are written with capital letters as $\mathbf{A}$ while curly capital letters $\mathcal{A}$ denote tensors (multi-linear matrices).

Number of data objects: N

Number of classes: C

Number of data objects in class c: N_c

Number of modes: P

Size of the p^{th} data mode (or dimension): J_p

Number of components/factors for mode p: K_p

Number of channels: M

Number of temporal samples in an entire EEG recording: T

Number of temporal samples in each epoch: T_e

Identity matrix of unspecified dimension: I

Identity matrix of dimension q: I_q

Transpose operator: $\top$

The i^{th} row of a matrix $\mathbf{X}$: $\mathbf{X}_{i,:}$

The j^{th} column of a matrix $\mathbf{X}$: $\mathbf{X}_{:,j}$

The element in the i^{th} row and j^{th} column of $\mathbf{X}$: $\mathbf{X}_{i,j}$

Regularisation matrix: $\mathbf{R}$

Matrix containing model errors: $\mathbf{E}$

Tensor containing model errors: $\mathcal{E}$

Projection matrix for mode p: $\mathbf{U}^{(p)}$

Class of an observation, $\mathcal{X}_n$: $class(\mathcal{X}_n)$

Frobenious norm: $\|\cdot\|_F$

Matricising along mode p: $(\cdot)_{(p)}$ or $[\cdot]_{(p)}$

Projection vector for binary class separation: $\boldsymbol{u}$

Projection matrix for multi-class separation of data: $\mathbf{U}$

Projection matrix for the p^{th} mode for class separation of tensor data: $\mathbf{U}^{(p)}$

EEG data matrices are represented as $channels \times time$ so that each row corresponds to an EEG electrode.

The following abbreviations are used throughout the thesis:

ICA: Independent Component Analysis

IC: Independent Component

SSD: Spatio-Spectral Decomposition

TDSEP: Temporal Decorrelation source SEParation

DATER: Discriminant Analysis with TEnsor Representation

CMDA: Constrained Multi-linear Discriminant Analysis

DGTDA: Direct General Tensor Discriminant Analysis

IC_MARC: Independent Component Multi-class ARtifact Classification

EEG: Electroencephalography

ERP: Event-Related Potential

BCI: Brain-Computer Interface

CSP: Common Spatial Patterns

Part I

Introduction

The word electroencephalography (EEG) has its roots in the Greek words: "electro", "encephalo", and "gram", which mean electrical, brain, and picture, respectively [Tyner et al., 1989]. This is a fitting description of EEG, as a picture of the electrical brain. In 1924, Hans Berger was the first to record the human EEG in order to study the functioning of the human brain [Tyner et al., 1989]. EEG data is recorded from the scalp with a number of electrodes, that can range from as few as two [Hugdahl and Westerhausen, 2010] or three [Rejer and Gorski, 2015] to as many as 512 [Hugdahl and Westerhausen, 2010, Ryynänen et al., 2005] electrodes. An EEG electrode registers the electrical potential between itself and a reference electrode. The electric potentials measurable on the scalp are mainly generated by post-synaptic potentials in dendrites of pyramidal neurons since such neurons are relatively large and aligned perpendicularly to the cortical surface [Buzsáki et al., 2012, da Silva, 2009, Kirschstein and Köhling, 2009]. When sufficiently many pyramidal neurons have similar activation patterns, i.e. nearly synchronous firing patterns, their electrical currents sum up and become measurable at the scalp [da Silva, 2009].

Artefacts in EEG

While EEG recordings measure electrical brain activity, they also include artefacts, which may be of technical or biological origin. The large amplitudes of artefacts relative to the electrical activity generated by brain processes complicate analyses of EEG data. We give some examples of artefacts in the following.

Technical artefacts are caused by equipment or other external sources. Examples of technical artefacts are line noise, loose electrodes, and gel or sweat short-circuiting electrodes [Fisch and Spehlmann, 1999]. These artefacts exhibit different characteristics. For example, line noise causes rhythmic activity at 50 or 60 Hz (power-grid dependent) while a loose electrode can cause spikes or long-duration slow waves in EEG recordings [Fisch and Spehlmann, 1999].

Biological artefacts originate in the subject's own body and include eye movements, chewing, facial- and neck muscle contractions, and the electrical heart beat [Fisch and Spehlmann, 1999]. A heart beat gives rise to an artefact with a frequency ranging from about 15-32 Hz [Jiang et al., 2007] and a spatial pattern exhibiting a smooth gradient from positive to negative across the scalp [Radüntz et al., 2015]. Blinks and eye movements generate somewhat different signatures in the EEG [Plöchl et al., 2012]. While the different ocular artefacts have different temporal signatures and frequency content, they are all strongly localised in the frontal electrodes. Muscle artefacts' spatial patterns look dipolar. That is, they have a strong positive value next to a negative value, and these are

concentrated in a small area, usually near the edge of the scalp [Radüntz et al., 2015]. Their temporal frequency is mostly above 20 Hz [Muthukumaraswamy, 2013].

Event-related potentials

Event-related potentials (ERPs) are patterns in the EEG time-locked to stimuli or responses [Luck, 2005]. As described above, EEG data also contains background noise unrelated to the stimulus, making it diffult to see ERPs in single trials and complicating analyses of EEG data. Since Hans Berger's times, though, the ready availability of computers has made analyses of EEG data more approachable. For example, a standard way to handle random noise in EEG recordings is to record many trials of a subject responding to stimuli. By rejecting the most noisy trials and averaging across the rest, much of the EEG activity that is not time-locked to the stimulus can be cancelled out [Luck, 2005]. This should leave the ERP as the strongest remaining signal. The averaging method is widespread due to its efficiency and ease of use.

Typical ERP responses in specific experimental paradigms, such as the P300 and N1 ERPs, have been discovered using the averaging method described above [Luck, 2005]. The waveforms now known to be elicited by certain stimuli can be used to investigate how the brain responds in more complex experimental paradigms [Kim and Osterhout, 2005, Ye et al., 2006, Eskelund et al., 2015, Paynter et al., 2010]. However, the amplitudes of ERPs of some subjects may be quite low [Luck, 2005, ch. 1]. This can potentially complicate analyses of experiments by making differences between averages with and without ERPs smaller, thus making it more difficult to detect whether or not an ERP was present. ERPs are also used as the basis of various Brain-Computer Interfaces (BCIs) [Guan et al., 2004, Zhu et al., 2010, Tobimatsu et al., 1999, Höhne et al., 2011, Chen et al., 2015].

Disadvantages of trial-rejection and averaging

The rejection of trials that are too noisy entails loss of data, leading to higher uncertainty in the final conclusions of analyses. Instead of rejecting entire trials, the noise can be subtracted from data. One approach to estimating non-neural activity in EEG recordings is to obtain auxilliary recordings of known artefactual sources such as eye movements and the electrocardiogram (ECG) simultaneously

with the EEG recording [Geetha and Geethalakshmi, 2012, Plöchl et al., 2012, Hoffmann and Falkenstein, 2008, Quilter et al., 1977]. Although this approach improves signal quality [Jervis et al., 1989, Croft and Barry, 2002], it has some inherent problems. Auxilliary recordings require more equipment and possibly added discomfort for the subject. Also, some auxilliary recordings, such as the electrooculogram for eye movements, might not be possible to record exclusively, without neural activity. When noise is subtracted based on an auxilliary signal, some neural activity can therefore also be lost. This is not a problem for all auxilliary recordings, though. The ECG, for example is unlikely to record neural signals. Another approach is to estimate noise signals based on the recorded EEG and subsequently subtract the estimated noise. Makeig et al. [1996] showed that Independent Component Analysis (ICA) is able to achieve a high degree of separation of artefactual and neural signals into different Independent Components (ICs). Since then, many researchers have used ICA as a pre-processing step [Jung et al., 2000, Vialatte et al., 2008], [Schomer and Da Silva, 2012, pp. 1087,1185]. In order to use ICA for data cleaning it is necessary to identify the ICs that should be projected out of the data. Manual classification of ICs is a laborious task, making automatic classification methods appealing.

Similarly, averaging across trials causes a loss of information about variations in the spatio-temporal structure. For example, a long waveform with low amplitude resulting from averaging could either be due to all responses being of this type or by high variability in the on-sets of waveforms of large amplitudes and short durations [Luck, 2005]. By representing data as tensors, which are multidimensional generalisations of matrices, the inherent data structure can be retained and exploited during analyses. This approach has been used to investigate the structure of ERPs [Verleger et al., 2013].

Furthermore, the necessity of averaging across several trials before classifying whether an ERP was present limits the information transfer rate of BCIs. Tensor methods might be good candidates for improving single-trial classification performances since they utilise a larger part of the data structure than methods that suppress the multi-linear nature of EEG data. Also, a reliable single-trial classifier's performance on data from a subject in a standard experimental setting known to produce an ERP could be used as a measure for the degree to which the subject produces an ERP. This could be used to screen subjects in neuroscience experiments in an objective manner.

Contributions

An automatic classifier of ICs is described in Chapter 4. Whereas others have worked on the problem of distinguishing neural from artefactual ICs, we extended the classification to distinguish between various types of artefacts. By distinguishing between different types of artefactual ICs, we investigated the different artefacts' effects on motor imagery based BCIs and the artefact distributions in experimental paradigms that simulated out-of-the-lab BCI. These analyses are described in Chapter 5. By distinguishing between neural and artefactual sources, we compared various decomposition methods on their ability to separate neural from artefactual activity. This work is summarised in Chapter 6.

Although others have proposed supervised tensor methods for feature extraction [Li and Schonfeld, 2014, Yan et al., 2005, Tao et al., 2007], they employed heuristic optimisation methods. We proposed to rigorously optimise different objective functions and used the conjugate gradient method as provided in the *ManOpt* [Boumal et al., 2014] toolbox for Matlab. To enforce orthonormality constraints on projection matrices, the proposed optimisation was performed on cross-products of Stiefel manifolds, on which orthonormal matrices lie. We compared the classification performances to unsupervised feature extraction and to a method that combines feature extraction and classification [Dyrholm et al., 2007]. Additionally, we completed the methods to encompass both the flexible Tucker structure and the more rigid PARAFAC structure. This allowed us to investigate whether it is sufficient to model data with the PARAFAC structure. The PARAFAC structure only allows a factor estimated for one data dimension (columns/rows for matrices) to interact with one factor from each of the other dimensions, making this model more easily interpretable. Conversely, the Tucker structure allows all factors to interact across dimensions. The investigations of tensor methods are summarised in Chapter 8.

Structure of thesis

In Part II, theory used in the remainder of the thesis is described. Next, the classifier of ICs and the work employing this classifier is described in Part III. Then our work on supervised tensor decomposition methods is summarised in Part IV. Finally, the work as a whole is discussed in Part V. The appendices contain the papers that the thesis is based on and some mathematical details.

Part II

Background theory

Chapter 1

Modeling EEG data

When modeling EEG data, linear models are often used. Let $\boldsymbol{x}_t$ be the recording at time t for an electrode. Then an often used linear model is:

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{b}_k^\top \mathbf{S}_{k,t} + \boldsymbol{e}_t,$$

where the matrix $\mathbf{S}$ holds the degree of activation of each of K sources at each time point, t, and $\boldsymbol{e}_t$ is the reconstruction error, i.e. deviance between the model and observed data. The vector $\boldsymbol{b}^\top$ models how the sources are mixed when recorded on the scalp by the electrode. This formulation of the linear model assumes that the mixing of the sources is instantaneous in addition to being linear. Due to the frequency of EEG signals, quasi-static approximations hold such that instantaneous mixing of signals from inside the brain at the scalp holds [Hyvärinen et al., 2001, p. 409]. Since volume conduction is thought to be linear, the assumption of linear mixing is also valid [Lee, 1999, p. 147]. When observations from multiple channels are available, observation vectors $\boldsymbol{x}^\top$ from each channel are collected as rows in the matrix $\mathbf{X}$ while vectors $\boldsymbol{b}$ corresponding to each channel are collected as rows in the matrix $\mathbf{A}$. The matrix $\mathbf{A}$ is referred to as the mixing matrix [Hyvärinen et al., 2001, ch. 7] since it mixes the unobserved sources, contained in the matrix $\mathbf{S}$, to form the

observed signals. Collecting the errors in the matrix $\mathbf{E}$, the matrix formulation of the model is:

$$\mathbf{X} = \mathbf{AS} + \mathbf{E}. \tag{1.1}$$

The i^{th} source is characterised by its spatial expression on the scalp (column $\mathbf{A}_{:,i}$) and its temporal activation (row $\mathbf{S}_{i,:}$). The formulation $\mathbf{X} = \mathbf{AS} + \mathbf{E}$ is referred to as the forward model while, disregarding the error term in the following,

$$\mathbf{S} = \mathbf{WX}$$

is referred to as the backward model [Haufe et al., 2014]. The columns of $\mathbf{A}$ are referred to as (activation) patterns while the rows of $\mathbf{W}$ are referred to as (extraction) filters [Haufe et al., 2014]. Each pattern shows how the corresponding source is expressed on the scalp while each filter gives the linear combination of electrodes needed to isolate the source's time series. Filters cannot be interpreted in a biophysical manner since the linear combination of electrodes required to extract a source signal is likely to include electrodes that are irrelevant to the generating source. Such electrodes might instead have recorded a noise signal that needs to be taken into account for source extraction. The following example, inspired by Haufe et al. [2014], illustrates the above statement.

Assume two electrodes, x_1 and x_2, record mixes of a source signal of interest, s_1, and a noise source, s_2, such that $x_1 = s_1 + 2 \cdot s_2$ and $x_2 = s_2$. Then the filter necessary to extract the signal of interest would be $\boldsymbol{w}_1 = [1, -2]$. If the magnitudes of the filter coefficients were interpreted as being biophysically meaningful, we would wrongly conclude that electrode x_2 detected most of the interesting signal. Conversely, we would have:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} s_1 \\ s_2 \end{bmatrix},$$

such that the pattern for s_1 would be $[1, 0]$, meaning that s_1 only projects to electrode x_1, as is the case.

1.1 Relation between filters and patterns

If $\mathbf{A}$ is square and invertible, it is easy to convert between the forward- and backward models since $\mathbf{W} = \mathbf{A}^{-1}$ in this case. If $\mathbf{A}$ and $\mathbf{W}$ are not square, the backward and forward models still uniquely determine each other, but less trivially:

$$\mathbf{A} = Cov(\mathbf{X})\mathbf{W}^\top (Cov(\mathbf{S}))^{-1} \text{ [Haufe et al., 2014].}$$

This relation is necessary to interpret extracted filters and we used it to interpret both spatial and temporal projection matrices in multi-linear (tensor) classification models [Frølich et al., 2016].

A common assumption is that sources are uncorrelated, in which case $Cov(\mathbf{S})$ is a diagonal matrix. In this case, the effect of the source covariance is simply to scale each column of $\mathbf{A}$. Since this scaling is already ambigous in the model $\mathbf{X} = \mathbf{AS}$, it does not influence the information on the sources' spatial distributions contained in the model. By taking this term out, a simpler conversion formula is obtained [Haufe et al., 2014]:

$$\mathbf{A} = Cov(\mathbf{X})\mathbf{W}^\top.$$

Formally, the covariance matrices in the above equations are the true population covariances. However, we have substituted the sample estimates since this is what would be used in practice.

The conversion from a filter to a pattern can be understood intuitively, for example by considering the simple case of a filter consisting of only one non-zero value, e.g. in the i^{th} position, with that value being 1 ($\boldsymbol{w} = (0, 0, \ldots, 1, 0, \ldots, 0)$). The effect of such a filter would be to extract all data recorded from channel i, and nothing else, as a source. The pattern corresponding to this filter is then the i^{th} column of the data covariance matrix. This result is in accordance with intuition since activation of this source would be entirely determined by activation of channel i, and the spread of activation across the scalp corresponding to recordings by this channel is exactly expressed as the covariance between channel i and all other channels.

1.1.1 Connection with least squares regression

The forward model $\mathbf{X} = \mathbf{AS}$, with the source time series matrix $\mathbf{S}$ already estimated so that these can be considered fixed, can alse be considered as a least squares problem [Parra et al., 2005, Haufe et al., 2014]. In this context, we wish to determine the parameter matrix $\mathbf{A}$ such that $\sum_{t=1}^{T}(\mathbf{X}_{:,t} - \mathbf{AS}_{:,t})^\top(\mathbf{X}_{:,t} - \mathbf{AS}_{:,t})$ is minimised. The maximum likelihood estimator of the parameter matrix $\mathbf{A}$ is (see e.g. [Bishop, 2006, section 3.1.5] for details)

$$\hat{\mathbf{A}}_{ML} = \mathbf{XS}^\top(\mathbf{SS}^\top)^{-1}.$$

Since we have $\mathbf{S} = \mathbf{WX}$, we can write this as

$$\hat{\mathbf{A}}_{ML} = \mathbf{XX}^\top\mathbf{W}^\top(\mathbf{SS}^\top)^{-1}.$$

With zero-mean data, this expression reduces to $Cov(\mathbf{X})\mathbf{W}^\top(Cov(\mathbf{S}))^{-1}$, which is what we arrived at before.

Chapter 2

Linear (matrix) methods

We now give an overview of the matrix methods that were used in this thesis. We describe the unsupervised decomposition methods Independent Component Analysis (ICA), Spatio-Spectral Decomposition (SSD), and Fourier-ICA. Finally, we describe the supervised method Common Spatial Patterns (CSP).

2.1 Unsupervised matrix decompositions

2.1.1 Independent Component Analysis

2.1.1.1 The model

We only describe and use the noise-free ICA model throughout this thesis. For descriptions of more general ICA models, see e.g. [Hyvärinen, 1998, Voss et al., 2013]. The noise-free Independent Component Analysis (ICA) model is:

$$\mathbf{X} = \mathbf{AS}.$$

In ICA, the sources are usually referred to as Independent Components (ICs).

2.1.1.2 Assumptions

Stationarity of source time series and mixing matrix All observations in a row of $\mathbf{S}$ are assumed to be realisations of random variables with the same distribution and the mixing matrix $\mathbf{A}$ is assumed to be constant [Hyvärinen, 2012, Comon, 1994]. Together, these two assumptions imply that all observations in a row of $\mathbf{X}$ should be realisations of random variables with the same distribution. If the rows of $\mathbf{X}$ and $\mathbf{S}$ contain stochastic processes, these assumptions imply that the time series must be stationary. These assumptions allow ICA algorithms to use observations from all time points to estimate a de-mixing matrix, $\mathbf{W}$, yielding maximally independent source time series ($\mathbf{S} = \mathbf{WX}$).

While EEG source time series cannot be assumed stationary over long stretches of time, the stationarity assumption is likely to be fulfilled to a higher degree for several short EEG recordings of a subject performing the same task [Neuper and Klimesch, 2006, Korats et al., 2012]. An algorithm that takes non-stationarity into account has been proposed [Palmer et al., 2008]. Although the algorithm outperforms other ICA variants on a measure of dipolarity of extracted sources, the advantage is modest [Delorme et al., 2012]. The mixing matrix $\mathbf{A}$ represents mixing of sources resulting from the physical structure of the head and brain so this can be assumed to be constant [Hyvärinen et al., 2001, p. 409].

Linear and instantaneous mixing The assumptions of linear and instantaneous mixing are apparent from the formulation of the ICA model and largely fulfilled by EEG data [Hyvärinen et al., 2001, p. 409], and [Lee, 1999, p. 147].

Statistical independence ICA is built on the assumption that sources are statistically independent [Hyvärinen et al., 2001, p. 152]. The assumption of statistical independence allows the use of all moments of the source time series in the quantificaton of their independence. This contrasts with Principal Component Analysis (PCA) whose goal is to decorrelate sources such that their second moments are zero. The stronger assumption of all cross-moments being zero provides additional information for source separation. If the rows of $\mathbf{X}$ and $\mathbf{S}$ contain time series, this assumption should be understood as instantaneous independence, i.e. that at each time instant each IC is statistically independent from the other ICs at that time instant. Some extensions to the basic ICA model make other assumptions on the time structure or also demand zero cross-correlation between ICs at different time points [Ziehe and Müller, 1998].

Since some brain processes may be statistically independent from each other at each time instant, ICA should be able to identify spatial filters that extract such sources. Some artefactual sources of activity are also likely to be independent from each other and brain processes. This is a better assumption for some artefacts than for others, though. While the assumption seems to be a good description for technical artefacts such as power grid noise and loose electrodes, it may be less appropriate for biological artefacts such as eye movements, which can affect input to the visual cortex, and eye blinks, which are known to be preceded by a signal to ignore visual input [Johns et al., 2009, Ridder and Tomlinson, 1993].

Non-Gaussian sources The assumption of at most one Gaussian source is necessary for identifiability of all sources [Hyvärinen et al., 2001, p. 153],[Gutch and Theis, 2007, Comon, 1994]. If more than one source is Gaussian, such sources cannot be separated from each other since their third- and higher order moments are non-existent. Hence the assumption of independence only implies lack of correlation between Gaussian sources, not independence, and any rotation of the estimated de-mixing matrix will also result in zero correlation between the Gaussian sources. This means that the part of a de-mixing matrix that is related to Gaussian sources can only be determined up to a rotational ambiguity. All the non-Gaussian sources can, however, still be separated from each other and the Gaussian sources [Hyvärinen et al., 2001, section 7.5].

Time series of biological artefacts are typically non-Gaussian. Eye blink time series, for example, are mostly zero, with occasional non-zero periods with characteristic shapes when a blink occurs. The same can be said of muscle artefacts and lateral eye movements, i.e. that their time series are mostly zero, except when the artefact occurs. Technical artefacts also have non-Gaussian time series in most cases. Contamination from the alternating current is systematic with the frequency of the alternating current, and hence non-Gaussian. A loose electrode is typically not loose during the whole EEG recording but when it is, its time series can exhibit various, typically non-Gaussian, characteristics (see Part I). Additionally, non-Gaussian neural sources seem to be common [Gómez-Herrero et al., 2008, Makeig et al., 2002]. Since both artefactual and neural non-Gaussian sources exist, it is reasonable to apply ICA to EEG data. After performing ICA, the extracted ICs' time series can be subjected to tests of normality. If such a test indicates an estimated source is Gaussian, it should be kept in mind that it may be a mix of several (Gaussian) sources.

Left invertible mixing matrix The $J_1 \times K$ mixing matrix $\mathbf{A}$ for K ICs must be invertible, or at least left invertible [Gutch and Theis, 2007, Comon, 1994,

Vigário et al., 1998],[Faugeras et al., 2012, p. 100]. This is fulfilled if $J_1 \geq K$ and $\mathbf{A}$ has rank at least K. Since the rank of $\mathbf{A}$ cannot be larger than the rank of the data matrix, at most as many sources as the rank of the data matrix can be estimated.

While the number of neural and artefactual sources contributing to the recorded EEG is generally larger than the number of EEG electrodes, we can ensure that the estimated mixing matrix is left-invertible by estimating at most as many ICs as the rank of the data matrix, $\mathbf{X}$.

2.1.1.3 Ambiguities

The ICA model has inherent sign and scale ambiguities since the model, $\mathbf{X}_{i,:} = \mathbf{A}_{i,:}\mathbf{S}$, is indistinguishable from $\mathbf{X}_{i,:} = (-\mathbf{A}_{i,:}) \times (-\mathbf{S})$ and $\mathbf{X}_{i,:} = \left(\frac{1}{\alpha}\mathbf{A}_{i,:}\right) \times (\alpha S)$ for some real number α. One way of resolving the scale ambiguity is to define rows of $\mathbf{S}$ to have unit variance. The sign ambiguity, however, persists.

2.1.1.4 Non-orthogonality

ICs are not necessarily orthogonal implying that removing one IC can affect dimensions of data that are also influenced by other ICs. Hence removing (or adding) an IC can either increase or decrease data variance in the space spanned by the ICs. Figure 2.1 shows (orthogonal) PCs and (non-orthogonal) ICs of randomly generated data to illustrate the difference between ICs and PCs.

2.1.1.5 Number of ICs

The desirable number of ICs to estimate is the true number of independent, generating sources. However, that number is, in general, unknown.

To determine the number of components in PCA, one heuristic is to establish a threshold of data variance that the retained Principal Components (PCs) must explain. After sorting the PCs in order of decreasing explained data variance, the first PCs that together explain as much data variance as the threshold are kept. The basis for this method is that PCs are uncorrelated since PCA forces PCs to be orthogonal, making their variances additive. This method determines not just the number, but the exact PCs to retain. Since ICs are not necessarily orthogonal, this simple PCA heuristic cannot be used directly for ICA. However, it can be used during the whitening pre-processing step, which

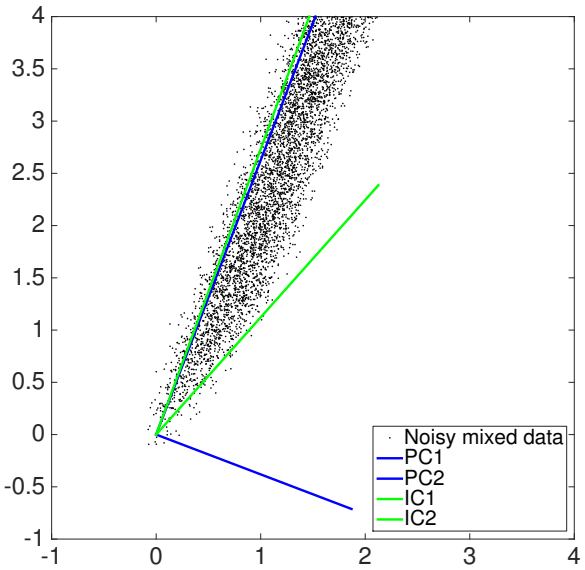


Figure 2.1: Orthogonal principal components and non-orthogonal independent components of randomly generated data. The top of the figure was cut to show axes on the same scale without sacrificing detail where the components meet.

many ICA algorithms employ. In this step, data are subjected to PCA to obtain a zero-mean uncorrelated data representation. In this step, a number of PCs or a threshold on the explained data variance to retain can be enforced to obtain a lower-dimensional representation of data. Then the number of PCs retained in this step is the number of ICs estimated [Hyvärinen et al., 2001, p. 269]. Other methods to determine the optimal number and which ICs to retain have also been proposed [Hyvärinen and Ramkumar, 2013, Cheng et al., 2012], but no gold standard is commonly agreed upon.

Since there are many ($>> J_1$) independent sources in EEG data, often as many ICs as possible are estimated. Hence the number of electrodes, J_1, is the limiting factor on the number of estimable ICs. However, usually only the first 20-30 components can be interpreted as neural or biophysical processes or technical artefacts. The remaining ICs represent mixes of different processes or unexplained data variance.

2.1.1.6 Algorithms

Some of the commonly used ICA algorithms are the Extended Infomax procedure [Lee et al., 1999], FastICA [Hyvärinen, 1999], and TDSEP [Ziehe and Müller, 1998]. These all have the aim of finding maximally statistically independent components of data, but approach the problem in different ways. The idea behind the Extended Infomax algorithm is to maximise the mutual information between the inputs and outputs of a neural network [Lee et al., 1999]. The learning rule is derived from maximum-likelihood considerations which constitute an equivalent formulation of the problem [Lee et al., 1999]. FastICA, on the other hand, focuses on maximising non-Gaussianity of each estimated source by maximising an estimate of its negentropy [Hyvärinen, 1999]. This idea is founded in the Central Limit Theorem, which states that a sum of (same-distribution) random variables is more Gaussian than any individual one. Hence, by maximising the non-Gaussianity of a linear mixture of observations, each linear mixture should end up being equal to one of the original sources [Hyvärinen et al., 2001, ch. 8]. Finally, TDSEP minimises a weighted sum of the cross-correlation of sources across several time lags and a measure of dependence between sources across several time lags [Ziehe and Müller, 1998].

2.1.2 Spatio-spectral decomposition

Spatio-Spectral Decomposition (SSD) aims at finding oscillatory sources in a pre-specified frequency band [Nikulin et al., 2011]. To do this, SSD finds linear combinations of the observed signals such that the power of the linear mixtures is maximised in a pre-specified frequency band relative to surrounding narrow (1-2 Hz) frequency bands. The optimal projection vectors are found as the generalised eigenvectors corresponding to the highest generalised eigenvalues in the following generalised eigenvalue problem [Nikulin et al., 2011]:

$$\mathbf{\Sigma}_s \mathbf{W} = \lambda \mathbf{\Sigma}_n \mathbf{W},$$

where $\mathbf{\Sigma}_s$ is the average (over time) covariance matrix of the band-pass filtered data in the frequency range of interest. The matrix $\mathbf{\Sigma}_n$ is the time-averaged covariance matrix of the sum of the band-pass filtered data in the two narrow frequency bands surrounding the frequency range of interest.

SSD is likely to be a good choice for extracting neural activity expected to be oscillatory in one or several relatively narrow frequency bands. On the other

hand, SSD can not be expected to extract artefactual components since these are found at a wide range of frequencies and most are not oscillatory.

2.1.3 Fourier-ICA

Similar to SSD, the goal of Fourier-ICA is to extract oscillatory sources [Hyvärinen et al., 2010]. Fourier-ICA splits data into short time windows and finds their Fourier-transforms. To retain the matrix structure, the time window indices and Fourier coefficient indices are concatenated for each channel. This new data matrix is subjected to FastICA, optionally allowing for complex mixing coefficients. Since FastICA maximises non-Gaussianity, the estimated sources will be sub- or super-Gaussian. The mixing coefficients for super-Gaussian sources will be sparse, i.e. either be zero or far from zero [Hyvärinen et al., 2010]. Hence such sources consist of contributions from a small number of frequencies. For the Fourier transform, the frequency band for which coefficients should be retained must be chosen since only a finite number of Fourier coefficients can be used.

The pre-specified frequency band for Fourier-ICA can be quite wide [Hyvärinen et al., 2010], making Fourier-ICA a relevant method when the components of interest are expected to be oscillatory, but only limited knowledge of their frequencies is available beforehand. As for SSD, Fourier-ICA should not be expected to extract artefactual sources since artefacts are typically not oscillatory.

2.2 Supervised matrix method

2.2.1 Common Spatial Patterns

CSP finds sources whose variance (or power, for zero-mean data) differs maximally between two conditions [Blankertz et al., 2008]. Hence, if the neural activity that differs between two conditions is expected to be oscillatory and have different spatial expressions on the scalp, CSP is a fitting method. This is precisely the case for motor-imagery based tasks [Blankertz et al., 2008].

Formally, filters $\boldsymbol{w}$ that maximise the following objective function are sought:

$$\underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{\boldsymbol{w}^T \boldsymbol{\Sigma}_1 \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{\Sigma}_2 \boldsymbol{w}}, \tag{2.1}$$

where the matrices $\mathbf{\Sigma}_i$ are the covariance matrices of the two trial classes. The optimal filters are the generalised eigenvectors of the generalised eigenvalue problem $\mathbf{\Sigma}_1 \boldsymbol{w} = \lambda \mathbf{\Sigma}_2 \boldsymbol{w}$. The eigenvectors corresponding to the largest values of λ maximise the ratio of the variance in class 1 to that in class 2 while the eigenvectors corresponding to the smallest values maximise the ratio of the variance in class 2 to that in class 1.

There are different ways to calculate the channel $\times$ channel matrices $\mathbf{\Sigma}_i$, e.g. as the average of covariance matrices from all condition i trials or as the covariance matrix of concatenated trials of type i [Kawanabe and Vidaurre, 2009, Allefeld et al., 2008, Yger et al., 2015]. Since covariance matrix estimates are vulnerable to outliers, it may be advantageous to use a robust estimate of the covariance matrices through regularisation [Kawanabe et al., 2014].

2.2.1.1 Regularisation of spatial filter estimates

The basic formulation of CSP is only rarely used since performance can usually be improved by regularising the spatial filters with a matrix $\mathbf{R}$:

$$\frac{\boldsymbol{w}^T \mathbf{\Sigma}_1 \boldsymbol{w}}{\boldsymbol{w}^T ((1-\lambda)\mathbf{\Sigma}_2 + \lambda \mathbf{R}) \boldsymbol{w}}. \tag{2.2}$$

A simple choice for $\mathbf{R}$ is the identity matrix [Lotte and Guan, 2011]. Alternatives use information from previous trials or other subjects [Lotte and Guan, 2011, Samek and Muller, 2014]. For regularised CSP, the cost-function ratio is no longer symmetric between the two classes. Hence Equation 2.2 must also be evaluated with switched roles of the covariance matrices [Lotte and Guan, 2011].

2.2.1.2 Choosing which filters to use and classification

CSP can find as many filters as there are linearly independent channels, but usually only a few are needed to capture the class-discriminative information. When used in motor-imagery based BCI paradigms, usually two classes are used. Common classes are motor imagery of moving the left vs. right hand. With two classes, three filters are often chosen to maximise each of the variance ratios ($\frac{\boldsymbol{w}^T \mathbf{\Sigma}_1 \boldsymbol{w}}{\boldsymbol{w}^T ((1-\lambda)\mathbf{\Sigma}_2 + \lambda \mathbf{R}) \boldsymbol{w}}$ and $\frac{\boldsymbol{w}^T \mathbf{\Sigma}_2 \boldsymbol{w}}{\boldsymbol{w}^T ((1-\lambda)\mathbf{\Sigma}_1 + \lambda \mathbf{R}) \boldsymbol{w}}$), resulting in a total of six filters [Blankertz et al., 2008].

Features for each trial are found by spatially filtering the data with each of the chosen filters and calculating the logarithm of the band-power [Blankertz et al., 2008]. These features can then be used for single-trial classification.

Chapter 3

Multi-linear (tensor) methods

Tensors are multidimensional arrays that generalise the concepts of vectors and matrices. Zeroth order tensors are scalars, first order tensors (one-modal) tensors are vectors, second order (two-modal) tensors are matrices, third-order (three-modal) tensors can be thought of as layers of matrices, etc. Representing data as tensors during analysis allows exploitation of multidimensional relations in data. Similar to the matrix methods described in Chapter 2, unsupervised tensor methods aim to find factors that reconstruct data with the least possible error while supervised tensor methods aim to find subspaces that separate the projections of tensors from different classes maximally.

For EEG data, this means that sources can be extracted or trials be classified while taking the variation in the spatio-temporal structure over trials into account. Since a, to some degree, consistent spatio-temporal structure is likely to exist in typical ERP and motor imagery paradigms, multi-linear methods should be able to model EEG data more efficiently than standard methods. The considerations described in Chapter 1 regarding linear modeling of EEG data also apply to multi-linear modeling.

One observation from Chapter 1 might be worth revisiting and extending to the multi-linear case. We have already seen why spatial patterns, and not their corresponding filters, should be used to interpret the spatial expression on the

scalp of a source. Sources isolated by multi-linear methods are characterised by filters/patterns for all modes. Assuming data consists of *channel* $\times$ *time* matrix observations, this means that a source is characterised both by a spatial filter/pattern (as for the linear methods) and a temporal filter/pattern. The unsupervised multi-linear methods used in this thesis estimate the forward model directly. Hence the factors extracted by these methods correspond to patterns and can be interpreted directly as the spatial and temporal expressions of a source. Conversely, the supervised multi-linear methods extract filters that optimise some objective function of the filtered data. Such filters contain both coefficients that enhance signals that improve the objective function and coefficients that minimise the influence of disrupting signals. Hence the filters cannot be interpreted directly. Analogously to the linear case, we pre-multiply the spatial and temporal filters by the spatial and temporal covariance matrices, respectively, to interpret the sources.

3.1 Tensor operations and characterisations

We denote the number of modes by P and let J_p denote the dimension of mode p. An index of mode p is referred to as j_p.

3.1.1 Modes

The modes of a tensor correspond to the rows and columns of a matrix. A three-dimensional tensor can be viewed as a line-up of standing matrices. In a 3D tensor, elements with the same mode-ond index are arranged in the same horizontal stratum (rows in a matrix), elements with the same mode-two index are arranged in the same vertical stratum (columns in a matrix), and elements with the same mode-three index are in the same line in the line-up of matrices.

3.1.2 Fibres

A fibre is a sequence of scalars, i.e. a vector. A fibre of a tensor is obtained by fixing every index except one. In a matrix, each row and column is a fibre. The fibre obtained when keeping all indices except j_p fixed is referred to as a mode-p fibre and there are J_p mode-p fibres [Mørup, 2011, Li and Schonfeld, 2014].

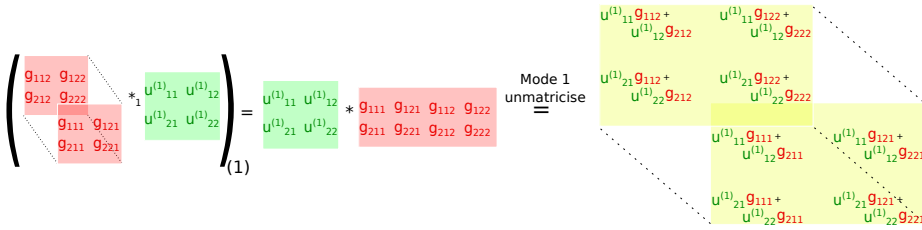


Figure 3.1: Example of mode-1 multiplication of a tensor by a matrix.

3.1.3 Matricising

When a tensor is matricised, its elements are rearranged in the form of a matrix. This operation is also referred to as flattening. A tensor can be matricised along each of its modes, so which mode is being used must be specified. Matricising tensor $\mathcal{X}$ along mode p is written as $\mathbf{X}_{(p)}$ and referred to as mode-p matricising. Matricising is performed by arranging all mode-p fibres as the columns of a matrix whose dimensions then become $J_p \times \prod_{q=1,2,\ldots,p-1,p+1,\ldots,P} J_q$ [Mørup, 2011, Li and Schonfeld, 2014].

In a two-dimensional tensor (a matrix), matricising along mode one has no effect while matricising along mode two transposes the matrix.

3.1.4 Mode-p multiplication

The multiplication of a tensor $\mathcal{X}$ along its p^{th} mode by a matrix $\mathbf{U}$ is written as $\mathcal{Y} = \mathcal{X} \times_p \mathbf{U}$. The multiplication can be performed through standard matrix multiplication by matricising the tensor to obtain $\mathbf{Y}_{(p)} = \mathbf{U} \times \mathbf{X}_{(p)}$. The final result is found by unmatricising the matrix product as illustrated for an example of mode-one multiplication in Figure 3.1.

We use the following notation for multiplication of a tensor along several modes with matrices $\mathbf{U}^{(p)}$:

$$\mathcal{X} \times_1 \mathbf{U}^{(1)} \ldots \times_P \mathbf{U}^{(P)} = \mathcal{X} \times_{p=1}^{P} \mathbf{U}^{(p)}$$

3.1.5 Kronecker and Khatri-Rao products

The Kronecker product of an $m \times \ell$ matrix $\mathbf{A}$ with elements $a_{i,j}$ and a $p \times q$ matrix $\mathbf{C}$ is defined as Mørup [2011]:

$$\mathbf{A} \otimes \mathbf{C} = \begin{pmatrix} a_{1,1}\mathbf{C} & a_{1,2}\mathbf{C} & \cdots & a_{1,\ell}\mathbf{C} \\ a_{2,1}\mathbf{C} & a_{2,2}\mathbf{C} & \cdots & a_{2,\ell}\mathbf{C} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}\mathbf{C} & a_{m,2}\mathbf{C} & \cdots & a_{m,\ell}\mathbf{C} \end{pmatrix}.$$

The Khatri-Rao product is the column-wise Kronecker product and is denoted by $\odot$. That is, column j of the matrix $\mathbf{D} = \mathbf{A} \odot \mathbf{C}$ is the Kronecker product of the j^{th} columns of $\mathbf{A}$ and $\mathbf{C}$: $\mathbf{D}_{:,j} = \mathbf{A}_{:,j} \otimes \mathbf{C}_{:,j}$ Mørup [2011]. Hence the matrices $\mathbf{A}$ and $\mathbf{C}$ must have the same number of columns for their Khatri-Rao product to be defined.

3.2 Unsupervised tensor decompositions

The multi-linear generalisation of the linear model (1.1) of EEG data is:

$$\mathcal{X} = \mathcal{G} \times_{p=1}^{P} \mathbf{U}^{(p)} + \mathcal{E}, \tag{3.1}$$

where the tensor $\mathcal{G}$ is referred to as the core array and determines how the $\mathbf{U}^{(p)}$ matrices interact. The tensor $\mathcal{E}$ contains model errors and is sought minimised, often under the Frobenius-norm using Alternating Least Squares [Mørup, 2011, Kiers et al., 1999]. An example with a 3-modal $2 \times 2 \times 2$ data tensor and core array is shown in detail in Appendix F.

Some tensor models are described in the following.

3.2.1 Tucker

The Tucker model assumes that the systematic variations in data can be described by a product such as

$$\mathcal{X} \approx \mathcal{G} \times_{p=1}^{P} \mathbf{U}^{(p)}$$

without further assumptions on $\mathcal{G}$ or $\mathbf{U}^{(p)}$ [Tucker, 1966, Mørup, 2011]. This model is invariant to multiplication of the matrices $\mathbf{U}^{(p)}$ by invertible matrices. To constrain the amount of rotational invariance, the matrices $\mathbf{U}^{(p)}$ can be required to be orthonormal. However, this still leaves the model invariant to multiplication by orthonormal matrices [Mørup, 2011]. The number of factors, i.e. number of columns in the $U^{(p)}$ matrices, to choose in each mode does not have a well-defined answer and does not have be the same for all modes.

3.2.2 Tucker2

The Tucker2 model assumes the same data structure as the Tucker model, but leaves one mode uncompressed [Kolda and Bader, 2009, Mørup, 2011, Tucker, 1966]. That is, there is one mode for which a $U^{(p)}$ matrix is not estimated, but the identity matrix is used instead. This model can be used to compress each observation, $\mathcal{X}_n \in \mathbb{R}^{J_1 \times J_2 \times \ldots J_P}$ in a collection of N observations $\mathcal{X} \in \mathbb{R}^{J_1 \times J_2 \times \ldots J_P \times N}$ by compressing all modes except mode $P+1$ [Liu et al., 2011].

3.2.3 PARAFAC

The PARAFAC model imposes some restrictions on the form of the product $\mathcal{G} \times_{p=1}^{P} \mathbf{U}^{(p)}$ [Harshman, 1970, Carroll and Chang, 1970]. This model is also referred to as the PARAFAC1 model [Kiers et al., 1999, Bro et al., 1999]. Non-zero entries in the core array $\mathcal{G}$ are only allowed on the diagonal (the elements $\mathcal{G}_{i,i,\ldots,i}$) and the number of columns must be the same in all $\mathbf{U}^{(p)}$ matrices. The k^{th} columns in the matrices $\mathbf{U}^{(p)}$ are sometimes collectively referred to as the k^{th} factor. These restrictions remove much of the invariance in the Tucker model so that the PARAFAC model is invariant up to scaling, signs, and the ordering of factors [Mørup, 2011]. This simplifies the example multiplication given in Appendix F to result in the following approximation of $\mathcal{X}$:

$$\tilde{\mathcal{X}}^{(3)}_{:,:,1} \approx \begin{pmatrix} u^{(3)}_{11}u^{(2)}_{11}u^{(1)}_{11}g_{111} + u^{(3)}_{12}u^{(2)}_{12}u^{(1)}_{12}g_{222} & u^{(3)}_{11}u^{(2)}_{21}u^{(1)}_{11}g_{111} + u^{(3)}_{12}u^{(2)}_{22}u^{(1)}_{12}g_{222} \\ u^{(3)}_{11}u^{(2)}_{11}u^{(1)}_{21}g_{111} + u^{(3)}_{12}u^{(2)}_{12}u^{(1)}_{22}g_{222} & u^{(3)}_{11}u^{(2)}_{21}u^{(1)}_{21}g_{111} + u^{(3)}_{12}u^{(2)}_{22}u^{(1)}_{22}g_{222} \end{pmatrix}$$

$$\tilde{\mathcal{X}}^{(3)}_{:,:,2} \approx \begin{pmatrix} u^{(3)}_{21}u^{(2)}_{11}u^{(1)}_{11}g_{111} + u^{(3)}_{22}u^{(2)}_{12}u^{(1)}_{12}g_{222} & u^{(3)}_{21}u^{(2)}_{21}u^{(1)}_{11}g_{111} + u^{(3)}_{22}u^{(2)}_{22}u^{(1)}_{12}g_{222} \\ u^{(3)}_{21}u^{(2)}_{11}u^{(1)}_{21}g_{111} + u^{(3)}_{22}u^{(2)}_{12}u^{(1)}_{22}g_{222} & u^{(3)}_{21}u^{(2)}_{21}u^{(1)}_{21}g_{111} + u^{(3)}_{22}u^{(2)}_{22}u^{(1)}_{22}g_{222} \end{pmatrix}.$$

From the above example we see that the only difference between $\mathcal{X}_{:,:,1}$ and $\mathcal{X}_{:,:,2}$ is the contributions from $\mathbf{U}^{(3)}$, i.e. the trial-mode ($\mathcal{X}_{:,:,i}$ corresponds to trial i). Hence the only difference between trials is the strength of the factors' presence, which is given by $\mathbf{U}^{(3)}_{n,k}$ for the k^{th} factor in the n^{th} observation. This implies that the PARAFAC model is well-suited to finding components with patterns that are consistent over all observations, but differ in their magnitudes of contribution to each observation.

We now give some different formulations of the PARAFAC model that imply different ways of understanding the model. Let the vector $\boldsymbol{g} = diag(\mathcal{G})$ consist of the diagonal elements of the core array. If there are K factors and the data has two modes such that the tensorr $\mathcal{X}$ containing all observations has three modes, the elementwise reconstruction of $\mathcal{X}$ is

$$x_{hij} = \sum_{k=1}^{K} u^{(1)}_{hk}u^{(2)}_{ik}u^{(3)}_{jk}\boldsymbol{g}_k + e_{hij}.$$

The above expression shows how the factors combine to reconstruct data. From this expression, it is clear that the k^{th} column of $\mathbf{U}^{(p)}$ can only interact with the k^{th} columns of the matrices estimated for the other modes. Since $\boldsymbol{g}_k$ only interacts with the k^{th} column of each matrix $\mathbf{U}^{(p)}$, we can set $\tilde{u}^{(3)}_{:,k} = u^{(3)}_{:,k}g_k$ and simplify the above expression to

$$x_{hij} = \sum_{k=1}^{K} u^{(1)}_{hk}u^{(2)}_{ik}\tilde{u}^{(3)}_{jk} + e_{hij}.$$

The above elimination of the core array makes it straight-foward to rewrite the model as a matrix product, where each observation $\mathcal{X}_n$ is a matrix :

$$\begin{aligned}\mathcal{X}_n &\approx \mathbf{U}^{(1)} diag(\mathbf{U}^{(3)}_{n,:})\mathbf{U}^{(2)\top} \\ &= \sum_{k=1}^{K} \mathbf{U}^{(1)}_{:,k}(\mathbf{U}^{(2)}_{:,k}\mathbf{U}^{(3)}_{n,k})^\top \text{ for matrix observations.}\end{aligned}$$

This expression makes it clear that each matrix observation is expressed as a sum of K outer products of estimated factors for the row- and column-modes, weighted by an observation-specific weight for the component contribution. For a general number of modes, P, the reconstruction of $\mathcal{X}$ can be written as

$$\mathcal{X} \approx \mathcal{I} \times_{p=1}^{P} \mathbf{U}^{(p)} = \sum_{k=1}^{K} \mathcal{I} \times_{p=1}^{P} U^{(p)}_{:,k},$$

where $\mathcal{I}$ is the P-modal identity tensor of size $K \times K \ldots K$.

3.2.4 PARAFAC2

The PARAFAC2 model is an extension of PARAFAC that is more flexible. For simplicity, we assume that data is represented as a 3D tensor with observations stacked over the third mode. Then, instead of requring the factors for the second mode to be the same for all observations, PARAFAC2 assumes a constant cross-product of the second-mode factors. The PARAFAC2 model for observation n can be written as:

$$\mathcal{X}_n \approx \mathbf{U}^{(1)} diag(\mathbf{U}^{(3)}_{n,:})\mathbf{U}^{(2)\top}_n.$$

In the PARAFAC2 model, the constant term $\mathbf{U}^{(2)}$ is replaced by the observation-specific matrices $\mathbf{U}^{(2)}_n$, with the above-mentioned constraint: $\mathbf{U}^{(2)}_n\mathbf{U}^{(2)\top}_n = \mathbf{Q}\,\forall\, n$. By factorising $\mathbf{U}^{(2)}_n = \mathbf{P}_n\mathbf{F}$, with orthonormal $\mathbf{P}_n$, a matrix, $\mathbf{F}$ common to all observations is obtained [Kiers et al., 1999].

3.3 Supervised tensor methods

Before describing existing tensor methods for finding projections that optimally separate classes, we describe corresponding methods for vector observations. Assuming equal covariances and numbers of observations in two classes, Fisher

proposed to optimise the following measure of class separation to find the projection vector, $\boldsymbol{u}$, that best separates observations from different classes:

$$f_{fisher}(\boldsymbol{u}) = \frac{\boldsymbol{u}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{u}}{\sum_{c=1}^{2}\sum_{\{n:class(\boldsymbol{x}_n)=c\}}\boldsymbol{u}^\top(\boldsymbol{x}_n - \boldsymbol{\mu}_c)(\boldsymbol{x}_n - \boldsymbol{\mu}_c)^\top\boldsymbol{u}}, \tag{3.2}$$

where $\boldsymbol{\mu}_c$ is the mean of observations from class c [Fisher, 1936]. The optimal projection vector is found to be $(\sum_{c=1}^{2}\sum_{\{n:class(\boldsymbol{x}_n)=c\}}(\boldsymbol{x}_n - \boldsymbol{\mu}_c)^2)^{-1} \times (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

This can also be derived by assuming the classes are normally distributed with the same covariance matrix but different means, as done in Linear Discriminant Analysis (LDA). By manipulating the expression for the log of the ratio of the probabilities of an observation belonging to one class relative to the other, an expression linear in the observation vector $\boldsymbol{x}$ is obtained [Hastie et al., 2009, p. 108]. The coefficient (or projection) vector that optimally separates observations from the two classes turns out to be $\boldsymbol{\Sigma}^{-1} \times (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, where $\boldsymbol{\Sigma}$ is the estimated common covariance matrix. Since, with equal numbers of observations in each class, $\boldsymbol{\Sigma} \propto \sum_{c=1}^{2}\sum_{\{n:class(\boldsymbol{x}_n)=c\}}\boldsymbol{u}^\top(\boldsymbol{x}_n - \boldsymbol{\mu}_c)[\boldsymbol{u}^\top(\boldsymbol{x}_n - \boldsymbol{\mu}_c)]^\top$, the projection vector obtained from the likelihood perspective under the LDA assumptions is the same as that obtained by Fisher. LDA can be extended to the multi-class case of C classes by changing the Fisher criterion to sum over all classes in the denominator, making the common covariance estimate based on all classes, and changing the numerator so that it is the sum of the squared differences between the class means and the overall mean [Bishop, 2006]:

$$f_{LDA,multi-class}(\mathbf{U}_{LDA}) = tr((\mathbf{U}_{LDA}\mathbf{W}_{LDA}\mathbf{U}_{LDA}^\top)^{-1}\mathbf{U}_{LDA}\mathbf{B}_{LDA}\mathbf{U}_{LDA}^\top), \tag{3.3}$$

where $\mathbf{U}_{LDA}$ is a matrix containing projection vectors in its columns, the matrix

$$\mathbf{W}_{LDA} = \sum_{c=1}^{C}\sum_{\{n:class(\boldsymbol{x}_n)=c\}}(\boldsymbol{x}_n - \boldsymbol{\mu}_c)(\boldsymbol{x}_n - \boldsymbol{\mu}_c)^\top$$

is an estimate of the common covariance within classes, and the matrix

$$\mathbf{B}_{LDA} = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top$$

quantifies the variance between classes. Then the optimal projection matrix $\mathbf{U}_{LDA}$ is defined by the eigenvectors corresponding to the highest eigenvalues of the generalised eigenvalue problem: $\mathbf{B}_{LDA}\mathbf{U}_{LDA} = \lambda\mathbf{W}_{LDA}\mathbf{U}_{LDA}$. Other functions quantifying class differences are also possible [Bishop, 2006].

The Fisherface algorithm [Belhumeur et al., 1997] is an example of a different criterion employing $\mathbf{B}_{LDA}$ and $\mathbf{W}_{LDA}$. The aim in [Belhumeur et al., 1997] was to classify vectorised images into one of multiple classes, via the criterion:

$$f_{fisherface}(\mathbf{U}_{LDA}) = \frac{\det\left(\mathbf{U}_{LDA}^{\top}\mathbf{B}_{LDA}\mathbf{U}_{LDA}\right)}{\det\left(\mathbf{U}_{LDA}^{\top}\mathbf{W}_{LDA}\mathbf{U}_{LDA}\right)}. \tag{3.4}$$

This criterion is also solved by the generalised eigenvalue problem: $\mathbf{B}_{LDA}\mathbf{U}_{LDA} = \lambda\mathbf{W}_{LDA}\mathbf{U}_{LDA}$.

The first suggestion of a supervised algorithm retaining the matrix structure of observations instead of vectorising them was described by Li and Yuan [2005] with the proposal of 2-dimensional LDA (2DLDA). The algorithm 2DLDA combines the idea of retaining the natural matrix representation of data (images) as suggested by Yang et al. [2004] with the idea of supervising the learning of discriminant features as done by Belhumeur et al. [1997]. Instead of vectorising matrix observations when calculating the within- and between-class scatter matrices, observations were kept as matrices by Li and Yuan [2005] such that

$$\begin{aligned} \mathbf{W}_{2DLDA} &= \textstyle\sum_{c=1}^{C}(\bar{\mathbf{X}}_c - \bar{\mathbf{X}})(\bar{\mathbf{X}}_c - \bar{\mathbf{X}})^{\top} \\ \mathbf{B}_{2DLDA} &= \textstyle\sum_{c=1}^{C} N_c \sum_{\{n:class(\mathbf{X}_n)=c\}}(\mathbf{X}_n - \bar{\mathbf{X}}_c)(\mathbf{X}_n - \bar{\mathbf{X}}_c)^{\top}, \end{aligned} \tag{3.5}$$

where $\bar{\mathbf{X}}_c$ is the matrix mean of class c and $\bar{\mathbf{X}}$ is the overall mean of all observations. Li and Yuan [2005] optimised the multi-class LDA objective:

$$f_{2DLDA}(\mathbf{U}) = tr\left((\mathbf{U}^{\top}\mathbf{W}_{2DLDA}\mathbf{U})^{-1}\mathbf{U}^{\top}\mathbf{B}_{2DLDA}\mathbf{U}\right). \tag{3.6}$$

For the criteria previous to f_{2DLDA}, the matrix $\mathbf{W}_{LDA}$ was formed from a sum of rank-one matrices (cross-products of vectors). Hence $\mathbf{W}_{LDA}$ would be singular unless there were more terms in the sum than elements in each observation vector. However, the within-class scatter matrix used in 2DLDA, $\mathbf{W}_{2DLDA}$ is a sum of matrix cross-products. As shown by Li and Yuan [2005], the requirement for $\mathbf{W}_{2DLDA}$ to be non-singular is $N \geq C + \frac{J_2}{J_1+J_2}$, where N is the number of

observations, C the number of classes, and J_1 and J_2 the numbers of rows and columns, respectively. Hence the approach taken with the 2DLDA algorithm allows interactions across modes to be taken into account and alleviates problems associated with high-dimensional observations [Hastie et al., 2009, ch. 18].

3.3.1 Feature extraction

3.3.1.1 CMDA and DATER

While only the column-dimension of matrices was compressed in 2DLDA [Li and Yuan, 2005], two suggestions of generalisation to an arbitrary number of modes, compressing all modes, were proposed soon after by two different teams [Yan et al., 2005, Visani et al., 2005]. The more general method proposed by Yan et al. [2005] was called Discriminant Analysis with TEnsor Representation (DATER). The proposal by Visani et al. [2005] was very similar. While DATER optimises by solving a generalised eigenvalue problem, the algorithm proposed by Visani et al. [2005] is optimised by solving a corresponding standard eigenvalue problem. While DATER did not have convergence guarantees, a modified version, referred to as Constrained Multilinear Discriminant Analysis (CMDA), with a convergence guarantee was proposed in 2014 [Li and Schonfeld, 2014]. Both DATER and CMDA seek to optimise the scatter ratio objective:

$$f_{sr}(\mathbf{U}^{(p)}|_{p=1}^{P}) = \frac{\sum_{c=1}^{C} N_c \|(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{p=1}^{P} \mathbf{U}^{(p)\top}\|_F^2}{\sum_{c=1}^{C} \sum_{\{n:\, class(\mathcal{X}_n)=c\}} \|(\mathcal{X}_n - \bar{\mathcal{X}}_c) \times_{p=1}^{P} \mathbf{U}^{(p)\top}\|_F^2}, \tag{3.7}$$

where $\bar{\mathcal{X}}$ is the mean of all observations, N the number of observations, N_c the number of observations in class c, $\bar{\mathcal{X}}_c$ the mean of observations in class c, C the number of classes, $\{n : class(\mathcal{X}_n) = c\}$ the set of indices of observations from class c, and $\mathbf{U}^{(p)}$ is the projection matrix for the p^{th} mode. For 2DLDA, the projection matrix for the first mode would be the identity matrix [Yan et al., 2005]. We now give some technical details about CMDA and DATER.

Isolating a projection matrix The Frobenius norms in the scatter ratio and scatter difference (used in the method DGTDA) objective functions can be rewritten using the Kronecker product since the Frobenius norm of a tensor is the same as the Frobenius norm of its unfolding along any mode. This allows us to move a projection matrix outside the tensor product [Li and Schonfeld, 2014]:

$$
\begin{aligned}
\|(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{p=1}^{P} \mathbf{U}^{(p)\top}\|_F^2 &= \\
\text{[Cichocki et al., 2009, Eq. 1.102]} &= \|\mathbf{U}^{(p)\top} \times (\bar{\mathcal{X}}_c - \bar{\mathcal{X}})_{(p)} \times \left(\otimes_{q=P,q\neq p}^{1} \mathbf{U}^{(q)\top}\right)^{\top} \|_F^2 \\
&= tr\Bigg[\left(\mathbf{U}^{(p)\top} \times (\bar{\mathcal{X}}_c - \bar{\mathcal{X}})_{(p)} \times \left(\otimes_{q=P,q\neq p}^{1} \mathbf{U}^{(q)\top}\right)^{\top}\right) \dots \\
&\quad \left(\mathbf{U}^{(p)\top} \times (\bar{\mathcal{X}}_c - \bar{\mathcal{X}})_{(p)} \times \left(\otimes_{q=P,q\neq p}^{1} \mathbf{U}^{(q)\top}\right)^{\top}\right)^{\top} \Bigg] \\
&= tr\Bigg[\left(\mathbf{U}^{(p)\top} \times (\bar{\mathcal{X}}_c - \bar{\mathcal{X}})_{(p)} \times \left(\otimes_{q=P,q\neq p}^{1} \mathbf{U}^{(q)\top}\right)^{\top}\right) \dots \\
&\quad \left((\bar{\mathcal{X}}_c - \bar{\mathcal{X}})_{(p)} \times \left(\otimes_{q=P,q\neq p}^{1} \mathbf{U}^{(q)\top}\right)^{\top}\right)^{\top} \mathbf{U}^{(p)} \Bigg] \\
&= tr\Bigg[\mathbf{U}^{(p)\top} \left[(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{q=1,q\neq p}^{P} \mathbf{U}^{(q)\top}\right]_{(p)} \dots \\
&\quad \left[(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{q=1,q\neq p}^{P} \mathbf{U}^{(q)\top}\right]_{(p)}^{\top} \mathbf{U}^{(p)} \Bigg]
\end{aligned}
$$

Define the between-class scatter, $\mathbf{B}_p^{\tilde{p}}$, and within-class scatter, $\mathbf{W}_p^{\tilde{p}}$ matrices, that do not project unto the p^{th} mode as

$$
\begin{aligned}
\mathbf{B}_p^{\tilde{p}} = &\quad \sum_{c=1}^{C} N_c \left[(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{q=1,q\neq p}^{P} \mathbf{U}^{(q)\top}\right]_{(p)} \left[(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{q=1,q\neq p}^{P} \mathbf{U}^{(q)\top}\right]_{(p)}^{\top} \\
\mathbf{W}_p^{\tilde{p}} = &\quad \sum_{c=1}^{C} \sum_{\{n:\, class(\mathcal{X}_n)=c\}} \left[\mathcal{X}_n - \bar{\mathcal{X}}_c) \times_{q=1,q\neq p}^{P} \mathbf{U}^{(q)\top}\right]_{(p)} \dots \\
&\left[\mathcal{X}_n - \bar{\mathcal{X}}_c) \times_{q=1,q\neq p}^{P} \mathbf{U}^{(q)\top}\right]_{(p)}^{\top}.
\end{aligned}
$$

Then the scatter ratio objective function can be written as

$$
f_{sr}(\mathbf{U}^{(p)}|_{p=1}^{P}) = \frac{tr(\mathbf{U}^{(p)\top} \mathbf{B}_p^{\tilde{p}} \mathbf{U}^{(p)})}{tr(\mathbf{U}^{(p)\top} \mathbf{W}_p^{\tilde{p}} \mathbf{U}^{(p)})} \tag{3.8}
$$

Optimisation As seen in Appendix G, the scatter ratio objective does not have an analytical solution. Hence DATER and CMDA use the iterative scheme shown in Algorithm (1). DATER updates the projection matrix $\mathbf{U}^{(p)}$ by setting its columns equal to the first K_p eigenvectors resulting from the generalised eigenvalue problem $\mathbf{B}_p^{\tilde{p}} \mathbf{U}^{(p)} = \mathbf{W}_p^{\tilde{p}} \mathbf{U}^{(p)} \Lambda_p$ while CMDA finds $\mathbf{U}^{(p)}$ as the first K_p left singular vectors of $\mathbf{W}_p^{\tilde{p}^{-1}} \mathbf{B}_p^{\tilde{p}}$, where K_p is the number of components fitted for the p^{th} mode. Hence the $\mathbf{U}^{(p)}$ matrices found by CMDA are

guaranteed to be orthonormal. On the other hand, the eigenvectors found by DATER are $\mathbf{W}_p^{\tilde{p}}$-orthogonal ($\mathbf{U}^{(p)\top}\mathbf{W}_p^{\tilde{p}}\mathbf{U}^{(p)} = \mathbf{\Lambda}$). This also implies that the constraints are different for the different modes. If $\mathbf{W}_p^{\tilde{p}}$ is invertible then the generalised eigenvalue problem solved by DATER is equivalent to the standard eigenvalue problem $\mathbf{W}_p^{\tilde{p}^{-1}}\mathbf{B}_p^{\tilde{p}}\mathbf{U}^{(p)} = \mathbf{U}^{(p)}\Lambda_p$, which is the one solved by Visani et al. [2005]. Since projection matrices defined by eigenvectors are orthonormal, the constraints are the same as in CMDA in this case. In this case, since the left singular vectors of a square matrix, $\mathbf{C}$, are the eigenvectors of $\mathbf{C}\mathbf{C}^\top$, DATER finds the eigenvectors of $\mathbf{W}_p^{\tilde{p}^{-1}}\mathbf{B}_p^{\tilde{p}}$ while CMDA finds the eigenvectors of $\mathbf{W}_p^{\tilde{p}^{-1}}\mathbf{B}_p^{\tilde{p}}\mathbf{B}_p^{\tilde{p}\top}\mathbf{W}_p^{\tilde{p}\top}$.

The two algorithms also differ in how they initialise the $\mathbf{U}^{(p)}$ matrices and their convergence criteria. DATER initialises each $\mathbf{U}^{(p)}$ as an identity matrix while CMDA uses matrices where all elements are 1 for initialisation. DATER uses the criterion $\|\mathbf{U}^{(p),it} - \mathbf{U}^{(p),it-1}\| < K_p J_p \epsilon \,\forall\, p$ to check convergence. Here, $\mathbf{U}^{(p),it}$ is the projection matrix obtained for the p^{th} mode at iteration it and ϵ is a predetermined tolerance parameter. CMDA uses the stopping criterion $\sum_{i=1}^{N} \|U_n^{it} U_n^{it-1T} - I\| \leq \epsilon$, where $\mathbf{I}$ is the identity matrix.

Algorithm 1 DATER and CMDA steps

procedure DATER/CMDA(Xs, classes, lowerdims)
Xs: collection of data. Each observation is a $J_1 \times J_2 \times \ldots \times J_P$ tensor
classes: class for each observation
K_1, K_2, …, K_P: the number of components (columns) to fit for each mode
  Initialise $\mathbf{U}^{(p)} \forall p \in \{1, 2, \ldots, P\}$ with dimensions $J_p \times K_p$
  while $its < maxits \wedge notconverged$ **do**
    for $p \leftarrow 1$, P **do**
      Calculate $\mathbf{B}_p^{\tilde{p}}$ and $\mathbf{W}_p^{\tilde{p}}$ using current $\mathbf{U}^{(q)}$, $q \neq p$ matrices
      Update $\mathbf{U}^{(p)}$ using $\mathbf{B}_p^{\tilde{p}}$ and $\mathbf{W}_p^{\tilde{p}}$
    end for
  end while
end procedure

3.3.1.2 DGTDA

Instead of using the ratio of the within- and between-class scatters, their difference could also be used as measure of the class separation of the projected observations. The formal formulation of this objective function is

$$\begin{aligned} f_{sd}(\mathbf{U}^{(p)}|_{p=1}^{P}) &= \sum_{c=1}^{C} N_c \|(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \prod_{p=1}^{P} \times_p \mathbf{U}^{(p)\top}\|_F^2 \\ &- \sum_{c=1}^{C} \sum_{\{n:\, class(\mathcal{X}_n)=c\}} \|(\mathcal{X}_n - \bar{\mathcal{X}}_c) \prod_{p=1}^{P} \times_p \mathbf{U}^{(p)\top}\|_F^2. \end{aligned} \tag{3.9}$$

This optimisation criterion was first proposed by Tao et al. [2007], where it was optimised iteratively by an algorithm referred to as General Tensor Discriminant Analysis (GTDA). Li and Schonfeld [2014] proposed a direct optimisation of the objective (3.9), only passing over each modality once. This algoritm was named Direct GTDA (DGTDA). DTGDA uses the within- and between-class scatters:

$$\begin{aligned} \mathbf{B}_{DGTDA}^{(p)} &= \sum_{c=1}^{C} N_c (\bar{\mathcal{X}}_c - \bar{\mathcal{X}})_{(p)} (\bar{\mathcal{X}}_c - \bar{\mathcal{X}})_{(p)}^{\top} \\ \mathbf{W}_{DGTDA}^{(p)} &= \sum_{c=1}^{C} \sum_{\{n:class(\mathcal{X}_n)=c\}} (\mathcal{X}_n - \bar{\mathcal{X}}_c)_{(n)} (\mathcal{X}_n - \bar{\mathcal{X}}_c)_{(p)}^{\top}. \end{aligned}$$

Then DGTDA finds the projection matrix $\mathbf{U}^{(p)}$ for each mode as the first K_p singular vectors from the singular value decomposition of $\mathbf{B}_{DGTDA}^{(p)} - \zeta \mathbf{W}_{DGTDA}^{(p)}$, where ζ is defined as the largest singular value of $\left(\mathbf{W}_{DGTDA}^{(p)}\right)^{-1} \mathbf{B}_{DGTDA}^{(p)}$. However, in personal correspondence with Qun Li, first author of Li and Schonfeld [2014], the choice of ζ was stated to have little influence on the solution.

3.3.2 Direct optimisation of classification rate

Instead of optimising a measure of the class separation of the projections of observed data, an alternative is to directly optimise the classification rate. This was done by Dyrholm et al. [2007] by assuming a PARAFAC structure of data and substituting this for the standard linear combination of explanatory variables in logistic regression, where the objective function consists of maximising the log-likelihood. In the case of matrix observations, the log-likelihood becomes:

$$\begin{aligned} f_{BDCA}(w_0, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \; & \sum_{n=1}^{N} y_n (w_0 + \sum_{k=1}^{K} \mathcal{X}_n \times_1 \mathbf{U}_{:,k}^{(1)} \times_2 \mathbf{U}_{:,k}^{(2)}) - \\ & \log(1 + \exp(w_0 + \sum_{k=1}^{K} \mathcal{X}_n \times_1 \mathbf{U}_{:,k}^{(1)} \times_2 \mathbf{U}_{:,k}^{(2)})), \end{aligned} \tag{3.10}$$

where y_n denotes the class of observation n, which can take the values zero and one. The expected value of y_n, or equivalently, the probability of observation $\mathcal{X}_n$ belonging to class one, is then:

$$P(y_n = 1) = \frac{1}{1 + \exp(-w_0 - \sum_{k=1}^{K} \mathcal{X}_n \times_1 \mathbf{U}^{(1)}_{:,k} \times_2 \mathbf{U}^{(2)}_{:,k})}.$$

The parameters w_0, $\mathbf{U}^{(1)}$, and $\mathbf{U}^{(2)}$ can be estimated by an optimisation method. The Damped Newton method was suggested as a good choice and the gradient and Hessian given by Dyrholm et al. [2007]. This method is referred to as Bilinear Discriminant Component Analysis (BDCA) [Dyrholm et al., 2007].

Part III

Automatic classification of artefactual independent components

This part is about automatic classification of independent components (ICs) of EEG data. We developed a method for automatic classification of ICs of EEG data into multiple artefact classes, which is described in Chapter 4. Next, some studies in which we made use of the classifier are described. In Chapter 5, two studies of the effect of different types of artefacts on motor-imagery based BCIs are summarised. A study comparing the ability of six linear decomposition methods, including three variants of ICA, to separate neural from artefactual sources is described in Chapter 6. Finally, the work on IC classification is summarised and the performance of IC_MARC discussed.

Chapter 4

Classification of independent components into multiple artefact classes

This chapter is based on the article "Classification of independent components of EEG into multiple artifact classes", in which we investigated classification of independent components (ICs) of EEG into six different classes [Frølich et al., 2015a]. The post-print of this paper is included as Appendix A. We considered the following classes: neural components, blinks, lateral eye movements, the heart beat, and muscle artefacts, and components containing noise or several types of activity. The mixed IC class consisted of all ICs that had not been manually labeled. Visual inspection of random examples from this class showed that the unlabeled ICs consisted of noisy ICs, ICs representing loose electrodes, and ICs containing both neural and artefactual activity or several types of artefactual activity, etc. Examples of scalp maps of ICs from each of these classes are shown in Figure 4.1. We used logistic regression as the classification method, so the classifier calculates probabilities for each IC of belonging to each class.

Figure 4.1: Two examples of scalp maps for each of the six classes in each of the two studies. The figure is from [Frølich et al., 2015a].

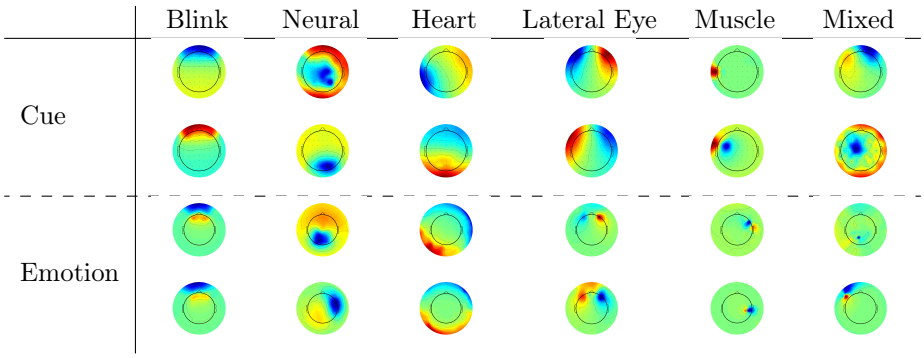


Table 4.1: Data summary, taken from [Frølich et al., 2015a]

Study	Subjects	Channels	Reference	ICA	Sampling rate	Duration	ICs
Emotion	34	250	Active	Infomax	256Hz	8-88 minutes	7255
Cue	12	64	Linked mastoids	Infomax	500Hz	56-66 minutes	768

4.1 Data

The classifier was trained and tested on two data sets, kindly provided by Scott Makeig, Julie Onton and Klaus Gramann [Onton and Makeig, 2009, Gramann et al., 2010], containing manually labelled ICs (summarised in Table 4.1).

4.2 Methods

We transformed time series and activation patterns of ICs into a common data space to minimise covariance shifts of features of ICs from different studies, including standardisation of patterns and time series to have mean zero and unit variance. All pre-processing steps are described in [Frølich et al., 2015a].

Features were selected from an initial pool of 65 spatial, spectral, and temporal features using two-layer cross-validation (CV) on the Emotion data. The outer CV consisted of leave-one-subject-out folds. Within each outer fold, features were selected using logistic regression with forward selection in five-fold CV.

4.3 Results

Figure 4.2 shows the number of outer CV folds that each feature was chosen. The two 14-feature sets (features chosen in at least 25 or 28 outer CV folds) are identical. This was chosen as the best feature set. Spatial features were chosen substantially more often than spectral and temporal features, which motivated us to inspect the classification performance when using only spatial features.

Figure 4.3 shows the confusion matrices obtained using only the spatial features from the initial feature pool in leave-one-subject-out CV on both studies and cross-study training and testing. Figure 4.4 shows the corresponding confusion matrices for the 14-feature set. There are some differences between the classification performances obtained with the two feature sets, but no feature set consistently outperforms the other. The neural class is well-classified by both feature sets. High performances on the binary problem of neural vs. non-neural IC classification has also been achieved by others [Winkler et al., 2011, Mognon et al., 2011, Viola et al., 2009]. The within-study, between-subject multi-class performances are also quite high for both feature sets, with balanced accuracy rates above 80%. However, when training on one data set and testing on the other, the multi-class performances decrease substantially. The cross-study performances are described in detail in the following.

The performance on blinks is worse when trained on Cue data and tested on Emotion data than vice versa. This could be due to the small number of blink ICs in the smaller Cue data set. When only spatial features are used, no blinks are classified correctly when trained on Cue and tested on Emotion data.

Using only spatial features, about 50% of the heart beat artefacts are classified correctly in both cross-study cases. When trained on the Cue data, they are confused with the muscle and mixed classes, while they are confused with the neural class when trained on the Emotion data. Heart beats are characterised by activity on the edges of the scalp map that changes smoothly across the scalp. The smooth change also characterises neural ICs while muscle ICs exhibit activity localised on the edges of the scalp. This may explain why heart beat ICs are confused with neural and muscle ICs when only spatial features are used. The large degree of confusion between the muscle and mixed classes also seems to spill over such that muscle-like heart beats are misclassified as mixed with the spatial feature set. With the 14-feature set, most (85%) heart beats are classified correctly while 15% are misclassified as lateral eye movements when trained on the Cue data. Another characteristic of heart beat artefacts is that they have opposing polarities on opposite sides of the scalp, which is also a defining characteristic of lateral eye movements. When trained on the Emotion data, all heart beat artefacts are misclassified as neural.

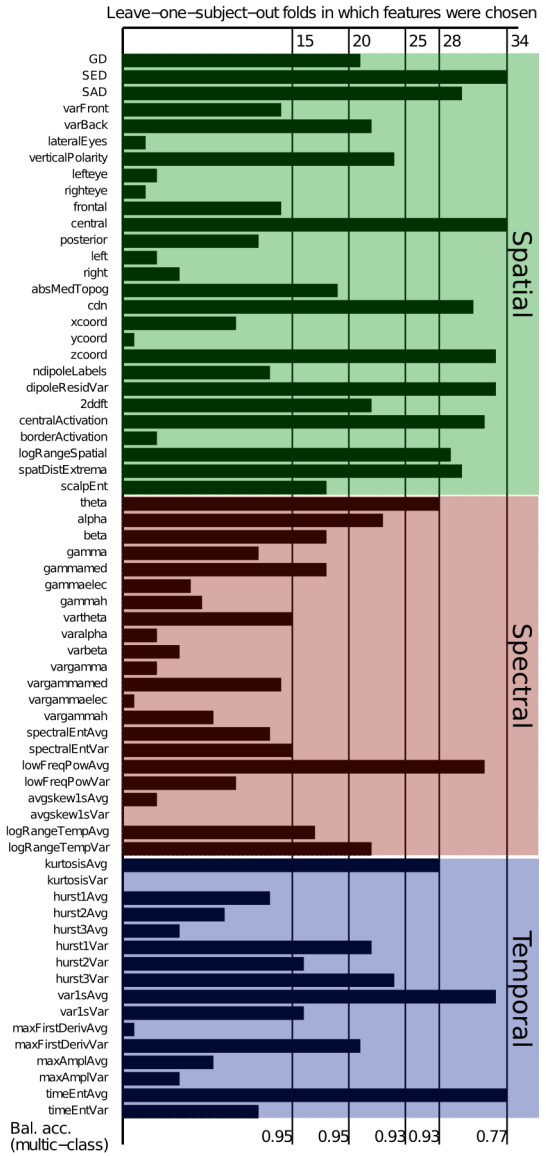


Figure 4.2: Barplot showing the number of folds each feature was chosen by forward selection in leave-one-subject-out CV on the Emotion data (34 subjects). Spatial features are shown first, followed by spectral and temporal features. The figure is from [Frølich et al., 2015a], with colors added to ease the distinction between feature types.

Lateral eye movements are classified quite accurately (between 78% and 100%) for both cross-study cases and feature sets.

The muscle class is classified well when trained on Emotion data, but confused with the mixed class when trained on Cue data with either feature set.

Mixed ICs tend to be confused with neural and muscle ICs. Since the labeling of ICs was performed as a pre-processing step in the studies described by Onton and Makeig [2009], Gramann et al. [2010] before analysing neural components, it is possible that some artefactual ICs were not labeled. These would then be treated as mixed. Hence the confusion between the mixed class and the other classes is not necessarily due solely to misclassifications. On the other hand, the confusion between the neural and mixed classes is more worrisome.

4.4 EEGLab plug-in: IC_MARC

An EEGLab plug-in, IC Classification into Multiple ARtefact Classes (IC_MARC), was built to make the classification method more accessible. The plug-in can be applied to an EEGLab data set with an ICA decomposition and channel locations. The feature set to use for classification can be chosen from the following three: 1) the 14-feature set described in the article which includes spatial, temporal, and spectral features, 2) a spatial feature set based on the spatial features from the 14-feature but further optimised heuristically, and 3) a heuristically optimised spatial feature set without dipole features, which are computationally demanding to calculate. After classification, the components' scalp maps can be visualised along with their classifications. By clicking on the assigned class label, a spectrogram and an ERP image of the component's activation over trials, if the EEGLab data set has a trial structure, are shown. The assigned class can be changed in this window, and the component can be marked for rejection.

4.5 Applications of classifier

The option to automatically remove only some artefact types opens up possibilities for diverse types of analyses. We have, for example, used the classifier to investigate the effects of various artefact groups on Brain-Computer Interfaces (see Chapter 5). An unexplored application is the use of time series of automatically identified eye blink ICs to detect eye blinks. This could be used to ensure that blinks are not time-locked with e.g. a stimulus. Additionally, the time series of ocular components could be used to detect drowsiness.

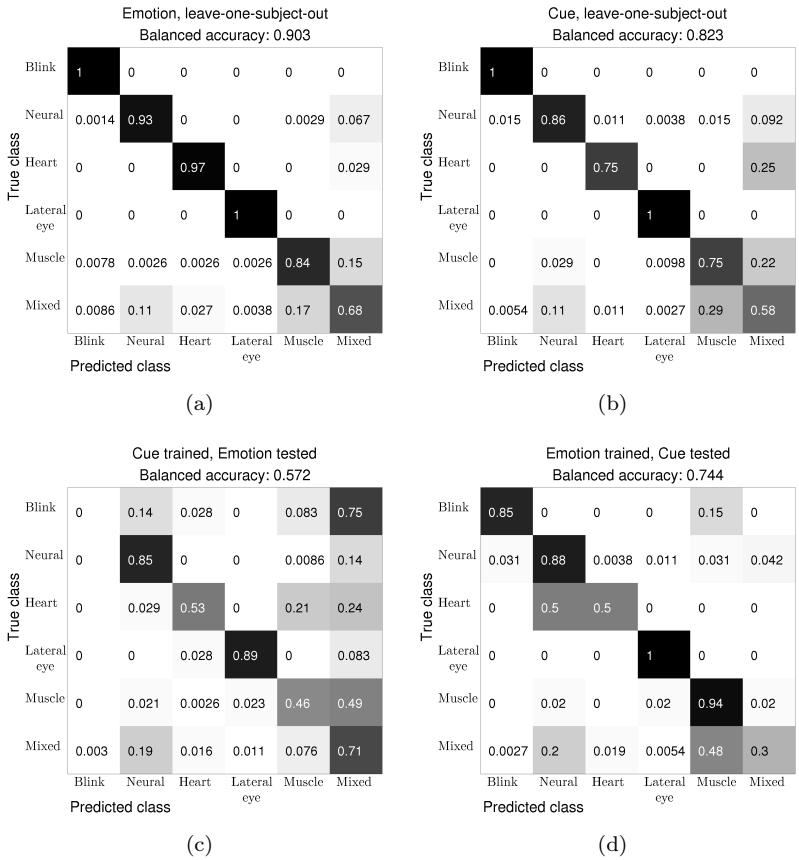


Figure 4.3: Confusion matrices showing classification performances with the spatial features from the initial pool of features in leave-one-subject-out CV on Emotion data (a), leave-one-subject-out CV on Cue data (b), training on Cue and testing on Emotion data (c), and training on Emotion and testing on Cue data (d). Black corresponds to higher and white to lower values. The figure is from [Frølich et al., 2015a].

4.6 Thoughts on possible improvements

An obvious way to improve the classifier would be to train the classifier on more data. This retraining could include another analysis of which features are best, but the classifier could also just be trained with the same features on more data. Optimally, several datasets containing ICs and their corresponding labels would be available. A leave-one-dataset-out CV could then be performed to assess the

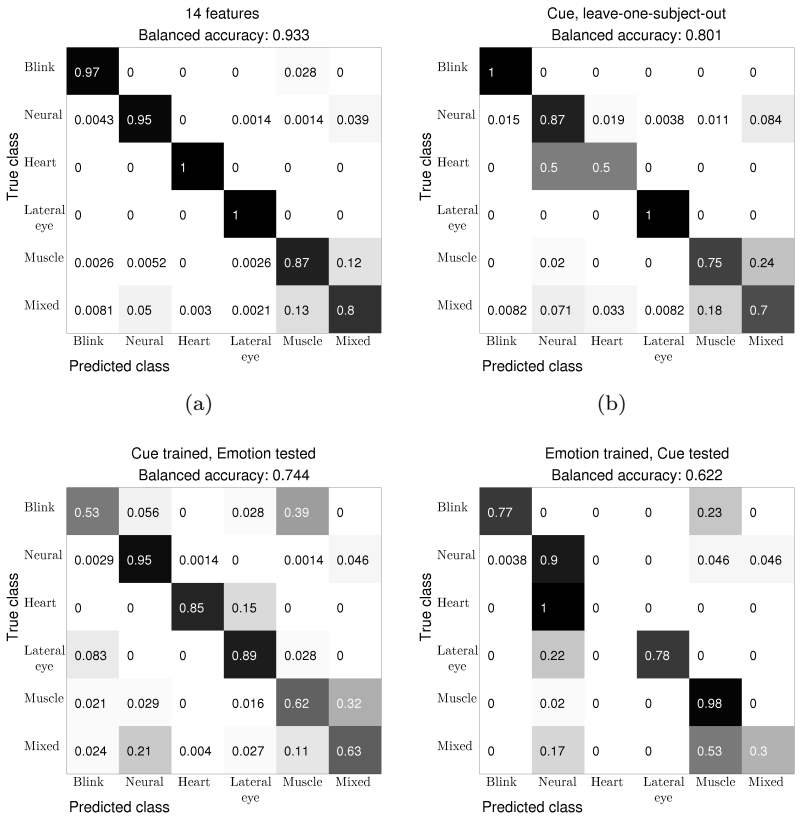


Figure 4.4: Confusion matrices showing classification performances with the 14-feature set in leave-one-subject-out CV on Emotion data (a) leave-one-subject-out CV on Cue data (b), training on Cue and testing on Emotion data (c), and training on Emotion and testing on Cue data (d). The balanced accuracies in the subtitles are the multi-class balanced accuracies. Black corresponds to higher and white to lower values. The figure is from [Frølich et al., 2015a].

generalisability and a final classifier could be trained on all datasets.

The EEGLab plug-in IC_MARC could be personalised by employing a user's changes to automatic classifications. Since the classifier gives probabilities of belonging to each class, such changes could be used to alter the thresholds at which ICs are classified. If, for example, a user often changes "mixed" classifications, the threshold necessary to classify an IC as "mixed" could be increased.

By normalising time series and scalp maps to have variance one, information on the total energy of ICs is lost. Normalising only either the scalp map or time series and using the variance of the unnormalised quantity as a feature might improve classification performance. In an undocumented analysis we included this feature after having completed all other feature selection. This did not improve the performance, so we did not re-run the feature selection analyses. However, if all analyses were to be performed again, such a feature could be included.

The EEGLab plug-in allows the user to label an IC as a loose electrode, even though this class is not known by the classifier. The plug-in could be improved by allowing users to add other such classes.

Since the spatial characteristics of heart beats are shared by several other classes, more temporal or spectral features would probably help to disentangle heart beats from the other classes. The reason that these features were not chosen in the automatic feature selection was probably due to the scarcity of available heart beat ICs. Even though observations were weighted to account for such class imbalance, the very small number of heart beat ICs might make it difficult to for the classifier to learn the temporal and spectral characteristics shared by heart beat ICs in general.

Sources extracted by other methods than ICA might vary on other parameters than ICs. For example, SSD and Fourier-ICA focus on extracting oscillatory components. Classification of such components might benefit from spectral features. Hence, the classifier could be made more general by training on sources extracted by several different methods.

Finally, it is easy to get access to many ICs while labels of ICs are more scarse. Hence semi-supervised learning methods might be suitable for this classification problem.

Chapter 5

Effects of artefacts on Brain-Computer Interfaces

In this chapter two papers are summarised in which we used the IC-classifier, IC_MARC, to investigate the effects of artefacts on motor-imagery based BCI systems.

5.1 Investigating effects of different artefact types on motor imagery BCI

This section is based on the paper "Investigating effects of different artifact types on Motor Imagery BCI", in which we used the classifier IC_MARC to distinguish between different types of artefactual ICs to investigate effects of different artefacts on motor imagery based Brain-Computer Interfaces (BCIs) [Frølich et al., 2015b]. The post-print of the paper is included as Appendix B.

5.1.1 Data

We used data from 80 BCI-novices performing motor imagery, described in [Blankertz et al., 2010]. Data were recorded at 1000 Hz from 119 electrodes placed according to the extended 10-20 system. We band-pass filtered data between 8-30 Hz and defined epochs as 0.75-3.5 s after event markers. Channels with excessively low or high variance in training data were automatically rejected.

5.1.2 Methods

We used CSP to find six spatial filters, three that maximised the ratio of variance for class one relative to class two and three for the opposite case. The filter matrices were regularised against artefactual directions as described below. We used Linear Discriminant Analysis for classification. The covariance estimates for each class were obtained as Euclidean averages of individual trial covariances.

For each subject, the following steps were taken. The best regularisation parameter for each regularisation method was found through 10-fold cross-validation (CV). This parameter was used to train the classifier on all CV folds, which was evaluated on test data. All analyses were performed both using all the channels available and using only 48 central channels previously reported to obtain a good performance Sannelli et al. [2010]. The final results are based on averages and tests of the results from each subject.

5.1.2.1 Independent components

For each subject, we ran ICA on the concatenated training data epochs using the Extended Infomax algorithm in EEGLab [Delorme and Makeig, 2004]. In the pre-processing step of Extended Infomax, we retained the first principal components that explained 99.9% of data variance. Hence the extracted ICs explain 99.9% of data variance. ICs were classified as belonging to the class for which the highest probability was predicted, except if the highest probability was for an ocular artefact class and that probability was less than 80%. Such ICs were classified as mixed. Figure 5.1 shows patterns from ICs classified by IC_MARC.[1] The figure also shows that muscle-artefact contamination is strongest on scalp edges.

[1]Except for the heartbeat class, the examples are good demonstrations of what one would expect in each class. Difficulty with the heart beat class was also found during the development of IC_MARC and CORRMAP Frølich et al. [2015a], Viola et al. [2009].

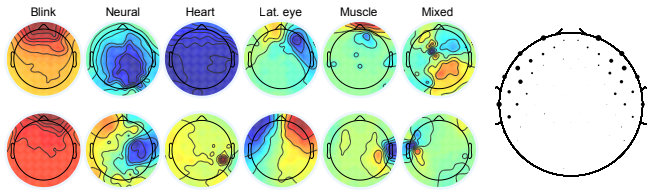


Figure 5.1: Left: Examples of patterns of automatically classified ICs. Right: Locations of most active electrode in muscle ICs from all subjects. Dot sizes represent the number of times electrodes were the most active in muscle ICs. The figure is from Frølich et al. [2015b].

5.1.2.2 Investigating artefacts' effects on BCI performance

Regularisation using ICs We considered several approaches to regularising the filter matrices by utilising the probabilities of class membership assigned to ICs by IC_MARC. In Appendix H, the regularisation methods initially considered are described. Since none resulted in substantial BCI performance improvements relative to regularisation with the identity matrix in initial analyses, we focused on a relatively simple regularisation approach. This method consisted of collecting patterns of artefactual ICs, normalised to have 2-norm one, as columns in a matrix, $\mathbf{A}_{art}$ and setting the regularisation matrix to $\mathbf{A}_{art}\mathbf{A}_{art}^{\top}$. This discourages CSP from finding filters that extract artefactual activity since this regularisation causes $\|\boldsymbol{w}^{\top}\mathbf{A}_{art}\|$ to be minimal implying that

$$\boldsymbol{w}^{\top}\mathbf{X} = \boldsymbol{w}^{\top}(\mathbf{AS}) = \boldsymbol{w}^{\top}(\mathbf{A}_{art}\mathbf{S}_{art} + \mathbf{A}_{neuro}\mathbf{S}_{neuro}) \approx \boldsymbol{w}^{\top}(\mathbf{A}_{neuro}\mathbf{S}_{neuro}),$$

where $\mathbf{S}_{art}$ denotes the matrix containing artefactual sources' time series and the matrices $\mathbf{A}_{neuro}$ and $\mathbf{S}_{neuro}$ contain the patterns and time series of non-artefactual ICs, respectively.

Removal of ICs We also looked at the BCI performance when training on only neural ICs (removing all ICs from each artefact group in turn) and when training on only the artefactual ICs as defined by each artefact group.

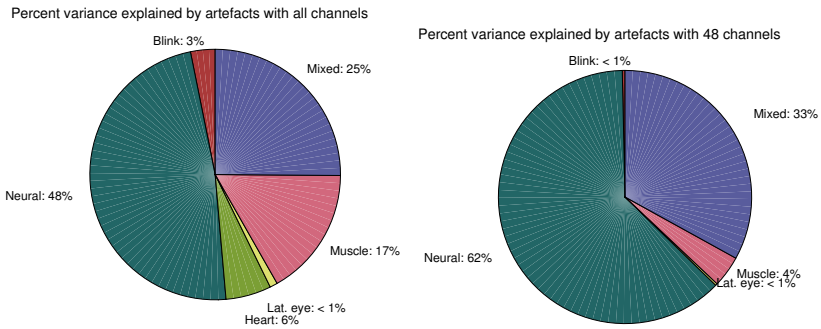


Figure 5.2: Percent variance of data explained by the various artefact groups for the all-channel electrode configuration (left) and the 48 centrally placed channels (right). The figure is from the oral presentation of the work at EMBC2015.

5.1.3 Results

Figure 5.2 shows the percent variance accounted for by each artefact class for both the all-channel and the 48-channel configurations. We see that eliminating the outer electrodes, which were in the all-channel but not the 48-channel configuration, substantially reduces the proportion of muscle artefacts in data and increases the proportion of neural data.

When training on artefacts, classification performances were significantly different from chance, but only a few percentage points above chance level. Hence artefacts seem to contain some class-discriminative information, but not enough to support the higher classification rates achieved when using all data.

With the all-channel configuration, some average (over subjects) performance improvement was obtained when regularising against muscle ICs, but not when removing or regularising against ocular or all non-neural ICs. No improvement was seen on the 48-channel configuration.

5.2 Brain-Computer Interfacing under Distraction: An Evaluation Study

This section is based on the paper "Brain-Computer Interfacing under Distraction: An Evaluation Study" [Brandl et al., 2016]. A pre-print of this paper is included in Appendix C. In this paper, we looked at the effects of various dis-

traction tasks while subjects performed a motor imagery task. The distractions simulating out-of-the-lab environments were:

- *Clean* and *Calibration*: No distraction.
- *Eyes closed*. Subjects had their eyes closed.
- *News*. Subjects attended to current news and news from 1994.
- *Numbers*. Sheets of paper with number-letter combinations were hung from the walls of the room in which the experiment was conducted. The subject had to locate a specific combination in each trial, requiring head movement.
- *Flicker*. Flicker in grey shades shown at 10 Hz on the computer screen.
- *Stimulation*. Vibratory stimulation at carrier frequencies of 50 and 100 Hz, each modulated at 9, 10, and 11 Hz.

The BCI classifier was trained on calibration data and tested on each distraction.

I was responsible for analyses of artefacts' effects (sections IIIe and IVc in the article), which are summarised here. For motivation for the distraction tasks, details on the experiment, and other analyses of these data, see [Brandl et al., 2016].

Artefact types were grouped into the five following groups:

- Muscular (muscle artefacts).
- Ocular (blinks and lateral eye movements).
- Non-neural (blinks, heart beats, lateral eye movements, muscle, and mixed artefacts).
- Muscular and mixed (muscle and mixed artefacts).
- Non-mixed artefacts (blinks, heart beats, lateral eye movements, and muscle artefacts).

Examples of scalp maps of ICs learned from the clean data and classified by IC_MARC are shown in Figure 5.3, where the groupings are also indicated. Except for the upper muscle artefact (which resembles a lateral eye movement), these scalp maps match our expectations of the classes they were classified to.

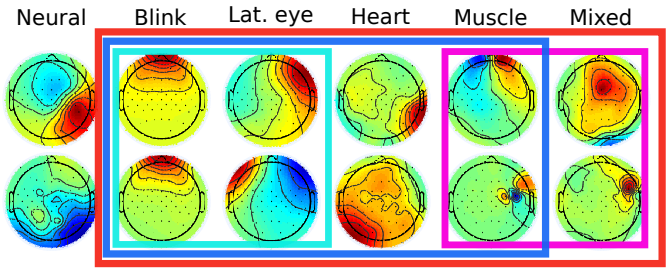


Figure 5.3: Examples of scalp maps of ICs classified by IC_MARC. We used the following groups of artefacts: muscular artefacts, ocular artefacts (cyan box), non-neural components (red box), muscle and mixed artefacts (purple box), and non-mixed artefacts (blue box). The figure is modified from [Brandl et al., 2016].

To clean data, the Extended Infomax ICA algorithm as implemented in EEGLab [Delorme and Makeig, 2004] was used to decompose calibration data. These ICs were classified by IC_MARC. ICs from each group then were removed in turn, both from the calibration data used to train the BCI classifier and from each distraction task's data, which the BCI classifier was tested on.

5.2.1 Properties of artefacts

To inspect differences between distractions, we quantified the proportion of data variance explained by each group of artefactual ICs for each distraction task. The means over subjects and their standard deviations are shown in Figure 5.4. The *Numbers* distraction stands out from the other distractions, including the calibration data, as having a substantially different artefact distribution. This may help explain why this task needed a different classifier from that used for the other distractions to reach acceptable performance levels [Brandl et al., 2016].

We also inspected the spectrum of each IC class, shown in Figure 5.5. This figure is in good accordance with expectations of the behaviour of the different IC types. For example, the neural ICs have highest power in the 8-13 Hz range, which is the frequency range that the investigated motor imagery task influences. Likewise, muscular ICs exhibit most power at high frequencies, which muscular artefacts are known to do [Muthukumaraswamy, 2013].

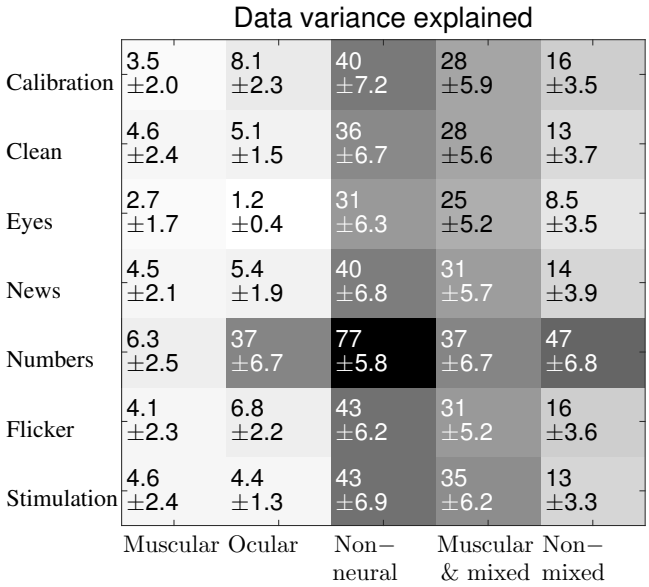


Figure 5.4: Percent data variance explained by each artefact group in each condition. The numbers give the mean over subjects plus/minus its standard deviation. The figure is from Brandl et al. [2016].

5.2.2 Classification after artefact removal

From Table 5.1, we see that removing artefacts does not substantially increase the classification rate for any distraction task/artefact group combination and statistical tests did not detect significant differences for any task when artefacts were removed. However, some improvement is obtained consistently for all artefact groups in the *News* distraction. This is in accordance with previous similar studies [Frølich et al., 2015b, Winkler et al., 2011].

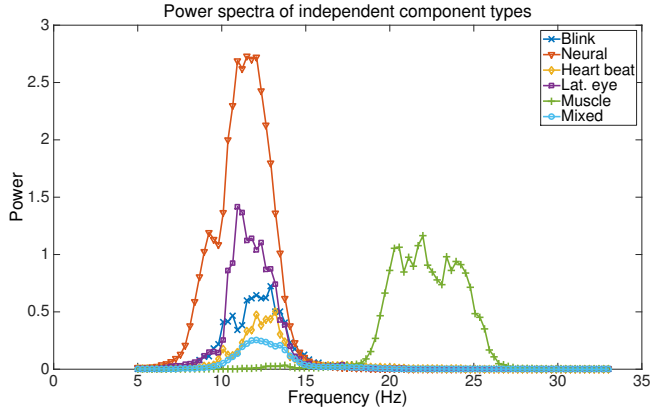


Figure 5.5: Power spectra of ICs from the six classes (blinks, neural, heart beats, lateral eye, muscle, and mixed). The power spectra were calculated for each epoch independently. Then the median was first taken over epochs for each subject, and then over subjects. The figure is from [Brandl et al., 2016].

Table 5.1: Mean classification accuracies for all distractions and removed artifact groups. For each experiment, the artifact group with highest (**bold**) and lowest (red) performance rates and performances better than baseline (blue) are highlighted. Lowest scores better than baseline are purple. The rightmost column is given for baseline comparison. The table is from [Brandl et al., 2016].

	Muscular	Ocular	Non-neural	Muscular and mixed	Non-mixed artefacts	Overall	Baseline
Overall	62.30	**62.69**	62.20	61.83	62.01	62.20	62.39
Clean	66.46	65.67	**66.97**	65.94	64.72	65.95	66.68
Eyesclosed	62.50	**63.80**	60.75	60.68	63.19	62.19	63.10
News	62.81	63.76	63.51	**63.94**	63.42	63.49	62.64
Numbers	53.56	55.29	**56.86**	54.69	55.12	55.10	53.81
Flicker	**67.10**	66.58	64.93	65.28	65.10	65.80	66.41
Stimulation	**61.35**	61.01	60.14	60.49	60.48	60.69	61.53

Chapter 6

Artefact removal using linear decompositions of EEG

In the article "Removal of muscular artifacts in EEG signals: A comparison of ICA and other linear decomposition methods", we compared several decomposition methods on their ability to separate artefactual from neural sources [Frølich and Winkler, 2016]. The aim was to remove a movement artefact while retaining neural activity. A pre-print of the article is included in Appendix D.

6.1 Data

We used data recorded from 18 subjects performing self-paced braking at approximate 1 s intervals for five minutes in a simulated driving experiment. It is known that motor preparation and execution induces Event-Related Desynchronisation (ERD) in the α (8-13 Hz) and β (15-30 Hz) bands [Neuper and Pfurtscheller, 2001]. However, we also observed a peak in the power at time zero, when time-locking to the brake press, in both bands. This is likely due to the subjects moving their heads when they press the brake. Since the phenom-

ena are well understood and the movement artefact is very prominent in the raw data, this data set is well suited to evaluating removal of this artefact.

6.2 Methods

We compared three variants of ICA (Extended Infomax, FastICA, and TDSEP), PARAFAC2, FourierICA, and SSD. Since FourierICA and SSD are guided toward frequencies of interest we gave more equal terms to all methods by running the ICA methods and PARAFAC2 both on the raw data subjected only to standard pre-processing and on the same data after it had been high-pass filtered with cut-off frequency just below the frequency range of interest. This was also motivated by previous work in which high-pass filtering at a high cut-off frequency improved separation of neural and artefactual sources [Winkler et al., 2015].

6.3 Results

Figure 6.1 shows grand averages of data time-locked to the EMG peak activity for the uncleaned data and data cleaned by each method, with and without high-pass filtering for the ICA methods and PARAFAC2. The intervals between -500ms, -300ms, -50ms, 50ms, and 300ms are emphasised by light and dark shades of grey. The top of Figure 6.1 shows results for the α band while the bottom part relates to the β band. Successful cleaning would result in the ERD (dip of the black line before time zero) being maintained throughout the movement, until 300ms. The numbers in the legend refer to a heuristic quantification of the ERD quality, described in [Frølich and Winkler, 2016]. The lower this measure is, the better the ERD is maintained during the brake press.

From the uncleaned data (solid black line), we see that the artefact is most prominent in the β band. However, the cleaning is also more effective in the β band for all methods. High-pass filtering improves the cleaning obtained by all the blind methods. While high-pass filtered Extended Infomax obtains the best performance for both bands, the performance of high-pass filtered FastICA is close to this performance in both bands, while FourierICA is also very close in the β band. From analyses of the ERD quality measures resulting from retaining between one and all components, we saw that these results are quite robust, and that there is no consistent difference between the methods when looking at all numbers of retained components [Frølich and Winkler, 2016]. We also saw that

the scalp maps of the cleaned data look reasonable in [Frølich and Winkler, 2016]. Surprisingly, PARAFAC2 consistenly had the worst performance.

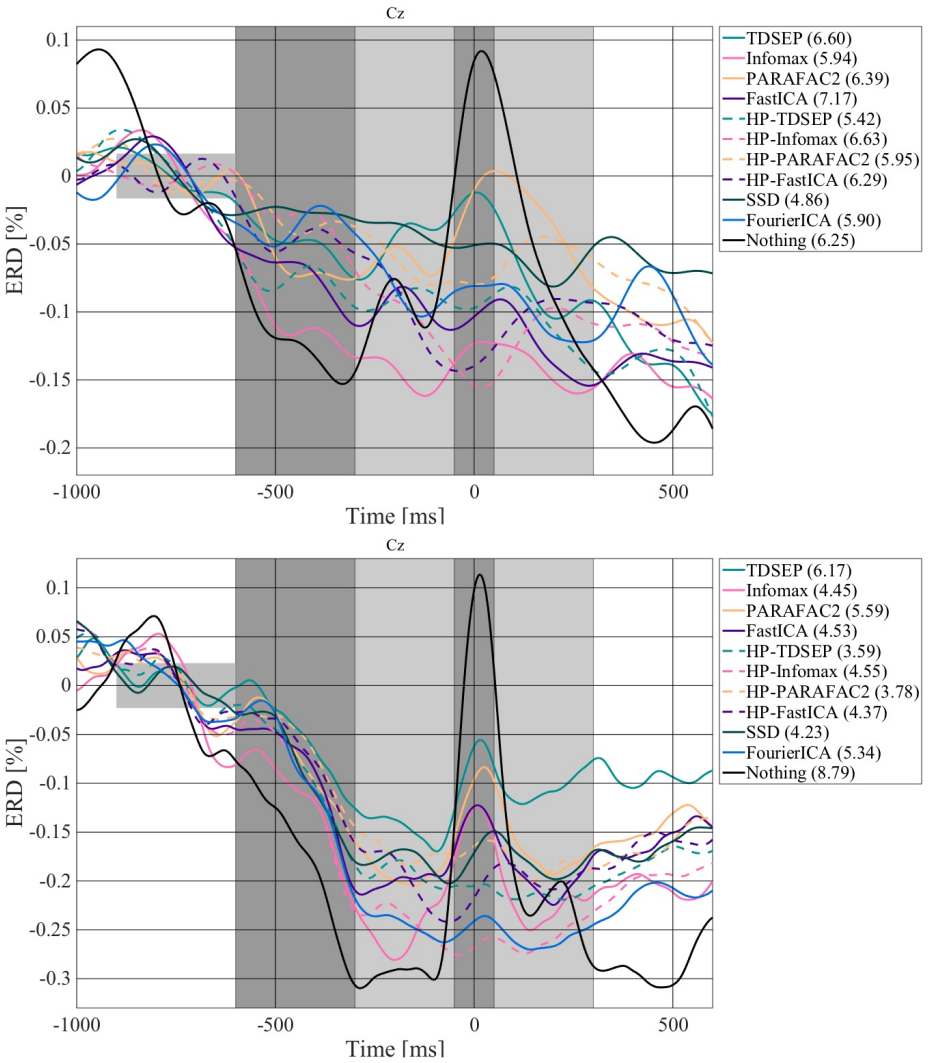


Figure 6.1: Grand-average ERD for 18 subjects recorded during self-paced foot movements in the alpha (7-14 Hz) (top) and beta band (15-30 Hz) (bottom), aligned to EMG peak activity. Time courses of data reconstructed from neural ICs (for SSD with the ten components with highest SNR) at channel Cz. The legend contains the ERD quality measure for each method, lower is better.

Chapter 7

Conclusion

While the developed IC classifier, IC_MARC, obtained high performances in the neural vs. artefact problem both within and across the two labeled studies, it had problems distinguishing between the various artefact classes in cross-study evaluation of the two studies with labeled ICs [Frølich et al., 2015a].

The high performance on the neural vs. artefact problem allowed us to have confidence in comparisons of decomposition methods on their ability to separate artefactual from neural sources, with the sources automatically labeled by IC_MARC [Frølich and Winkler, 2016]. Since several methods were applied on data from each of 18 subjects, a high number (≈11000) of sources needed to be classified. The automatic classification enabled us to perform analyses that would not have been practical with manual classification. We found that all methods' cleaning results improved if data were high-pass filtered at a high cut-off frequency before decomposing data. With this high-pass filtering, all three variants of ICA (FastICA, Extended Infomax, and TDSEP) had similar performances, which was also similar to FourierICA. Although SSD did not perform as well, its much lower running time might make it a good compromise between the quality of data cleaning and the time it takes to clean data.

In [Frølich et al., 2015b] and [Brandl et al., 2016], we used IC_MARC to investigate relations between artefactual contamination of data and performances of motor imagery based BCIs. We were unable to obtain notable BCI performance

improvements in both of these attempts. However, the inspection of artefact distributions could help explain why one distraction in [Brandl et al., 2016] requires a different classifier than the other distractions. The inability to improve the BCI performance by minimising or removing the use of artefactual directions in BCI classification could be due to wrong classifications. However, we know that, at least for the neural vs. artefact problem, IC_MARC has performed well on several other data sets. Another explanation could be that subjects accidentally use artefacts to control the BCI system, such that it is in fact controlled through muscle control and not EEG. In that case, we would expect to see performance decreases when removing artefacts. However, we did not observe this. Another reason could be that artefacts are nearly equally distributed across the motor imagery classes in the training data. If that is the case, then the BCI classifier would disregard the artefactual directions, implying that the removal of the artefacts would not make a difference.

Since the ability of IC_MARC to distinguish between different artefactual classes is not perfect, we inspected a random selection of classified ICs in [Frølich et al., 2015b] and [Brandl et al., 2016]. Although some ICs were misclassified, the majority of the inspected classifications seemed sound. In our analyses, we combined the different artefact classes in larger groups. Mis-classifications between groups should be less likely since several groups encompass classes that tend to be confused with each other (e.g. the mixed and muscle).

While removing artefacts did not improve BCI performance, it did not cause substantial performance decreases either. This is a sign that neural ICs relevant for class-discrimination are not misclassified by IC_MARC to any great extent. Furthermore, Figure 5.5 shows that the peaks in power of neural and muscle ICs are as expected. Since the ICs were classified using a spatial feature set, this also indicates that IC_MARC produces reasonable classifications. Additionally, our finding that the artefact distribution for the *Numbers* distraction differs from the other distractions is well in line with the finding that one classifier works well under all distractions except the *Numbers* distraction [Brandl et al., 2016]. This is furhter evidence that IC_MARC performs reasonably .

In our analyses relying on classified ICs, we only ran one ICA for each data set. While some ICs may be unstable due to local minima of the ICA solution, this should not be a problem for our analyses since we did not analyse single ICs in detail, but rather looked at group effects. If a few ICs in a group are unstable to the point of changing groups, it should not affect the group effect to a large extent since most ICs in the group are likely to be stable [Duann et al., 2006].

Part IV

Supervised tensor methods

Chapter 8

Multi-way Strategies for Single-trial Classification of Electroencephalography Data

In the paper "Multi-linear Discriminant Analysis with Tucker and PARAFAC Structures optimized on the Stiefel Manifold", we used multi-linear methods to classify EEG data on the single-trial level [Frølich et al., 2016]. A pre-print of the article is included as Appendix E. We compared classification performances from the following approaches: 1) feature extraction using four existing unsupervised decomposition methods (PARAFAC [Harshman, 1970, Carroll and Chang, 1970], Tucker [Tucker, 1966], Tucker2 [Tucker, 1966], and PARAFAC2 [Harshman, 1972, Kiers et al., 1999]) followed by logistic regression, 2) feature extraction using four new and four existing Multi-linear Discriminant Analysis (MDA) methods (Constrained MDA (CMDA) [Li and Schonfeld, 2014], Discriminant Analysis with TEnsor Representation (DATER) [Yan et al., 2005], Bilinear Discriminant Analysis [Visani et al., 2005], and Direct General Tensor Discriminant Analysis (DGTDA) [Li and Schonfeld, 2014]) followed by logistic regression, and 3) combined feature extraction and classification in a logistic regression framework using one new and one existing method (Bi-linear Discriminant Component Analysis (BDCA) [Dyrholm et al., 2007]).

We compared the methods on two data sets. One data set contained 16 subjects and has previously been analysed as Experiment 2 in [Stekelenburg and Vroomen, 2007]. We used control trials (no sound, gray box on screen) and non-speech auditory-only trials (clapping (103–107 ms) and tapping of spoon on cup (292–305 ms), gray box on screen). We balanced the data so that there were equally many (1302) trials of each type after rejecting trials with values greater than 150 μV or less than -150 μV. We used trials starting at stimulus onset and lasting until 0.5 s after the time-locking. This time interval was chosen since this is where differences between condition were seen in difference waves in Figure 2 in [Stekelenburg and Vroomen, 2007]. We selected channels that were common to all subjects (50 channels) in order to perform leave-one-subject-out cross-validation (CV). Since these analyses were performed as leave-one-subject-out CV, the classification results constitute an estimate of the across-subject generalisation of the classification methods.

The other data set was Dataset II [Schalk et al., 2004] from BCI Competition III [Blankertz et al., 2006]. This data set consisted of data from two subjects in a P300 speller paradigm. Each subject had a training data set with 85 characters and a test data set with 100 characters. There were 180 trials for each character. We extracted epochs from stimulus onset until 667 ms after onset. For each subject, we performed 5-fold CV on the training data to determine the best component number for each method. Each method was then trained on all the training data for each subject. Using the single-trial classifications of the test data (for which single-trial labels were not available) letters were predicted and compared to the correct letters.

8.1 Methods

8.1.1 Unsupervised feature extraction

One way to use multi-linear methods for classification is to extract features from data through a multi-linear decomposition and then apply a supervised classifier to those features in a following step. We did this for the Tucker, PARAFAC, and PARAFAC2 models by using estimated trial strengths as features. To avoid degenerate solutions [Stegeman, 2007], we imposed an orthogonal constraint on PARAFAC and PARAFAC2 in the trial mode. We chose to constrain the trial mode to ensure as large a span of data as possible in this mode since this is the mode used in the following classification step. Since these methods are unsupervised, all data can be used in the decomposition without being influenced by the true labels of the test data. To get estimated trial strengths

for all observations, it is necessary to use all data in the decomposition such that the trial-mode factor is estimated for all observations.

We also used the Tucker2 model to obtain projection matrices for the spatial and temporal modes. Using these matrices, all trials were then projected into a lower-dimensional core array representation and the core arrays were vectorised and used as features. It is not necessary to use the test data in the Tucker2 pipeline since the projection matrices estimated from training data can be applied to test data later. However, all data may be used to estimate the Tucker2 decomposition since it is unsupervised, and we did this in our analyses. The *nway* toolbox [Andersson and Bro, 2000] was used to optimise the four unsupervised methods.

8.1.2 Supervised feature extraction

Analogously to classification based on Tucker2 core arrays, the MDA methods also find projection matrices for trial-wise projection. Contrary to Tucker2, these methods are supervised and aim to find projection matrices that project trials into a lower-dimensional space that is maximally discriminative. As for Tucker2, these projection matrices are then used to obtain lower-dimensional core array representations of all trials, which are used for classification in their vectorised forms. Since the MDA methods are supervised, only the training data can be used to find the projection matrices.

In the following, we refer to the method proposed by Visani et al. [2005] as DATEReig since it solves the standard eigenvalue problem corresponding to the generalised eigenvalue problem solved by DATER, as explained in Section 3.3. While CMDA, DATER, DATEReig, and DGTDA optimise their objective functions with heuristic algorithms, we propose to perform the optimisation rigourously on a manifold using the derivatives of the objective functions. To avoid redundancy between the projection factors, we require them to be orthogonal as is also done in CMDA, DATEReig and DGTDA. Since the Stiefel manifold contains all orthonormal matrices, we want each projection matrix to lie on the Stiefel manifold corresponding to its dimensions. We ensure this is the case by optimising over the product of Stiefel manifolds of dimensions corresponding to those of the projection matrices. We used the Conjugate Gradient method in the *ManOpt* toolbox [Boumal et al., 2014] for Matlab to perform the optimisation. The four new methods correspond to four objective functions, each optimised as described above. The derivatives necessary for rigorous optimisation of our suggested objective functions are given in Appendix G. The four objective functions consist of a PARAFAC- and a Tucker-structure version of the scatter ratio LDA objective function (3.7) and the trace of matrix ratio objective (3.3).

We compare the scatter ratio and trace of matrix ratio objective formulations since these have been used in discriminant analysis previously [Bishop, 2006, Li and Schonfeld, 2014, Yan et al., 2005]. We propose Tucker-structure versions to make the models flexible and thus able to model complex interactions in data. However, such complex models might not be necessary. In such cases, the PARAFAC structure is suitable. Since the Tucker structure is invariant to rotations of the projection matrices and allows interactions between all factors across modes, it is difficult to interpret. The PARAFAC-structure only allows each factor in a mode to interact with one factor from each of the other modes, making it easier to interpret. Hence models with the PARAFAC-structure are appealing if they model data adequately. The formal definitions of the proposed objective functions are given in the following.

Manifold Tucker Discriminant Analysis with the scatter ratio objective function (ManTDA_sr):

$$\frac{\sum_{c=1}^{C} N_c \|(\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{p=1}^{P} \mathbf{U}^{(p)}\|_F^2}{\sum_{c=1}^{C} \sum_{\{n:class(\mathcal{X}_n)=c\}} \|(\mathcal{X}_n - \bar{\mathcal{X}}_c) \times_{p=1}^{P} \mathbf{U}^{(p)}\|_F^2} \tag{8.1}$$

Manifold PARAFAC Discriminant Analysis with the scatter ratio objective function (ManPDA_sr):

$$\frac{\sum_{c=1}^{C} N_c \|\mathcal{B}_{c,PARAFAC}\|_F^2}{\sum_{c=1}^{C} \sum_{\{n:class(\mathcal{X}_n)=c\}} \|\mathcal{W}_{n,PARAFAC}\|_F^2}, \tag{8.2}$$

where $\mathcal{W}_{n,PARAFAC}$ and $\mathcal{B}_{n,PARAFAC}$ are diagonal core arrays:

$$\begin{aligned}
\mathcal{W}_{n,PARAFAC_{k,k,\ldots,k}} &= (\mathcal{X}_n - \bar{\mathcal{X}}_{c:class(\mathcal{X}_n)=c}) \times_{p=1}^{P} \mathbf{U}_{:,k}^{(p)} \quad k \in \{1, 2, \ldots, K\} \\
\mathcal{B}_{c,PARAFAC_{k,k,\ldots,k}} &= (\bar{\mathcal{X}}_c - \bar{\mathcal{X}}) \times_{p=1}^{P} \mathbf{U}_{:,k}^{(p)} \quad k \in \{1, 2, \ldots, K\}.
\end{aligned}$$

Manifold Tucker Discriminant Analysis with the trace of matrix ratio objective (Man_TDA):

$$
\begin{aligned}
&Tr\Bigg[\Bigg(\sum_{c=1}^{C}\sum_{\{n:class(\mathcal{X}_n)=c\}} vec\Big((\mathcal{X}_n-\bar{\mathcal{X}}_c)\times_{p=1}^{P}\mathbf{U}^{(p)}\Big)\\
&vec\Big((\mathcal{X}_n-\bar{\mathcal{X}}_c)\times_{p=1}^{P}\mathbf{U}^{(p)}\Big)^{\top}\Bigg)^{-1}\cdots\\
&\Bigg(\sum_{c=1}^{C}N_c vec\Big((\bar{\mathcal{X}}_c-\bar{\mathcal{X}})\prod_{p=1}^{P}\times_p\mathbf{U}^{(p)}\Big)\\
&vec\Big((\bar{\mathcal{X}}_c-\bar{\mathcal{X}})\times_{p=1}^{P}\mathbf{U}^{(p)}\Big)^{\top}\Bigg)\Bigg].
\end{aligned}
\tag{8.3}
$$

Manifold PARAFAC Discriminant Analysis with the trace of matrix ratio objective (Man_PDA):

$$
\begin{aligned}
&Tr\Bigg[\Bigg(\sum_{c=1}^{C}\sum_{\{n:class(\mathcal{X}_n)=c\}} diag\Big(vec\Big(\mathcal{W}_{n,PARAFAC}\Big)vec\Big(\mathcal{W}_{n,PARAFAC}\Big)^{\top}\Big)\Bigg)^{-1}\\
&\Bigg(\sum_{c=1}^{C}N_c diag\Big(vec\Big(\mathcal{B}_{c,PARAFAC}\Big)vec\Big(\mathcal{B}_{c,PARAFAC}\Big)^{\top}\Big)\Bigg)\Bigg].
\end{aligned}
\tag{8.4}
$$

8.1.3 Feature extraction and classification in one step

By combining feature extraction and classification, the whole classification pipeline can be unified in the goal of optimising the final classification performance measure. Incorporating the PARAFAC structure in a logistic regression model was suggested in 2007, and was referred to as Bilinear Discriminant Component Analysis (BDCA). We suggest generalising this model to assuming the Tucker structure to allow more flexibility in modeling data, and refer to this method as $BDCA\_Tucker$. For simplicity, we assume that observations are matrices. We obtain the following log-likelihood for BDCA_Tucker:

$$
\sum_{n=1}^{N} y_n(w_0+\psi(\mathcal{X}_n))-\log(1+\exp(w_0+\psi(\mathcal{X}_n)), \tag{8.5}
$$

where $\psi(\mathcal{X}_n) = \sum_{k_1=1}^{K_1} \sum_{k_1=1}^{K_2} \mathbf{V}_{k_1,k_2} \mathcal{X}_n \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}$ and $\mathbf{V}_{k_1,k_2} = \delta_{k_1,k_2}$ to disambiguate scaling of the projection matrices $\mathbf{U}^{(p)}$ and matrix of interaction coefficients $\mathbf{V}$. As in BDCA, there are no constraints on the projection matrices with this method.

8.2 Results

Figure 8.1 shows the Area Under ROC Curve scores (AUCs) obtained when the classifiers were evaluated on training- and test data from the Stekelenburg&Vroomen data set.

When evaluated on training data (top of Figure 8.1), the ManTDA and BDCA methods obtained the highest performances. That ManTDA provides a better fit to training data was also reflected in substantially higher objective function values for this method than for DATER, DATEReig and CMDA [Frølich et al., 2016]. However, this better fit to training data can result in an overfitted model, which was the case for the Stekelenburg&Vroomen data, seen as a performance drop when the model was evaluated on test data (bottom of Figure 8.1). Overfitting was also a problem for the BDCA methods while the more restricted PARAFAC version of ManTDA, ManPDA, did not suffer from overfitting. The methods ManPDA, ManTDA_sr and ManPDA_sr, retained performances similar, albeit a little lower, to CMDA and DATER on the test data. The unsupervised methods and DGTDA performed considerably worse than the supervised MDA methods on both the test and training data. The Tucker2 method, which resembles the MDA methods by relying on single-trial projections into core arrays, outperformed the other unsupervised methods.

Table 8.1 shows the mean classification rates of letters over the two subjects from the BCI data. Each method was trained on all training data with the number of components that performed best in the five-fold CV on training data. Each method was then used to classify each trial in the test data, and these classifications determined the letter classifications.

The top part of Table 8.1 shows the compared methods in decreasing order of performance, and the bottom shows the five best results obtained in the BCI competition. Since there are many more trials in the BCI data than in the Stekelenburg&Vroomen data, the problem of overfitting is less likely to occur in this data set. Indeed, ManTDA and the BDCA methods achieve high classification rates here. As was also the case for the Stekelenburg&Vroomen data, the unsupervised methods and DGTDA perform substantially worse than the BDCA and MDA methods, except for DGTDA.

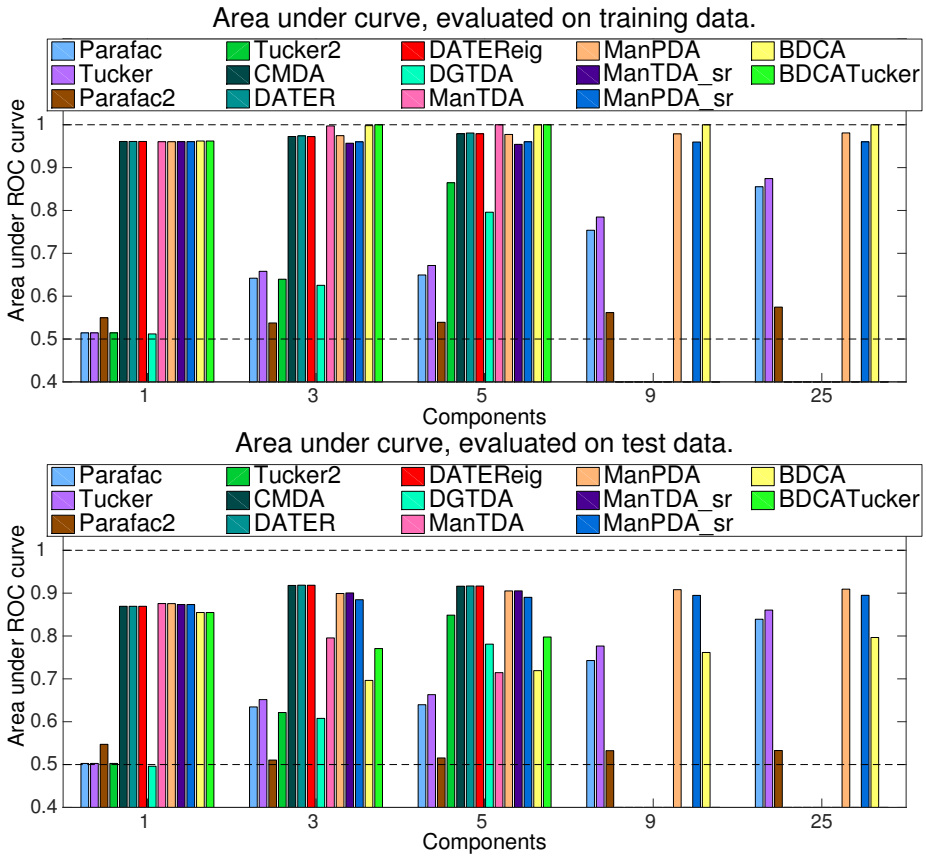


Figure 8.1: Area under ROC curve scores obtained when the methods were evaluated on (top:) training- and (bottom:) test data from the Stekelenburg&Vroomen data set. The figure is parts of figure in [Frølich et al., 2016].

	15 flashes	5 flashes
DATEReig	0.930	0.695
CMDA	0.925	0.695
DATER	0.925	0.670
ManTDA	0.915	0.660
ManPDA	0.910	0.645
BDCA	0.895	0.645
ManPDA_sr	0.890	0.555
BDCATucker	0.890	0.655
ManTDA_sr	0.880	0.555
Tucker2	0.605	0.260
DGTDA	0.595	0.320
Parafac	0.315	0.105
Tucker	0.300	0.090
Parafac2	0.025	0.005
Contributor	15 flashes	5 flashes
Alain Rakotomamonjy	0.965	0.735
Li Yandong	0.905	0.550
Zhou Zongtan	0.900	0.595
Ulrich Hoffmann	0.895	0.530
Lin Zhonglin	0.875	0.575

Table 8.1: Mean letter classification rates for Dataset II from BCI Competition III from the compared methods (top) and best five competition participants (bottom), copied from `http://www.bbci.de/competition/iii/results/index.html`. The table is from [Frølich et al., 2016]

Chapter 9

Conclusion

Our results showed that rigorous optimisation of MDA objective functions improves the fit to training data. This also increases the risk of overfitting, as was evident on the Stekelenburg&Vroomen data with few trials. The PARAFAC structure protected against overfitting, as did the heuristic optimisation approaches utilised by CMDA and DATER. The BDCA methods were also seen to be susceptible to overfitting, but performed well when this was not a problem. Apart from the overfitting problem, all the supervised methods, except DGTDA, obtained similar classification performances. The unsupervised methods, as well as the MDA method DGTDA, had substantially lower performances than the other methods. The unsupervised methods and DGTDA also had the worst performance when tested on letter classification on the BCI data. Likewise, all the supervised methods, except DGTDA, provided similar classification rates. Hence we conclude that supervising the process of finding subspaces is advantageous for classification, but the choice of supervised method is not so important. Likewise, similar performances were seen for the same method differing only on whether it used the PARAFAC or Tucker structure. Since models based on the PARAFAC structure are more easily interpretable, these methods are appealing.

Part V

Discussion & Conclusion

Chapter 10

Discussion

The work described in this thesis was about decompositions of EEG data, classification of estimated EEG sources, use of these classifications in further analyses, and single-trial classification.

In "Classification of independent components of EEG into multiple artifact classes", we developed a classifier for ICs of EEG data, IC_MARC, which classifies an IC as belonging to one of the classes: neural, blinks, lateral eye movements, heart beats, muscle contractions, or mixed. Classification rates were high on (binary) neural vs. non-neural decisions within and across studies. Multi-class performances were acceptable across subjects within a study, but dropped in cross-study testing.

Although rigorous evaluation of IC_MARC on the two data sets containing labeled ICs showed low multi-class performances when evaluated on the study that was not used for training, visual inspection of scalp maps [Frølich et al., 2015b, Brandl et al., 2016] and spectra [Brandl et al., 2016] of classified ICs from new data sets indicated reasonable multi-class classification behaviour of IC_MARC. Similarly, the distributions of artefact types in distraction tasks in [Brandl et al., 2016], as determined using IC_MARC, were in good accordance with other findings in that work. Specifically, the artefact distribution in one task was very different from the distributions in the other tasks, and this task also required a different classifier from the classifier used on all other tasks. This

is another indication that the multi-class classifications provided by IC_MARC are sound, although not perfect, even though the rigorus evaluation did not show this. More data sets with labels would enable new analyses of the performance of IC_MARC, and also allow training on more diverse data. Data sets with labeled ICs are likely to become available in the near future due to the crowd-sourcing initiative for IC labeling set in motion at the Swartz Center for Computational Neuroscience at `http://reaching.ucsd.edu:8000/tutorial/overview`.

Using IC_MARC, we performed analyses that would have been impractical with manual IC classifications. In "Investigating effects of different artefact types on Motor Imagery BCI" [Frølich et al., 2015b] and "Brain-Computer Interfacing under Distraction: An Evaluation Study" [Brandl et al., 2016], we inspected effects of artefacts on motor imagery based BCIs. Although some improvement was obtained by restricting the influence of muscle artefacts [Frølich et al., 2015b], the same advantage was gained by removing electrodes on scalp edges in [Frølich et al., 2015b], where muscle contamination is most prominent. While we were not able to improve BCI performances by removing and regularising against artefactual ICs, these strategies did not lead to substantial decreases in BCI performance either. This indicates that cleaning BCI data by removing artefactual ICs does not interfere with the neural signals used for classification. The inability to improve performance by removing artefactual ICs may be a sign that artefacts are similarly distributed across classes and hence ignored by the Common Spatial Patterns algorithm when identifying the spatial filters optimal for class discrimination.

In "Artefact removal using linear decompositions of EEG", we used neural vs. non-neural classifications of estimated EEG sources to compare six decomposition methods and the effect of high-pass filtering data prior to decomposition [Frølich and Winkler, 2016]. PARAFAC2 had the lowest performance while Fourier-ICA and SSD provided cleaning close to that achieved by ICA. High-pass filtering at high cut-off frequencies improved performances of PARAFAC2 and the included ICA variants(Extended Infomax, FastICA, and TDSEP). When data were high-pass filtered prior to decomposition, we did not observe consistent differences between the ICA variants. This implies that the choice of ICA method is not so important when used for cleaning data in certain frequency ranges, whereas pre-processing steps may have larger effects on the cleaning obtained. Since SSD is computed by solving a generalised eigenvalue problem, decomposition via SSD is faster than using ICA methods, which are iterative. The low computational time required by SSD may make SSD a good alternative to ICA methods since the cleaning obtained by SSD was close to that obtained by the ICA methods.

Finally, we compared multi-linear classification methods on single trials of raw EEG data [Frølich et al., 2016]. We compared classification performances from

1) feature extraction using four existing unsupervised methods followed by logistic regression, 2) feature extraction by four existing and four new Multi-linear Discriminant Analysis (MDA) methods followed by logistic regression, and 3) combined feature extraction and classification by one existing and one new logistic regression based method. The four new MDA methods were proposed in a collective framework in which MDA objective functions were optimised using the conjugate gradient algorithm (as provided in the *ManOpt* toolbox [Boumal et al., 2014] for Matlab) on a cross-product of Stiefel manifolds to enforce orthonormality constraints on the projection matrices. The four new MDA methods consisted of two objective function structures, each formulated as a PARAFAC and a Tucker version, optimised within the manifold-optimisation framework. We found that rigorous optimisation provided higher objective function values for the MDA methods. However, this did not translate into better classification performances. This might be due to overfitting issues, implying that a regularisation scheme combined with rigorous optimisation would likely obtain better classification performances. Similarly, both the existing (BDCA) and the new method endowing BDCA with a Tucker structure were optimised rigorously and also showed signs of being susceptible to overfitting. These observations indicate that the heuristically optimised methods have inherent protection against overfitting while explicit regularisation should be applied in the rigorously optimised MDA and BDCA methods. The performances obtained with the Tucker and PARAFAC structures were similar, indicating that the PARAFAC model is a good model of EEG data, which is fortunate since the PARAFAC structure is easier to interpret than the Tucker structure.

In future work, multi-linear classification methods could be used for IC classification. This would enable incorporation of the temporal variation in ICs' time series by e.g. calculating temporal and spectral features in small time windows, resulting in a feature $\times$ time matrix. Multi-linear methods might be able to exploit structure in such data to a higher degree than is possible when vectorising feature vectors.

Code for the IC classifier and the compared tensor classification methods is available at `http://www2.compute.dtu.dk/~lffr/publications/indexpub.php`. The tensor classification code is customised for matrix observations.

Chapter 11

Conclusion

Independent Component classification performances were high on (binary) neural vs. non-neural decisions within and across studies. Multi-class performances were acceptable across subjects within a study, but dropped in cross-study testing.

We were not able to improve BCI performances substantially for any artefact group, but removing muscle artefacts had the largest effect [Frølich et al., 2015b]. The same effect was obtained by removing outer electrodes, though.

Another study showed that high-pass filtering data at high cut-off frequencies improved the cleaning obtained by removing estimated artefactual sources and provided similar performances of the ICA variants, which outperformed the other methods.

In comparisons of classification performances with different multi-linear methods, we found that supervised methods outperformed unsupervised methods by a large margin, but that differences between supervised methods were small. Rigorous optimisation provided higher objective function values but did not lead to better classification performances. Finally, the Tucker and PARAFAC structures provided similar performances.

Part VI

Articles

Appendix A

Classification of independent components of EEG into multiple artifact classes

Classification of independent components of EEG into multiple artifact classes

Laura Frølich[a,*], Tobias S. Andersen[a], Morten Mørup[a]

[a]*Section for Cognitive Systems, DTU Compute, Technical University of Denmark, Matematiktorvet, Building 321, 2800 Kgs. Lyngby, Denmark*

Keywords: EEG, artifact, independent component, multi-class classification, cross-study generalization

*Corresponding author, Tel.: +45 4525 3431, Fax: +45 4588 1399

Email addresses: lffr@dtu.dk (Laura Frølich), toban@dtu.dk (Tobias S. Andersen), mmor@dtu.dk (Morten Mørup)

URL: http://www2.imm.dtu.dk/~lffr/ (Laura Frølich), http://www.mortenmorup.dk/ (Morten Mørup)

Abstract

In this study, we aim to automatically identify multiple artifact types in EEG.

We used multinomial regression to classify independent components of EEG data, selecting from 65 spatial, spectral, and temporal features of independent components using forward selection. The classifier identified neural and five non-neural types of components.

Between subjects within studies, high classification performances were obtained. Between studies, however, classification was more difficult. For neural vs. non-neural classifications, performance was on par with previous results obtained by others.

We found that automatic separation of multiple artifact classes is possible with a small feature set.

Our method can reduce manual workload and allow for the selective removal of artifact classes. Identifying artifacts during EEG recording may be used to instruct subjects to refrain from activity causing them.

1. Introduction

EEG data is generally contaminated by artifactual, non-neural electrical activity stemming from non-physiological sources such as electrical background noise and loose electrodes, and physiological sources such as subjects' heartbeat, muscle or eye movements. Such non-neural activity can, to some extent, be separated from the data using Independent Component Analysis (ICA), which is a widely used method in data analysis (Comani et al., 2004; Di et al., 2007; Hyvärinen and Oja, 2000; Kim et al., 2003; Kong et al., 2008; Tsai and Lai, 2009). Particularly, it is commonly used for pre-processing and analyzing EEG data (Acar et al., 2008; Erfanian and Erfani, 2004; Ullsperger and Debener, 2010). ICA extracts spatial patterns with statistically independent behavior over time from the raw EEG data (Hyvärinen and Oja, 2000). These patterns and their corresponding time series are referred to as independent components (ICs).

Non-neural activity in EEG is typically considered a nuisance and the main purpose of separating it from the data using ICA is to exclude it by filtering (Jung et al., 2000). Other approaches to cleaning data include identifying heavily contaminated channels or epochs of EEG data, and then removing such channels or epochs prior to analysis (Citi et al., 2010; Nolan et al., 2010; Ypparila et al., 2004). Unfortunately, this may lead to unnecessary data loss. Simultaneous reordings of e.g. the electrooulogram (EOG) and electrocardiogram along with EEG can also be used to remove artifats (Fatourechi et al., 2007; He et al., 2004; Nolan et al., 2010) but this approach is not useful for all types of artifacts and requires the additional labor of mounting auxiliary sensors. Therefore, we only consider approaches using ICA based solely on EEG data in the current study.

1.1. State of the art

Presently, classification of ICs for artifact detection in EEG is often done manually in a time-consuming and subjective process. While work on fully automated supervised classification methods has increased over the past years, most of this work has focused on the binary problem of distinguishing between neural and non-neural ICs (Bartels et al., 2010; Halder et al., 2007; LeVan et al., 2006; Mognon et al., 2010; Tangermann et al., 2009; Viola et al., 2009; Winkler et al., 2011), some using multiple classes as an intermediate step (Bartels et al., 2010; Halder et al., 2007; Mognon et al., 2010) with only few studies evaluating performance for the multi-class problem (Halder et al., 2007; Viola et al., 2009).

Several studies have used simulated artifacts as a ground truth to which they compare their automatic classification e.g. (Delorme et al., 2007; Nolan et al., 2010). This is problematic as real artifacts may affect data in ways different from the simulation. Therefore, we limit our review to those studies that compare performance with human experts' classifications of real artifacts in real data.

The most important factor in performance evaluation is generalization. For a method to be fully automated it should perform well when tested on data from a study that was not used to train the method. Automatic thresholding at e.g. a pre-determined Z-score for a certain feature is one approach that allows this (Mantini et al., 2008; Nolan et al., 2010). Another approach is to train a classifier on data from one study and make sure that it performs well on data from another study. This would allow the method to be trained once and then applied to new data without manual intervention. Few studies have tested

supervised classifiers for artifact detection at this level of generalization, and only for the binary problem of distinguishing artifactual from neural ICs (Mognon et al., 2010; Winkler et al., 2011).

Winkler et al. (2011) built a classifier based on an initial pool of 38 features from the spatial, spectral, and temporal domains. They compared several classification methods and found that regularized linear discriminant analysis with three spectral, one temporal, and two spatial features obtained the best classification results. We describe most of the features in their initial pool in detail in Appendix A. They reported error rates of 8.9% and 14.7% within and between studies, respectively.

The method ADJUST (Mognon et al., 2010) uses Gaussian densities for classification and incorporates features from the spatial and temporal domains of ICs. ADJUST employs class-specific classifiers for eye blinks, vertical eye movements, horizontal eye movements and generic discontinuities (non-biological artifacts) to solve the classification problem by classifying an IC as being non-neural if one or more class-specific classifiers labeled it as artifactual. The evaluation measure reported for ADJUST was the percentage of data variance explained by correctly classified ICs. On test data, the ADJUST performance measure was 99.0%, 96.0%, 99.2% and 97.7% for the class-specific classifiers for blinks, vertical eye movements, horizontal eye movements, and generic discontinuities, respectively. In classifying neural vs. non-neural ICs, the ADJUST performance measure was 95.2%. Several classes were considered in ADJUST, and so the method is appropriate to be used for multi-class classification purposes. Since an IC may be assigned to several classes, ADJUST can, strictly speaking, not be tested in the multi-class problem in its current form.

A few studies have addressed multi-class identification of artifacts (Mantini et al., 2008; Viola et al., 2009) at the level of across-subject-generalization within a study. This level of generalization could certainly also be useful as it would allow automatic artifact classification of future subjects once manual classification of some subjects has been achieved.

Viola et al. (2009) introduced the method CORRMAP, which solely uses the scalp map of an IC to classify it as representing a blink, a lateral eye movement, or the heartbeat. CORRMAP classifies an IC using the correlation between the spatial topography of the IC and template topographies from ICs with known classes. If the correlation is higher than a certain threshold, which can either be set manually or determined automatically, then the IC is classified as being of the same class as the template IC.

In Viola et al.'s study, classification rates were calculated for electrode configurations with 30, 68, and 128 channels for three classes: blinks, lateral eye movements, and heartbeats. The mean correlation over electrode arrays between CORRMAP and human experts for these three classes were 0.90, 0.88, and 0.47, respectively. The classification rates for blinks and lateral eye movements were higher for the less dense electrode arrays, while classification of heartbeats improved with denser electrode arrays.

A new fully automated method using the same principle as CORRMAP, of using the correlation between spatial maps as sole feature, has recently been presented (Bigdely-Shamlo et al., 2013) for the identification of eye-related ICs. An area under the receiver operating characteristics (ROC) curve of 0.993 was obtained on ICs from a study that was not used during training. This result shows that the principle behind CORRMAP is a very promising method for automatic artifact identification at the highest level of generalization, namely across studies.

Table 1: Data summary

Study	Subjects	Channels	Reference	ICA	Sampling rate	Duration	ICs
Emotion	34	250	Active	Infomax	256Hz	8-88 minutes	7255
Cue	12	64	Linked mastoids	Infomax	500Hz	56-66 minutes	768

Mantini et al. (2008) used thresholding of a single feature, the approximate entropy of IC time series, to classify ICs of MEG as non-cerebral biological artifacts (low approximate entropy), neural (medium approximate entropy) or environmental noise components (high approximate entropy). They obtained very good results with the area under the ROC curves being above 0.9 with labels by human experts as ground truth. As this method separates artifacts into biological and non-biological ICs it does address the multi-class problem but it is unknown whether it is suitable for a further division of these classes into more specific classes such as lateral eye movements versus eye blinks.

1.2. Purpose of study

By distinguishing between multiple types of artifacts such as eye movements and the electrical heartbeat artifact, more diverse uses of an automatic classification method can be imagined since some artifacts may be informative for some purposes, or it may be desirable to remove only some artifact types. The heartbeat, for example, may be an informative signal in some settings, or eye-related ICs could be used to detect drowsiness. Automatic detection and identification of multiple types of artifacts during EEG recording would also allow researchers to instruct subjects to refrain from the activity causing those artifacts.

The purpose of the study is to develop a multi-class artifact detection system covering four diverse artifact classes: eye blinks, horizontal eye movements, heartbeat artifacts and muscle artifacts, as well as ICs consisting of mixed neural and artifactual activity. Importantly, we test the performance of the system at two levels of generalization: between subjects within a study and between studies. A good performance across subjects would allow a classifier to be trained for the first subjects in a study, and then used to automatically classify ICs for the subsequent subjects. A good performance across studies would mean that the classifier can be used on arbitrary studies and subjects without re-calibration. We are also interested in determining the features most relevant to classifying ICs. Hence we aim to answer the following research questions:

1. Which features are important for a high performance in multi-class classification of ICs?
2. Is it possible to distinguish between multiple classes of ICs between subjects within a study?
3. Will a classifier generalize between studies?

2. Data

Two data sets containing manually labeled ICs were kindly made available by Scott Makeig, Julie Onton and Klaus Gramann (Gramann et al., 2010; Onton and Makeig, 2009).

One data set was acquired for the purpose of studying the EEG during different emotional states (Onton and Makeig, 2009). Subjects were seated in a dimly lit room with eyes

closed, imagining emotional states. This study contained recordings from 34 subjects from a Biosemi[1] 250 channel active reference system (Onton and Makeig, 2009). Channels that showed highly abnormal activity had been removed manually before performing ICA, leaving 134-235 channels for each subject. The ICA decompositions for this data were obtained by "full-rank decomposition by extended infomax ICA" (Onton and Makeig, 2009). The 34 data sets were between eight and eighty-eight minutes long after concatenating the recordings for the various emotions imagined. We will refer to this data set as the *Emotion* data or study.

The other data set was recorded to investigate how attention is guided early in visual processing. This was recorded from 64 scalp channels "referenced to Cz and re-referenced off-line to linked mastoids" from 12 subjects during a visual task (Gramann et al., 2010). ICA was performed with the implementation of the ICA infomax algorithm in the Brain Vision Analyzer software from Brain Products GmbH[2]. The data sets we had access to were between 56 and 66 minutes long for the different subjects. We will refer to data from this study as the *Cue* data or study. The data sets are summarized in Table 1.

The two data sets differed in various ways (see Table 1). The number of electrodes was much higher in the Emotion study than in the Cue study, implying a higher spatial sampling of the EEG. The Emotion study also contained more subjects, resulting in a total of almost ten times as many ICs in the Emotion study as in the Cue study. Also, different sampling rates and analogue filters were used and the lengths of recordings differed. Additionally, the experimental tasks differed. In the Emotion study, an eyes-closed task was performed while a task requiring responses to visual cues was used in the Cue study. These differences are likely to cause covariate shifts in the data, i.e. differences in distributions of features between training data and future data (Sugiyama and Kawanabe, 2012), in the features across studies if features are calculated naively from the raw data. We discuss how we handle this issue in Section 3.

Both studies contained ICs labeled by experts with the labels "eye blink", "neural", "heart", "lateral eye movement", and "muscle". Two experts, one in each study, performed the manual classification of ICs. Figure 1 shows examples of scalp maps from the different classes. Neural ICs are the ICs that correspond to activity generated by neural sources within the brain. ICs with the label "heart" represent the electrical heartbeat artifact. The ICs that were not labeled represented, based on visual inspection, mixed ICs containing both artifactual and neural signals. We will refer to the unlabeled ICs as "mixed" ICs. We chose to include mixed ICs in our analysis since mixed ICs will almost always be present in real data. Not including this class would then force mixed ICs to be classified as one of the four artifact, or neural classes. Since mixed ICs have different characteristics from neural ICs, it is likely that many would be classified as artifactual. This is undesirable since mixed ICs also contain traces of neural activity, meaning that the removal of mixed ICs would imply a loss of neural activity in data. The inclusion of mixed ICs can also be seen as a step toward making the classifier mimic human expert classifications as much as possible.

Some types of ICs are much more common than others, which presents a challenge to classification methods as described in Section 3. Mixed ICs, for example, make up the

[1] `http://www.biosemi.com/`

[2] `http://www.brainproducts.com/`

Figure 1: Two examples of scalp maps for each of the six classes in each of the two studies

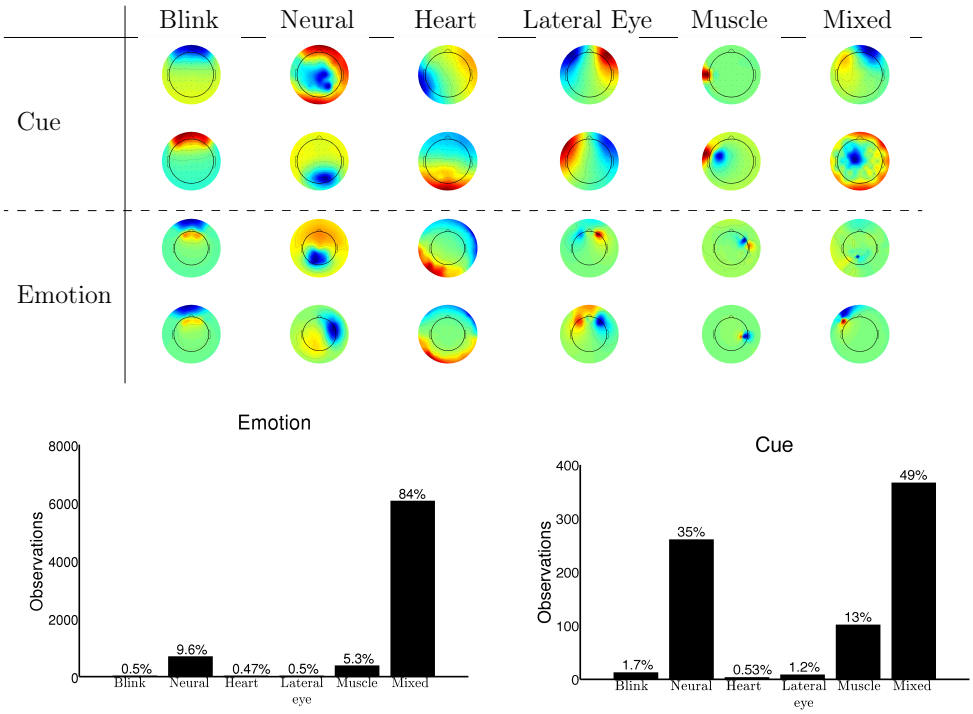


Figure 2: The number (and percentage over each bar) of ICs in the six classes for the Emotion and Cue datasets. These distributions reflect the experts' manual classifications.

majority of available ICs. The numbers and proportions of the different types of ICs in each study are shown in Figure 2.

3. Methods

Figure 3 shows the pipeline used to train and validate our IC classifier. Each of the steps is described in detail in the remainder of this section.

We first discuss the steps taken during pre-processing to avoid covariate shifts between studies due to differences in experimental setups. Next, we discuss our feature set. We then describe our classification and feature selection procedures. Finally, we outline how we investigated the research questions posed in the introduction.

3.1. Pre-processing

Different EEG studies use different sampling rates, analogue filters, and electrode arrays during recordings, and durations of recordings vary. If features that are influenced by such differences are used, it is improbable that a classifier will generalize across studies.

Higher sampling rates enable spectral features to be determined for higher frequencies. Likewise, different analog filters during recording of EEG cause the spectral content of signals

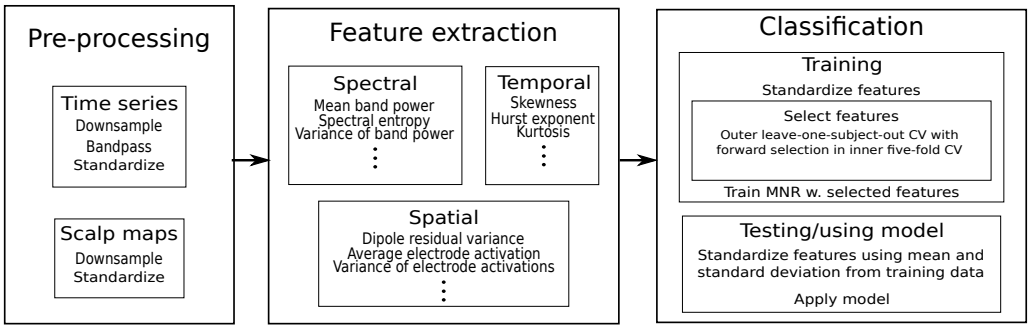


Figure 3: Processing pipeline for ICs from EEG data. The abbreviations CV and MNR stand for cross-validation and multinomial regression, both explained in section 3.3.

to vary systematically. To avoid such differences, we filter and resample all signals before calculating features. We require that any data given as input was recorded with a sampling rate of at least 200Hz, and that the analogue filter used during recording had a low edge of 3Hz or lower and a high edge of 90Hz or higher. With these requirements in place, it is safe to band-pass filter the signal between 3Hz and 90Hz and downsample all input signals to 200Hz. This ensures that all feature calculations are performed on signals with the same spectral content.

Different durations of recordings entail different uncertainties in the calculation of temporal and spectral features. Invariance to this effect is achieved by using the means and variances of temporal and spectral characterstics of the signal over one-second intervals as temporal and spectral features.

Some features are based on distances between electrodes and are thus clearly influenced by electrode array density. We require that recordings were performed using an array with at least 64 electrodes to ensure a good spatial coverage. We spatially downsample all scalp maps to the 10-20 system electrode array with 64 electrodes. The spatial downsampling is performed with Gaussian kernels using spherical distances between electrodes. We use a standard deviation of 0.5 cm and a head radius of 9 cm.

Before calculating features derived from the spatial distribution of an IC, we standardized the spatial map. Each column of the mixing matrix was standardized to have variance one and mean zero. This ensures that only patterns in the spatial map, and not its scale, determine the features calculated. This is desirable since the magnitude of the mixing matrix cannot be uniquely determined due to an inherent ambiguity in the scaling of the mixing matrix and the matrix of activation time series of ICs. We also standardized time series before calculating temporal and spectral features.

3.2. Features

An IC consists of a scalp map containing the contribution of the IC to each EEG channel, and a time series that shows how active the spatial pattern is over time. To quantify the characteristics of an IC, features based on both the spatial and temporal representations have been shown to be relevant (Mognon et al., 2010; Viola et al., 2009; Winkler et al., 2011). Spectral (frequency domain) characteristics of the time series have also been shown

to be informative (Winkler et al., 2011). Hence we use features from the spatial, temporal, and spectral domains. We included most of the features described in two recent studies of the binary classification problem (Mognon et al., 2010; Winkler et al., 2011). Descriptions of features are given in Appendix A. Before training we standardized the features in the training set to have mean zero and variance one. We standardized the test data using the mean and variance from the training data, which is the standard approach (Hastie et al., 2009; Jayalakshmi and Santhakumaran, 2011).

3.3. Classification

We used the linear classifier multinomial logistic regression (MNR) since this was found to obtain good results and linear classifiers are desirable both for their interpretability and fast training. Linear classifiers have previously shown good performance in the binary classification of ICs (Winkler et al., 2011).

As is evident from Figure 2, the class of mixed ICs makes up the large majority of ICs in both studies. Thus a classifier would achieve a high classification rate by classifying all ICs as mixed. This problem of imbalanced classes is well known, and various approaches to solving it have been proposed (López et al., 2012; Zadrozny et al., 2003). We weighted observations by the reciprocal of their class proportion during training such that the penalty of misclassification was higher for ICs from smaller classes. This weighting scheme can be considered a proxy for optimizing balanced accuracy. Balanced accuracy is a performance measure that weighs all classes equally since it is defined as the mean over classes of the proportion of correct classifications in each class. In the binary case, balanced accuracy is thus the mean of specificity and sensitivity.

Previous studies on the binary classification problem found that only few features are necessary to distinguish between classes (Mognon et al., 2010; Tangermann et al., 2009; Winkler et al., 2011). This motivated us to investigate research question 1 of whether only few features are sufficient in the multi-class problem as well. This was done in a two-level cross-validation (CV). In the outer level, leave-one-subject-out CV was performed over the 34 subjects in the Emotion data. In each outer fold, features were chosen using forward selection in an inner 5-fold stratified CV by adding features to an MNR model until the test error stopped decreasing. The use of stratified CV ensured that class proportions were as equal as possible across partitions. For each feature, we counted the number of outer CV folds in which it was selected. This number reflects the importance or pertinence of the feature. We then created 35 sets of features consisting of the features that had been selected in at least 0, 1, 2, ..., 34 outer CV folds. For each subject, the classifier was trained on each of these feature sets using the 33 other subjects, and tested on the left-out subject. The classes of ICs predicted for each subject in this manner were used to calculate a balanced accuracy for each feature set. As the best feature set we chose the sparsest feature set with acceptable performance.

3.4. Investigation of research questions

Research question 1, concerning the features important for multi-class classification, was investigated by comparing classification performances with different feature sets. These feature sets were the ones constructed using the Emotion data as described in Section 3.3. The Emotion data was also used to choose the best feature set. To evaluate the sensitivity of

the classification performance to the choice of features, balanced accuracies were calculated in leave-one-subject out CV on the Cue data and across-study training and testing for each feature set constructed from the Emotion data. If new ICs to be classified have short time series, spectral and temporal features will likely be badly determined. In such cases, the exclusive use of spatial features would be preferable. For this reason, we also tested the classifier using only the spatial features.

Both research questions 2 and 3 were investigated using the feature set determined based on Emotion data. We investigated research question 2, concerning between-subject generalization within studies, through the leave-one-subject-out CV schemes on both the Emotion and Cue data sets. A high classification performance when testing on a subject not used during training would signify that it is possible for a classifier to generalize across subjects within a study, meaning that each class of ICs exhibits certain characteristics independently of the specific subject.

To answer question 3, concerning between-study generalization, we trained a model on each data set using the features selected using the Emotion data. The models were then tested on all subjects from the other study. A good performance on subjects from the other study would indicate that the classifier is able to generalize across studies.

We used confusion matrices to inspect the classification performance of the classifiers on a class-by-class basis. We also used the balanced accuracy rate to evaluate performance and compare to classification performances obtained by others.

4. Results

Figure 4 shows the number of times each feature was chosen by forward selection in the leave-one-subject-out CV scheme performed on the Emotion data. The balanced accuracies obtained using the features chosen in at least 15, 20, 25, 28, or 34 outer folds are also shown. The feature sets constructed using the thresholds 15, 20, 25, 28, and 34 contain 32, 23, 14, 14, and 3 features, respectively. The two 14-feature sets are identical.

Figure 5 shows the balanced accuracies obtained with each of the 35 feature sets. The variability of the curves in Figure 5 gives an idea of how sensitive the classification performance is to the choice of feature set.

Figure 6 shows the confusion matrices that arose from using the 32-feature set, the 23-feature set, the 14-feature set, and the 3-feature set in a leave-one-subject-out CV on Emotion data. This figure is included to show that the class-wise performances are stable over the different feature sets.

Figure 7 shows the confusion matrices obtained in leave-one-subject-out CV on both studies, and with cross-study training and testing using only the spatial features in the initial pool of features. This figure is shown to illustrate the classification performance that can be expected if only short time series of ICs are available, in which case non-spatial features may be unreliable.

Figure 8 shows the class-wise classification performances when the classifier with the feature set containing 14 features is used. The confusion matrix in the top row shows the performance with leave-one-subject-out evaluation on the Cue data and the two confusion matrices in the bottom row show the cross-study performances. This figure details the class-wise performances, which cannot be derived from the balanced accuracy rates shown in

Figure 5.

5. Discussion

Before analyzing the classification performance obtained by our classifier we discuss the classification performance of human experts, which sets the upper bound on the performance we might hope to achieve.

5.1. Performance of human experts

As the true underlying content of ICs, i.e. the ground truth, is unknown, we can only rely on classifications made by expert human observers when training and testing classifiers. Several studies have found that the agreement between human experts is generally less than perfect and that it differs for different types of artifacts (Klekowicz et al., 2009; Viola et al., 2009; Winkler et al., 2011). Although the agreement between experts is likely dependent on the particular method of ICA, the information available to the experts, the particular data sets and how experts are instructed to classify ambiguous cases, there seems to be a good agreement between studies on the inter-expert agreement rate (Klekowicz et al., 2009; Viola et al., 2009; Winkler et al., 2011).

Viola et al. (2009) had 11 independent experts classify ICs as eye blinks, lateral eye movements and heartbeat artifacts based solely on the scalp maps of ICs. The data came from three independent studies and observers were under the constraint that a maximum of three ICs could be identified as containing one particular artifact type. In terms of the binary correlation the inter-expert agreement was very high for eye-blinks (0.82 – 1.00), high but more variable for lateral eye movements (0.55 – 0.93) and low and very variable for heartbeat (0.02 – 0.73).

Winkler et al. (2011) had 2 experts classify ICs from a single study as artifactual or neural based on their spectrum, time series and spatial distribution on the scalp and found that the error rate was 10.6%. They also had one expert re-label the ICs from another study two years after the same expert's first labeling of the same data. The error rate between the two labelings was 13.2%. This is not much higher than the agreement between experts and the disagreement may thus reflect the inherent difficulty of the task rather than differences in technique or approach by different observers.

Klekowicz et al. (2009) made 22 comparisons between expert classifications (artifact vs. neural) based on the EEG time series from 7 polysomnographic recordings and found an agreement of 0.92 in terms of the area under curve of the best fitting ROC curve. Of the 22 comparisons, four were between classifications made by the same expert at different points in time. From their figures (Figure 6 in their article) these agreements were high compared to the agreement between different observers. Hence, their reported overall agreement between human classifications is a high estimate of the agreement between human experts.

The imperfect agreement between human experts should be kept in mind when evaluating automatic artifact detection systems as inter-expert agreement sets the upper limit for what we can hope to achieve through automatic classification. It is very promising that several studies have reported a good agreement between automatic IC classification and human experts, close to the agreement between experts.

5.2. Evaluation of classifier

Feature selection. In Figure 5, the blue curve shows the average leave-one-subject-out CV performance on the Emotion data, the same data used to construct the feature sets. This is also the curve used to determine the feature set to use in the classifier. The feature set resulting from requiring that features must have been included in 28 CV folds or more was chosen as the best feature set since classification performance starts to consistently decrease at this threshold. This feature set includes 14 features, consisting of nine spatial, two spectral, and three temporal features. The red and blue curves are biased upwards since testing for these curves was performed on the Emotion data, which was used to choose the feature sets, implying that the feature sets contain features especially well suited to describing ICs of different classes in the Emotion data. At threshold zero, when all features are included, there is no bias since no features were chosen based on the Emotion data at this point. Figure 5 shows that, when training and testing on subjects from the same study (blue and green curves), the performance is stable for most feature sets until the number of features becomes too small. This indicates that, within a study, overfitting to subjects in the training data is not a problem, even for the relatively small amount of data present in the Cue study. The lack of upwards or downwards trends in the performance when training on Emotion data and testing on Cue data (cyan curve) indicates that the Emotion study contains sufficient data that overfitting is avoided. Conversely, when training on Cue data and testing on Emotion data (red curve), the performance peaks with feature sets that are neither too small nor too large. One explanation of this is that there is not enough data in the Cue data set to prevent overfitting when very large feature sets are used. Another explanation is that, since features were chosen based on Emotion data, small feature sets help the model home in on characteristics that best discriminate classes of ICs in Emotion data. All curves indicate that underfitting occurs with feature sets that are too small. In summary, Figure 5 shows that the classification performance is quite robust to the specific choice of threshold when training and testing on subjects from the same study, whereas the performance is sensitive to the choice of threshold when training on one study and testing on the other study.

The inclusion of both spatial, spectral, and temporal features in nearly all feature sets (Figure 4) shows that all three types of features carry information on the classes of ICs. The features included in the 14-feature set are shown in Table 2, arranged according to the classes they should be good at detecting.

For the spatial feature set, the within-study performances were very similar to those obtained with the 14-feature set (compare confusion matrix (c) in Figure 6 and confusion matrix (a) in Figure 8 to confusion matrices (a) and (b) in Figure 7). In the between-study case, the performance improved when testing on Cue data and decreased when testing on Emotion data (compare confusion matrices (b) and (c) in Figure 8 to confusion matrices (c) and (d) in Figure 7). However, the performance when testing on Emotion data with the 14-feature set is biased upwards since features were chosen using the Emotion data. Thus the decrease seen when testing on Emotion data with the spatial feature set might be artificial, indicating that spatial features may be sufficient if across-study generalization is to be improved.

Classification performance with the 14-feature set. The following discussion of the classification performance is based on the results given for the classifier with the 14-feature set.

Table 2: Selected features

Class	
Blink	SAD, theta, lowFreqPowAvg, logRangeTempVar, var1sAvg, timeEntAvg
Neural	central, centralActivation
Heart	zcoord, timeEntAvg
Lateral eye	SED, SAD, theta, lowFreqPowAvg, logRangeTempVar, var1sAvg, timeEntAvg
Muscle	logRangeSpatial, spatDistExtrema, logRangeTempVar, var1sAvg, timeEntAvg
Mixed	cdn, dipoleResidVar

When classifying ICs in the within-study case into only two classes, artifactual or non-artifactual, we obtain balanced accuracy rates of 0.90 and 0.95. This is comparable to performances obtained by others. Balanced accuracy rates of 0.91 and 0.79 were obtained in Winkler et al. (2011) and LeVan et al. (2006), respectively, while Halder et al. (2007) and Bartels et al. (2010) report balanced accuracy rates above 0.90 without giving the exact numbers. Likewise, our classifier performs on par with others in the binary across-study case, obtaining balanced accuracy rates of 0.88. In the across-study case, Winkler et al. obtained a balanced accuracy of 0.86 (Winkler et al., 2011). These accuracy rates compare well with the inter-expert agreement seen in previous studies (Klekowicz et al., 2009; Viola et al., 2009; Winkler et al., 2011).

In the following, we discuss the multi-class performance. This is visualized in confusion matrix (c) in Figure 6 for the leave-one-subject-out CV on the Emotion data, and in Figure 8 for the leave-one-subject out CV on the Cue data and the cross-study training and testing. The performance on the class of lateral eye movements is high. This could be expected since eye-related ICs have previously been classified well by many others (Bigdely-Shamlo et al., 2013; Mognon et al., 2010; Viola et al., 2009). For the blink class, however, difficulty is experienced when training on Cue data and testing on Emotion data. This could be due to the low number of observations (14) of the blink class in the Cue data, making it difficult for the classifier to learn a good characterization of this class. The high performance on the neural class is also in good agreement with that found by others (Mognon et al., 2010; Winkler et al., 2011). When tested on Cue data, heartbeat ICs tended to be misclassified as neural. Difficulty with the heartbeat class has also been observed in previous work including this class, both for an automatic classifier and for human experts (Viola et al., 2009). The high degree of confusion between the classes of muscle and mixed ICs may partly be explained by the shared characteristic of highly peaked scalp maps in these two classes compared to the other classes. The class most often confused with other classes is that of mixed ICs, which is not surprising since mixed ICs are ICs that do not clearly belong to one class, but may contain characteristics of several classes. The classification of some mixed ICs as neural is arguably difficult to avoid as the contrast between neural and mixed will be based on a threshold, which may be poorly defined.

In general, the classifier performs better when trained on other subjects within the same study than when trained on subjects from another study. High classification performances with balanced accuracies of 93% and 80% for the Emotion and Cue data, respectively, were found in the within-study cases. Evaluation between studies, however, gave balanced accuracies of 74% and 62% when testing on Emotion and Cue data, respectively. Data from

more studies would probably help the between-study performance approach the within-study performance. Another way to improve the across-study performance could be to take into account the distributions of feature values in the test data set compared to the training data set.

Quality of ICA decomposition. Since the quality of an ICA decomposition depends on the pre-processing of data before running ICA, the usefulness of a classifier also depends on the pre-processing steps. If data is subjected to ICA with little pre-processing, many ICs are likely to be either mixed or noisy representations of individual classes. Since such ICs are difficult to classify, the performance of the classifier is likely to decrease. If ICs are truly mixed, classification into separate classes is not possible even for human experts. A future approach to tackling such cases could be to use the class probabilities given by MNR to decide how to handle mixed ICs. If, for example, an IC classified as mixed is also given somewhat high probabilities of representing blinks and lateral eye movements, the IC could be classified as being generally eye-related and discarded. On the other hand, mixed ICs could be retained if the probability of the neural class is above some pre-defined threshold.

5.3. Online capability

The reasonable performance of the classifier makes it possible to use it for online monitoring of artifact occurrence while recording EEG. A rule of thumb states that about $20 \times n^2$ samples are necessary to perform an ICA of n channels (Ullsperger and Debener, 2010). Hence an ICA and classification of resulting ICs can be performed every $20 \times n^2/f$ seconds, where f is the sampling rate. With 64 channels and a sampling rate of 512Hz, three minutes of recorded EEG provides sufficient data for an ICA decomposition. Using the *runica* algorithm in EEGLab (Delorme and Makeig, 2004) with at most 50 iterations, an ICA decomposition can be calculated in less than two minutes and calculating the features for an IC takes less than one minute. By distributing the feature calculations for the ICs over several threads, classified ICs can be provided online at a lag of about six minutes.

6. Conclusion

The presence of artifactual activity in EEG recordings is problematic in the analysis of data. While different approaches to removing such noise exist, these are either subjective and require lengthy manual processing of data or distinguish only between two classes. In this paper, we described an approach to automatic multi-class classification of artifactual ICs of EEG data. We considered neural ICs and five artifact classes: eye blinks, heartbeat, lateral eye movements, muscle, and mixed neural and artifactual activity. Using an initial pool of 65 spatial, spectral, and temporal features invariant to experimental setup, we investigated which features were important for classification of ICs. We found that features from all three spatial, spectral, and temporal domains carried information important for classification. However, we also saw that classification with a feature set consisting of only the spatial features had very similar performance to the 14-feature set when evaluating the classifier within studies. Across studies, the performance increased with the spatial feature set when testing on Cue data. The performance decreased when testing on Emotion data, but this was compared to the upwards biased performance estimate obtained with the 14-feature set

chosen based on Emotion data. The classifier generalizes well across subjects within studies, whereas across-study generalization is more challenging. Collapsing the multi-class classifications into binary classifications (artifact or neural), we obtain classification performances comparable to those found in previous studies both within and between studies (Bartels et al., 2010; Halder et al., 2007; LeVan et al., 2006; Mognon et al., 2010; Viola et al., 2009; Winkler et al., 2011). Thus the proposed classifier can be used for binary or multi-class classification interchangeably. The classification performance and speed of obtaining classified ICs allows online use of the classifier to detect artifacts while recording EEG so that subjects can be instructed to refrain from activity producing the detected types of artifacts. Although some artifacts such as the heartbeat are unavoidable, others may be mitigated in some paradigms, e.g. ERP studies, by the experiment being paused to allow subjects to blink or make them aware of muscle tension. Additionally, multi-class classification of artifactual ICs can make researchers aware of overly many artifacts of some class automatically. If possible, the experimental setup could then be redesigned to minimize the risk of such artifacts, e.g. by adjusting seating arrangements for participants to reduce eye and muscle tension. Additionally, the classifier could be used to identify artifacts typical of individual subjects in a short pilot run before performing an experiment.

We provide Matlab code for feature calculation and MNR classifiers trained on different feature sets online at `http://www2.imm.dtu.dk/~lffr/publications/IC_MARC.zip`. We hope that this will encourage others to further explore automatic classification of artifactual ICs and use this technique to ease data cleaning.

7. Acknowledgments

We would like to express our gratitude to Julie Onton, Klaus Gramann and Thomas Toellner for the use of their data (Gramann et al., 2010; Onton and Makeig, 2009), without which this study would not have been possible. Laura Frølich would also like to thank Scott Makeig for hosting her at the Swartz Center for Computational Neuroscience and for discussions about IC classification, and Christian Kothe for discussions and aid with programming. We would also like to thank two anonymous reviewers for their constructive comments which have improved the manuscript.

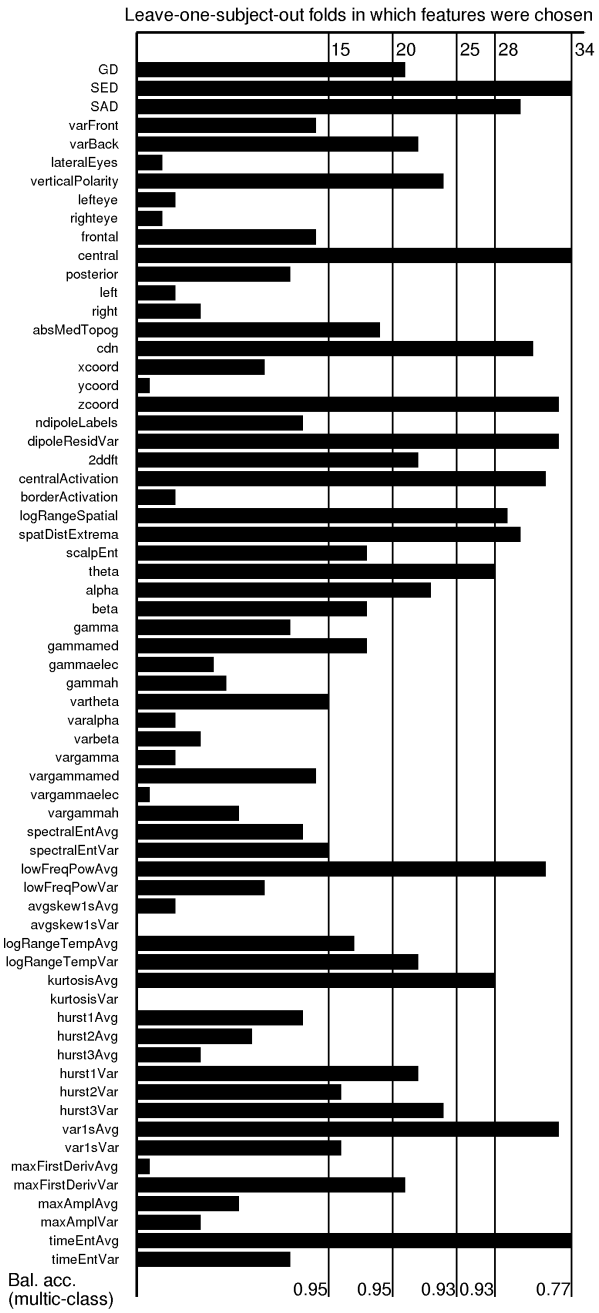


Figure 4: Barplot showing the number of folds each feature was chosen by forward selection in leave-one-subject-out CV on the emotion data (34 subjects).

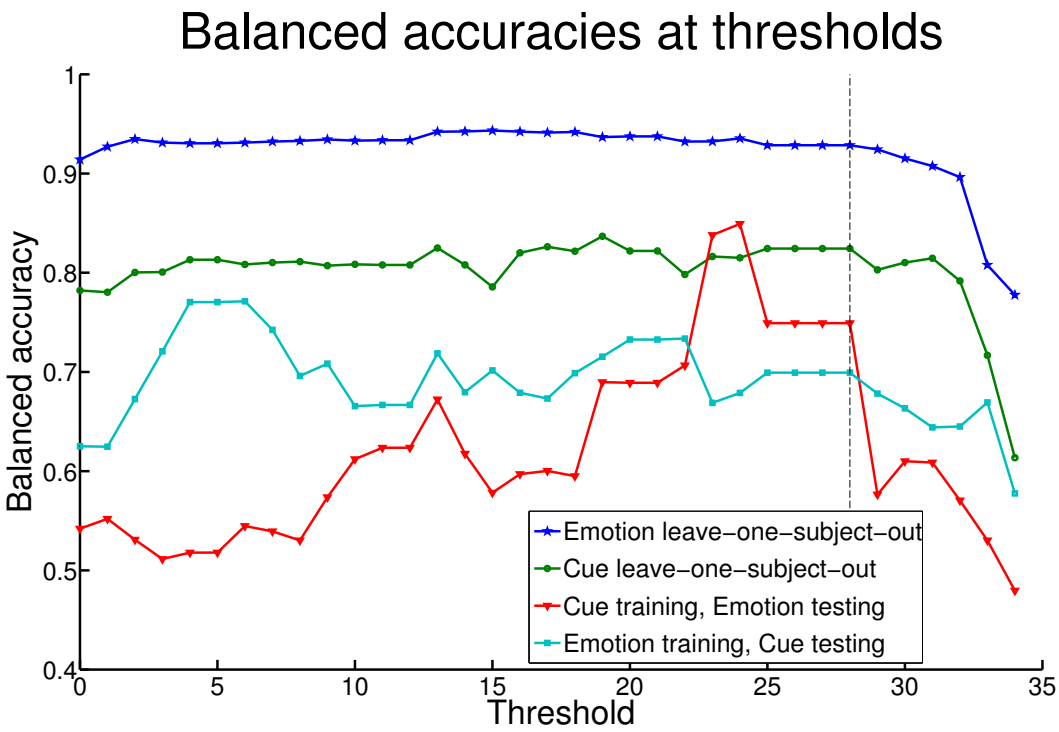


Figure 5: Balanced accuracy obtained with different feature sets constructed by varying the number of CV folds features must have been selected in to be included. The dashed line shows the cut-off chosen based on the blue curve. This choice led to a feature set containing 14 features.

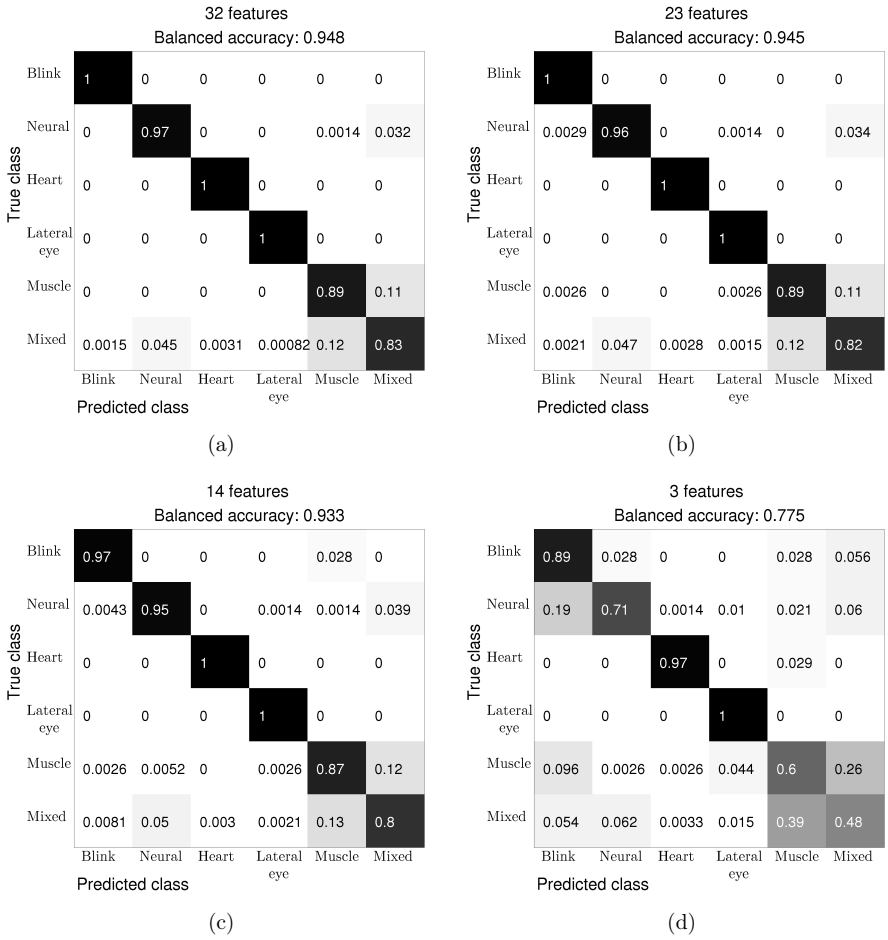


Figure 6: Confusion matrices and balanced accuracies of leave-one-subject-out classification performance on the Emotion data using the 32-feature set (a), the 23-feature set (b), the 14-feature set (c), and the 3-feature set (d). Black corresponds to higher and white to lower values. The balanced accuracies in the subtitles are the multi-class balanced accuracies.

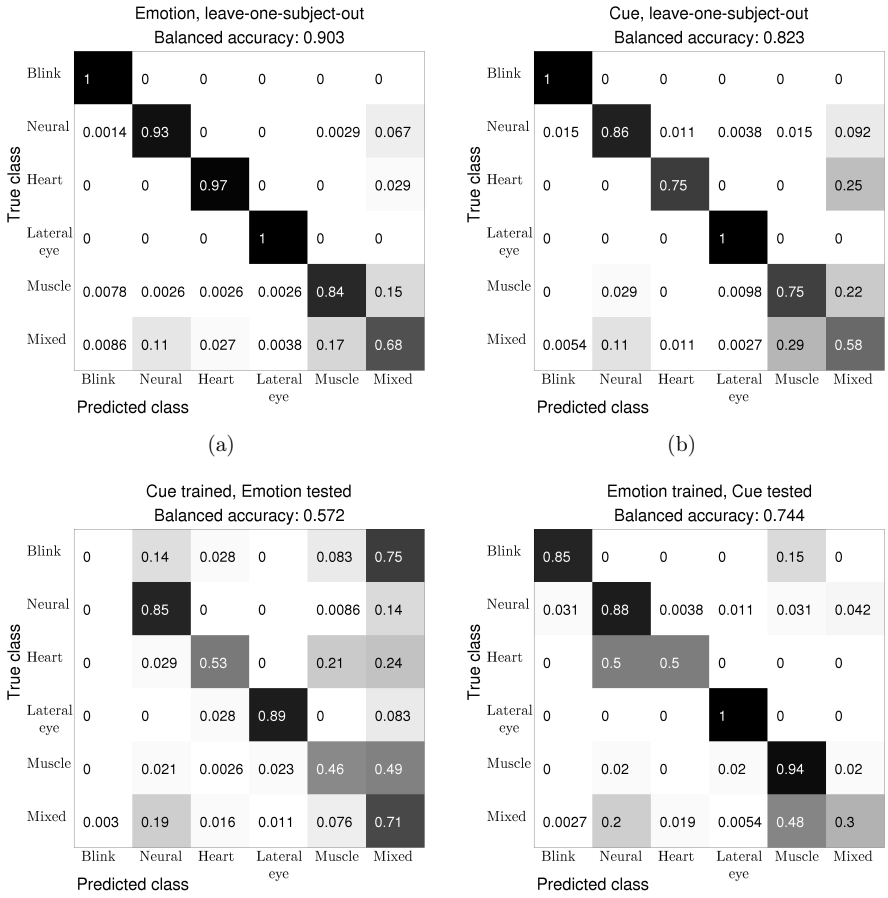
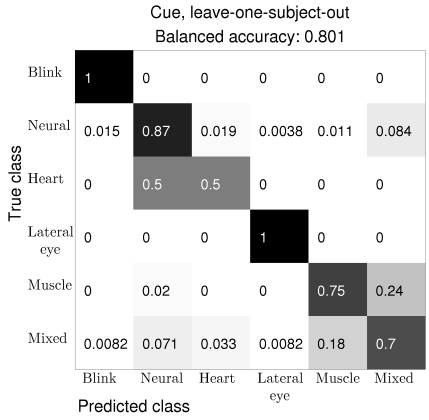
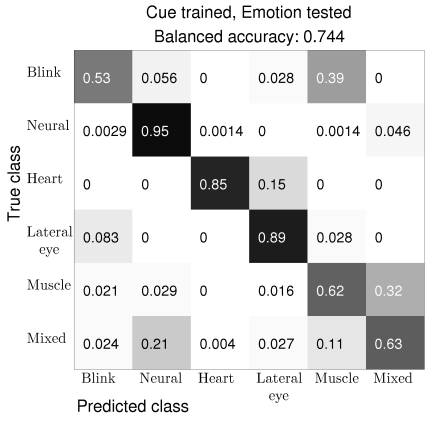


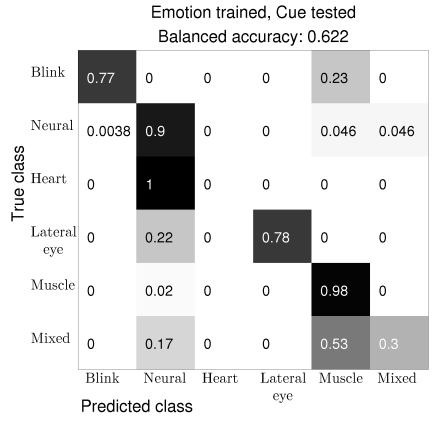
Figure 7: Confusion matrices for showing the classification performance using only the spatial features from the initial pool of features on leave-one-subject-out CV for the Emotion data set (a), leave-one-subject-out CV for the Cue data set (b), a model trained on Cue data and tested on all Emotion subjects (c), and a model trained on Emotion data and tested on all Cue subjects (d). Black corresponds to higher and white to lower values.



(a)



(b)



(c)

Figure 8: Confusion matrices showing the classification performance using the 14-feature set on leave-one-subject-out CV for the Cue data set (a), a model trained on Cue data and tested on all Emotion subjects (b), and model trained on Emotion data and tested on all Cue subjects (c). The balanced accuracies in the subtitles are the multi-class balanced accuracies. Black corresponds to higher and white to lower values.

References

Acar, Z.A., Makeig, S., Worrell, G.. Head modeling and cortical source localization in epilepsy. Conf Proc IEEE Eng Med Biol Soc 2008;2008:3763–3766.

Bartels, G., Shi, L.C., Lu, B.L.. Automatic artifact removal from eeg - a mixed approach based on double blind source separation and support vector machine. In: EMBC. 2010. p. 5383 –5386. doi:10.1109/IEMBS.2010.5626481.

Bigdely-Shamlo, N., Kreutz-Delgado, K., Kothe, C., Makeig, S.. Eyecatch: Data-mining over half a million eeg independent components to construct a fully-automated eye-component detector. In: IEEE Engineering in Biology and Medicine Conference, Osaka, Japan. 2013. .

Citi, L., Poli, R., Cinel, C.. Documenting, modelling and exploiting P300 amplitude changes due to variable target delays in Donchin's speller. J Neural Eng 2010;7(5):056006.

Comani, S., Mantini, D., Pennesi, P., Lagatta, A., Cancellieri, G.. Independent component analysis: fetal signal reconstruction from magnetocardiographic recordings. Comput Methods Programs Biomed 2004;75(2):163–177.

Delorme, A., Makeig, S.. Eeglab: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Meth 2004;134:9–21.

Delorme, A., Sejnowski, T., Makeig, S.. Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. Neuroimage 2007;34(4):1443 – 1449. URL: \url{http://www.sciencedirect.com/science/article/B6WNP-4MNHY2V-4/2/93de9223a58e55c80f19ecdb50e8dcfe}. doi:DOI:10.1016/j.neuroimage.2006.11.004.

Di, L., Rao, N., Kuo, C.H., Bhatt, S., Dogra, V.. Independent component analysis applied to ultrasound speckle texture analysis and tissue characterization. Conf Proc IEEE Eng Med Biol Soc 2007;2007:6524–6527.

Erfanian, A., Erfani, A.. ICA-based classification scheme for EEG-based brain-computer interface: the role of mental practice and concentration skills. Conf Proc IEEE Eng Med Biol Soc 2004;1:235–238.

Fatourechi, M., Bashashati, A., Ward, R.K., Birch, G.E.. Emg and eog artifacts in brain computer interface systems: A survey. Clin Neurophysiol 2007;118(3):480 – 494. URL: http://www.sciencedirect.com/science/article/pii/S1388245706015124. doi:10.1016/j.clinph.2006.10.019.

Gramann, K., Tollner, T., Muller, H.J.. Dimension-based attention modulates early visual processing. Psychophysiology 2010;47:968–978.

Halder, S., Bensch, M., Mellinger, J., Bogdan, M., Kubler, A., Birbaumer, N., Rosenstiel, W.. Online artifact removal for brain-computer interfaces using support vector machines and blind source separation. Comput Intell Neurosci 2007;:82069.

Hastie, T., Tibshirani, R., Friedman, J.. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. 2nd ed. Springer Series in Statistics. New York: Springer, 2009.

He, P., Wilson, G., Russell, C.. Removal of ocular artifacts from electro-encephalogram by adaptive filtering. Med Biol Eng Comput 2004;42(3):407–412.

Hyvärinen, A., Oja, E.. Independent component analysis: algorithms and applications. Neural Networks 2000;13(4-5):411 – 430.

Jayalakshmi, T., Santhakumaran, D.. Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering 2011;3(1):1793–8201.

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

Kim, C.M., Park, H.M., Kim, T., Choi, Y.K., Lee, S.Y.. FPGA implementation of ICA algorithm for blind signal separation and adaptive noise canceling. IEEE Trans Neural Netw 2003;14(5):1038–1046.

Klekowicz, H., Malinowska, U., Piotrowska, A.J., Wolynczyk-Gmaj, D., Niemcewicz, S., Durka, P.J.. On the robust parametric detection of EEG artifacts in polysomnographic recordings. Neuroinformatics 2009;7(2):147–160.

Kong, W., Vanderburg, C.R., Gunshin, H., Rogers, J.T., Huang, X.. A review of independent component analysis application to microarray gene expression data. BioTechniques 2008;45(5):501–520.

LeVan, P., Urrestarazu, E., Gotman, J.. A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification. Clin Neurophysiol 2006;117(4):912–27.

López, V., Fernández, A., Moreno-Torres, J.G., Herrera, F.. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. Expert Systems with Applications 2012;39(7):6585 – 6608. URL: http://www.sciencedirect.com/science/article/pii/S0957417411017143. doi:10.1016/j.eswa.2011.12.043.

Mantini, D., Franciotti, R., Romani, G.L., Pizzella, V.. Improving MEG source localizations: an automated method for complete artifact removal based on independent component analysis. Neuroimage 2008;40(1):160–173.

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;48(2):229–240. URL: http://dx.doi.org/10.1111/j.1469-8986.2010.01061.x. doi:10.1111/j.1469-8986.2010.01061.x.

Nolan, H., Whelan, R., Reilly, R.B.. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. J Neurosci Methods 2010;192:152–162.

Onton, J.A., Makeig, S.. High-frequency broadband modulation of electroencephalographic spectra. Front Hum Neurosci 2009;3(0):12. URL: \url{http://www.frontiersin.org/Journal/Abstract.aspx?s=537&name=humanneuroscience&ART_DOI=10.3389/neuro.09.061.2009}. doi:10.3389/neuro.09.061.2009.

Sugiyama, M., Kawanabe, M.. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation. 1st ed. Adaptive Computation and Machine Learning series. Cambridge, Massachusetts, USA: MIT Press, 2012.

Tangermann, M., Winkler, I., Haufe, S., Blankertz, B.. Classification of artifactual ICA components. Int J Bioelectromagnetism 2009;11(2):110–114.

Tsai, D.M., Lai, S.C.. Independent component analysis-based background subtraction for indoor surveillance. IEEE Trans Image Process 2009;18(1):158–167.

Ullsperger, M., Debener, S., editors. Simultaneous EEG and fMRI: Recording, Analysis, and Application; New York: Oxford University Press. p. 121–135.

Viola, F.C., Thorne, J., Edmonds, B., Schneider, T., Eichele, T., Debener, S.. Semi-automatic identification of independent components representing EEG artifact. Clin Neurophysiol 2009;120(5):868–77.

Winkler, I., Haufe, S., Tangermann, M.. Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals. Behav Brain Funct 2011;7:30.

Ypparila, H., Nunes, S., Korhonen, I., Partanen, J., Ruokonen, E.. The effect of interruption to propofol sedation on auditory event-related potentials and electroencephalogram in intensive care patients. Crit Care 2004;8(6):R483–490.

Zadrozny, B., Langford, J., Abe, N.. Cost-sensitive learning by cost-proportionate example weighting. In: Proceedings of the Third IEEE International Conference on Data Mining. Washington, DC, USA: IEEE Computer Society; ICDM '03; 2003. p. 435–. URL: http://dl.acm.org/citation.cfm?id=951949.952181.

Appendix A. Features

The 65 features of ICs used for classification are listed here. All these features are sign-invariant since the sign-ambiguity of spatial maps and time series of ICs cannot be resolved through normalization.

Appendix A.1. Spatial

- (GD) Generic discontinuity measure Mognon et al. (2010). This measure as used in ADJUST Mognon et al. (2010) is defined as

$$\max_n \left| a_n - \frac{1}{c-1} \sum_{m \neq n} \exp(-\|y_m - y_n\|) a_m \right|,$$

where y_m is the location of the m^{th} electrode on the scalp, a_m is the activation of the m^{th} electrode by the IC, and c is the number of electrodes. Hence this measure gives a high value if the IC activates any electrode a lot more than the neighboring electrodes, indicative of e.g. a loose electrode.

We use a slightly modified version of this measure to make the second term a weighted average. Our measure is defined as

$$\max_n \left| a_n - \frac{1}{\sum_{m \neq n} \exp(-\|y_m - y_n\|)} \sum_{m \neq n} \exp(-\|y_m - y_n\|) a_m \right|.$$

- (SED) Spatial eye difference Mognon et al. (2010). Absolute value of the difference between activation of electrodes around the left and right eye areas. The left eye area is defined to lie between the angles -61^o and -35^o with a radius larger than 0.3 (where the head radius is assumed to be one, the convention in EEGLab). The right eye area is defined to lie between the angles 34^o and 61^o, also at a radius larger than 0.3. Zero degrees is towards the nose and positive 90^o is at the right ear.

- (SAD) Spatial average difference Mognon et al. (2010). This feature is defined as the absolute value of the mean of frontal electrode activations minus the absolute value of the mean of posterior electrode activations. The frontal area is defined to be the electrodes with absolute angles less than 60^o and radii larger than 0.4. The posterior area consists of the electrodes with absolute angles larger than 110^o.

- (varFront and varBack) Variance of activation of frontal and posterior electrodes Mognon et al. (2010).

- (lateralEyes) Absolute value of the difference between activation of electrodes around the left and right eye areas. The left eye area is defined as the mean over all electrodes, weighted by a Gaussian bell with center at the location of Fp1 in the 10-20 electrode system. The right eye area is defined as the mean over all electrodes, weighted by a Gaussian bell with center at the location of Fp2 in the 10-20 electrode system. The standard deviation of both Gaussian bells is set to be 1 cm and a head radius of 9 cm is assumed.

- (verticalPolarity) Absolute value of the difference between activation of frontal and posterior electrodes. The frontal area is defined as the mean of all electrodes weighted by a Gaussian bell centered at the location of AFz in the 10-20 electrode system. The posterior area is defined as the mean of all electrodes weighted by a Gaussian bell centered at the location of POz in the 10-20 electrode system. The standard deviation of both Gaussian bells is set to be 2 cm and a head radius of 9 cm is assumed.

- (lefteye, righteye, frontal, central, posterior, left, right) These features give the absolute values of the mean activations of electrodes in various areas of the scalp. Each area is defined as the mean over all electrodes, where the contribution from each electrode to the mean is weighted by a Gaussian bell. For the areas around the eyes (lefteye and righteye), the standard deviation of the Gaussian bell is 1 cm. For all other areas, it is 2 cm. A 9 cm radius of the scalp is assumed. The Gaussian bells are centered at the locations of Fp1, Fp2, AFz, Cz, POz, C5, and C4, respectively.

- (absMedTopog) The absolute value of the median of the values in the scalp map.

- (cdn) Current density norm Winkler et al. (2011). The current density norm is a measure of the complexity of the current source distribution of an IC. A high complexity of the current source distribution indicates that the source of the IC is difficult to locate inside the brain, and thus that it is likely to be an artifact. This was one of the six final features included in the classifier described in Winkler et al. (2011), in which a more detailed description can be found.

- (xcoord, ycoord, and zcoord) X, Y, and Z coordinates of dipole fit Winkler et al. (2011). The dipole fit used returns a single dipole.

- (ndipoleLabels) Number of anatomical areas associated with dipole fit.

- dipoleResidualVariance

- (2ddft) Average logarithm of band power in high frequencies of spatial pattern Winkler et al. (2011).

- (centralActivation) Logarithm of mean of absolute values of activations of central electrodes of IC Winkler et al. (2011).

- (borderActivation) Binary feature to detect scalp maps with highest activity at an edge of the pattern. The most active electrode is the electrode for which the IC has the highest absolute value of activation. If the most active electrode in the pattern is in an outer group of electrodes, the feature is defined to be 1. Also, if the local maximum of an outer group is at the edge of the group, and its activation differs by more than two standard deviations from the group mean, then the feature is defined to be 1, too. Otherwise, it is defined to be -1 Winkler et al. (2011).

- (logRangeSpatial) Logarithm of range of activation of electrodes. This was one of the six final features included in the classifier described in Winkler et al. (2011).

- (spatDistExtrema) Euclidean distance in 3D coordinates between the two electrodes with minimal and maximal activation.

- (scalpEntropy) The entropy of the scalp map.

Appendix A.2. Spectral

- (theta, alpha, beta, gamma, gammamed, gammaelec and gammah) Mean over one-second intervals of the logarithm of band power in the θ (4-7Hz), α (8-13Hz), β (13-20Hz), lower γ (21-30Hz), middle γ (30-45Hz), γ around the power grid frequencies (both US and European) (46-65Hz), and higher γ (66-80Hz) bands. The average band power in the α-band was one of the six final features included in the classifier described in Winkler et al. (2011).

- (vartheta, varalpha, varbeta, vargamma, vargammamed, vargammaelec and vargammah) The variance over one-second intervals of the logarithm of the bandpower in the same bands as mentioned above.

- (spectralEntropyAvg and spectralEntropyVar) The entropy of the power distribution over the bands mentioned above is calculated for one-second intervals of the time series. The feature spectralEntropyAvg is then the average over these one-second intervals, while spectralEntropyVar is the variance of the spectral entropy over the one-second intervals.

- (lowFrequentPowerAvg and lowFrequentPowerVar) These features give the band power in the δ band (1-3Hz) relative to the total power in the time series. The spectrogram used for these features is calculated based on the downsampled but un-filtered time series since the filter removes frequencies lower than 3Hz. The spectrogram is calculated over one-second intervals, and the power in the δ band divided by the power over all frequencies is then found. The feature lowFrequentPowerAvg is the mean over the one-second intervals of this ratio, and lowFrequentPowerVar is the variance over the one-second intervals.

Appendix A.3. Temporal

- (skew1sAvg and skew1sVar) The skewness was calculated for one-second intervals of the time series of ICs. The feature skew1sAvg is the average over these one-second intervals and skew1sVar is the variance over these intervals. The feature skew1sAvg for 15 second intervals was one of the six final features included in the classifier described in Winkler et al. (2011).

- (logRangeTemporalAvg and logRangeTemporalVar) The range (maximum value minus minimum value) was calculated for one-second intervals. The feature logRangeTemporalAvg is the average over these one-second intervals and logRangeTemporalVar is the variance.

- (kurtosisAvg and kurtosisVar) As for the two above features, the feature kurtosisAvg is the average of the kurtosis in one-second intervals and kurtosisVar is the variance of the kurtosis in one-second intervals. This was also used in Winkler et al. (2011).

- (hurst1Avg, hurst2Avg, hurst3Avg, hurst1Var, hurst2Var and hurst3Var) We used the Matlab function *wfbmesti* in the Wavelet toolbox to get three different estimates of the Hurst exponent, which is a measure of the autocorrelation of a time series. These

three estimates of the Hurst exponent are found for one-second intervals. The features hurst1Avg, hurst2Avg, and hurst3Avg are the averages over these intervals, and hurst1Var, hurst2Var, and hurst3Var are the variances over the intervals.

- (var1sAvg and var1sVar) Again, the variance is found in one-second intervals of the time series. The features var1sAvg and var1sVar are the average and variance over these intervals, respectively. This was also used in Winkler et al. (2011).

- (maxFirstDerivAvg and maxFirstDerivVar) In each one-second interval, the maximum difference between consecutive values was found. The average over the intervals is maxFirstDerivAvg and the variance is maxFirstDerivVar. This was also used in Winkler et al. (2011).

- (maxAmplAvg adn maxAmplVar) In each one-second interval, the maximum amplitude (maximum absolute value in that interval) was found. The average over these intervals is maxAmplAvg and the variance is maxAmplVar. This was also used in Winkler et al. (2011).

- (timeEntropyAvg and timeEntropyVar) In each one-second interval, the entropy was found. The average over these intervals is timeEntropyAvg and the variance is timeEntropyVar. This was also used in Winkler et al. (2011).

Appendix B

Investigating effects of different artefact types on Motor Imagery BCI

Investigating effects of different artefact types on Motor Imagery BCI

Laura Frølich[1], Irene Winkler[2], Klaus-Robert Müller[3], *Member, IEEE*, and Wojciech Samek[4], *Member, IEEE*

***Abstract*— Artefacts in recordings of the electroencephalogram (EEG) are a common problem in Brain-Computer Interfaces (BCIs). Artefacts make it difficult to calibrate from training sessions, resulting in low test performance, or lead to artificially high performance when unintentionally used for BCI control. We investigate different artefacts' effects on motor-imagery based BCI relying on Common Spatial Patterns (CSP). Data stem from an 80-subject BCI study. We use the recently developed classifier IC_MARC to classify independent components of EEG data into neural and five classes of artefacts. We find that muscle, but not ocular, artefacts adversely affect BCI performance when all 119 EEG channels are used. Artefacts have little influence when using 48 centrally located EEG channels in a configuration previously found to be optimal.**

I. INTRODUCTION

Brain-Computer Interfaces (BCIs) allow a user to control a computer through his or her brain activity. The brain activity is often examined using electroencephalography (EEG) recordings, which offer a high temporal resolution and can be acquired with relatively low-cost, transportable equipment.

EEG signals show fluctuations of electrical acitivity as measured from electrodes placed on the scalp. These are also affected by electrical sources unrelated to brain activity, referred to as artefacts, which often produce larger potential differences than brain activity. Some artefacts are of physiological origin, such as eye movements, muscle contractions, the heartbeat etc. while others, such as loose electrodes and the power grid, are technical artefacts.

A. Motivation

An often cited goal of BCIs is to enable paralysed patients to communicate. Since healthy subjects are easier to recruit, development of BCIs is usually carried out on healthy subjects. If a BCI system developed on healthy subjects turns out to be controlled by artefacts, it will be of little use in patients. Even if a BCI system is developed for healthy subjects, artefacts may be problematic if the stimulus during training induces other artefacts than those from online use.

Some artefacts may affect BCI training more than others and the methods for remedying different artefacts' effects differ. By investigating artefacts' influence on BCIs we aim to identify those most detrimental to performance which can then be targeted to gain the largest improvements.

B. Previous work on artefacts' effects on BCIs

Only few studies have previously inspected the influence of artefacts on motor-imagery based BCIs. McFarland et al. inspected the presence of muscle artefacts in 10 BCI sessions of novices [1]. Muscle artefacts either caused or indicated frustration with lacking BCI control. Winkler et al. investigated the performance of a motor-imagery based BCI system as a function of the number of removed artefactual data dimensions [2]. No substantial decrease in performance was observed until fewer than 12 dimensions remained in training data. Others have proposed variations of Common Spatial Patterns (CSP) to cope with artefacts [3], [4], [5], [6], [7]. To the best of our knowledge, no study has previously attempted to quantify the influence of different types of artefacts on motor-imagery based BCI.

C. Aim and research questions

We wish to learn how various artefact types affect motor-imagery based BCI systems. Using data from an 80-subject BCI study, we applied Independent Component Analysis (ICA) to linearly transform EEG signals into a space of independent source components (ICs). We then used the recently developed multi-class classifier 'IC_MARC' to label each component as neural activity or as one of five artefact types (blinks, lateral eye movements, heartbeat artefact, muscle artefact, or mixed artefact) [8]. Mixed artefacts are artefacts that do not clearly belong to one of the other four artefact classes and may also include traces of neural activity. We answered the following research questions:

1) What types of artefacts are most common in training data (after automatic removal of noisy channels)?
2) Do participants use information contained in artefactual ICs to control the BCI system?
3) Does removing or regularising away from artefactual ICA directions improve BCI performance?
4) Do the answers for the above questions differ depending on whether all available EEG channels (119 channels) or only the 48 central channels found to be optimal by Sannelli et al. [9] are used?

*This work was supported by the Federal Ministry of Education and Research (BMBF) under the project Adaptive BCI (FKZ 01GQ1115) and by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education.

[1]Laura Frølich is with the Section for Cognitive Systems, DTU Compute, Technical University of Denmark, Matematiktorvet, Building 321, 2800 Kgs. Lyngby, Denmark lffr@dtu.dk

[2]Irene Winkler is with the Machine Learning Group, Technische Universität Berlin, Marchstr. 23, 10587 Berlin, Germany irene.winkler@tu-berlin.de

[3]Klaus-Robert Müller is with the Machine Learning Group, Technische Universität Berlin, Marchstr. 23, 10587 Berlin, Germany and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Republic of Korea klaus-robert.mueller@tu-berlin.de

[4]Wojciech Samek is with the Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany wojciech.samek@hhi.fraunhofer.de

II. Methods & Materials

A. Data

Data stem from Blankertz et al. [10], who recorded 80 BCI-novices in a classical motor-imagery paradigm. Subjects were paid 8 EUR per hour for participation [10]. Participants first performed motor imagery with the left hand, right hand and both feet in a training measurement. Every 8 s, the requested BCI task of the current trial was indicated by a visual cue. Following calibration of the system, the test data were recorded using the two classes that provided best discrimination. Participants controlled a 1D cursor application. For the training data 75 trials for each motor condition were recorded, while the test data contained 150 trials from each condition. All BCI performance tests were performed on test data for each participant, while ICA demixing and training of the BCI-classifier were based on calibration data.

EEG data were recorded from 119 electrodes placed according to the extended 10-20 system at a frequency of 1000 Hz. For our offline re-analysis, data were band-pass filtered between 8-30 Hz. Epochs were defined as 0.75-3.5 s after event markers. In the training data, channels with excessively low or high variance were automatically rejected.

B. Determining effects of artefacts on BCI performance

1) Common Spatial Patterns: Common Spatial Patterns is a standard feature extraction method for motor-imagery based BCIs [11]. CSP extracts spatial filters as linear channel combinations, $\mathbf{w}$, for which the variance differs most between conditions. Formally, CSP filters are the eigenvectors corresponding to the largest (and smallest) eigenvalues λ of the generalized eigenvalue problem $C_1\mathbf{w} = \lambda C_2\mathbf{w}$, found as:

$$\underset{\mathbf{w}}{\operatorname{argmax}} \frac{\mathbf{w}^T C_1 \mathbf{w}}{\mathbf{w}^T C_2 \mathbf{w}}. \tag{1}$$

The channel $\times$ channel matrix C_i is the average of covariance matrices from condition i trials. We used the filters from the three highest and lowest eigenvalues for classification.

2) Automatic classification of independent components: For each subject, we ran an ICA on the concatenated training data epochs. We used the extended Infomax algorithm in EEGLab [12] to extract enough ICs to account for 99.9% of data variance. Each IC consists of its time course and a spatial pattern which expresses the IC's influence on scalp electrodes. Subsequently, we used the previously developed automatic classifier "IC_MARC" to classify ICs [8].

IC_MARC uses multinomial regression to assign probabilities to ICs of belonging to each of six classes (blinks, lateral eye movements, electrical heartbeat, muscle, neural, or mixed artefact). We used features of the scalp maps for classification. This is, to the best of our knowledge, the only existing classifier allowing distinction between both ocular and muscular artefacts. Most other classifiers can distinguish between different ocular, but not muscular artifacts (e.g. [13], [14]), or cannot be used in a multi-class setting.

ICs were classified as belonging to the class for which the highest probability was predicted, except if the highest probability was for an ocular artefact class and that probability was less than 80%. Such ICs were classified as mixed. Fig. 1 shows patterns from ICs classified by IC_MARC.[1]

For the analysis presented here, we consider three groups of artefactual ICs: 1) muscle artefacts, 2) ocular artefacts (eye blink and horizontal eye movements), and 3) all non-neural components (eye blink, electrical heartbeat, lateral eye movement, muscle, and mixed artefacts).

3) EEG channel configuration: If only central channels are kept it is likely that some artefacts become less pronounced or disappear, as e.g. muscle artefacts affect outer electrodes most (see Fig. 1). Since artefacts may affect electrode configurations differently we analysed both the full electrode configuration and the electrode configuration found to be optimal by Sannelli et al. that consists of 48 centrally located electrodes [9].

4) BCI performance on artefactual and non-artefactual data: We applied CSP to the activity contained in artefactual ICs to quantify the amount of class-discriminative information in artefacts. We also investigated the BCI performance when different groups of artefacts were projected out.

5) BCI performance when artefacts are regularised against: Since artefactual ICs may contain traces of neural activity, we might expect CSP performance to increase when we regularise against artefactual directions instead of completely removing them. This should allow the CSP algorithm to find spatial filters in the artefactual directions if there is enough class-discriminative information to warrant this.

By introducing a channel $\times$ channel regularisation matrix K (and a regularisation parameter $\lambda \in \mathbb{R}$) in the CSP objective as follows, spatial filters that cause large variance along the directions of K are discouraged [15]:

$$\underset{\mathbf{w}}{\operatorname{argmax}} \frac{\mathbf{w}^T C_1 \mathbf{w}}{\mathbf{w}^T((1-\lambda)C_2 + \lambda K)\mathbf{w}}. \tag{2}$$

To regularise against artefactual directions, normalised patterns of artefactual ICs were collected as columns in a matrix, A_{art}. Analyses not reported here showed no significant difference in performance between making patterns or time series of ICs have norm one.

The penalty matrix K was set equal to $A_{art}A_{art}^T$ to find spatial filters $\mathbf{w}$ such that $||\mathbf{w}^T A_{art}||$ is minimal, where $||\cdot||$ denotes the euclidean norm. This choice can be understood by looking at the ICA decomposition of the EEG data X, given as $X = A_{art}S_{art} + A_{neuro}S_{neuro}$, where S contains the time courses of ICs in rows and the subscript $neuro$ denotes neural ICs. The source activity extracted by a spatial filter $\mathbf{w}$, given as $\mathbf{w}^\top X$, contains minimal contributions from artefactual activity S_{art} if $||\mathbf{w}^T A_{art}||$ is minimized. (For more information on the interpretation of patterns and filters we refer the reader to [16].)

For each subject, the regularisation parameter λ was chosen in a five-fold cross-validation on calibration data from the values 0, 2^{-16}, 2^{-15}, ..., 2^{-1}, 0.6, 0.7, ..., 1.

[1]Except for the heartbeat class, the examples are good demonstrations of what one would expect in each class. Difficulty with the heartbeat class was also found during the development of IC_MARC and CORRMAP [8], [13].

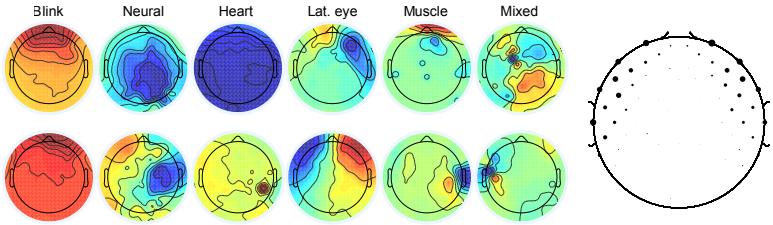


Fig. 1: Left: Examples of patterns of automatically classified ICs. Right: Locations of most active electrode in muscle ICs from all subjects. Dot sizes represent the number of times electrodes were the most active in muscle ICs.

III. Results

A. Most common artefacts

Mixed and muscle artefacts were the most and second most common artefact classes, respectively. Using all channels, out of 6428 (range over subjects: 39-107) ICs, 33 (0-14) were classifed as blinks, 1854 (4-43) as neural, 57 (0-4) as heartbeats, 80 (0-9) as lateral eye movements, 1773 (8-45) as muscular, and 2631 (5-86) as mixed. On the 48 channels, out of 2925 (range: 22-45) ICs, 7 (0-4) were classifed as blinks, 1320 (6-25) as neural, 21 (0-4) as lateral eye movements, 276 (0-12) as muscular, and 1301 (5-33) as mixed.

B. Class-discriminative information in artefacts

Using the Wilcoxon signed rank test we found that error rates significantly differed from chance (50%) when CSP was trained on muscular or all non-neural ICs ($p < 0.0001$, both channel configurations). When trained on ocular artefacts, the performance did not differ from chance (p-values of 0.39 and 0.75 for the all- and 48-channel configurations, respectively). This shows that only the muscle and non-neural artefact groups contain class-discriminative information.

We used a sign test to compare the performance for each subject when muscle artefacts were removed to the baseline by looking at whether each trial was correctly or incorrectly classified. On the full channel configuration, the performance of 19 subjects significantly changed when muscle artefacts were removed, 6 getting worse. On the 48-channel configuration the performance of 17 subjects changed, 9 getting worse. When removing all non-neural ICs, the performance of 12 and 17 subjects significantly decreased on the all-channel and 48-channel configurations while 10 and 12 subjects improved on the two configurations, respectively.

C. Does removing or regularising away from artefactual ICA directions improve BCI performance?

Table I shows error rates obtained from baseline CSP, CSP trained on non-artefactual activity, and CSP with artefact regularisation for all three artefact groups. Significance tests were calculated using the Wilcoxon signed rank test. Since subjects with the same performance in two methods are not included in the comparision, some differences between medians may be higher than others without showing corresponding significance. On the full channel configuration, the only significant difference from baseline CSP was obtained when regularising against muscle ICs, which improved performance. Removing muscle ICs did not result in a significant difference from the CSP baseline although the median performance was better than that obtained with regularisation. This shows that regularising gives a more consistent improvement across subjects. With regularisation, however, artefactual activity could still be used to gain artificially high levels of BCI control.

On the 48-channel configuration muscle artefacts were not as prominent, which is reflected by the lack of performance improvement with regularisation against muscle ICs. In line with the observation that ocular artefacts did not contain class-discriminative information for either channel configuration, we observed that regularising against or removing ocular ICs did not significantly impact performance.

Fig. 2 shows the relationship between improvements in performance when removing non-neural ICs and the CSP classification performance when training only on those non-neural ICs, on the 48-channel configuration. A higher error rate from training on artefacts implies less class-discriminative information in the artefacts. Hence removing such artefacts should make the neural signal clearer without removing class-relevant data. This is indeed what the figure shows since the improvement with artefact removal increases with the error rate from training on artefacts.

When artefacts contain class-discriminative information it could be due to traces of neural activity in the artefacts or to the user employing artefacts to control the BCI. Fig. 3 shows

TABLE I: Error Rates

	Muscle	Ocular	All non-neural
48 channels			
CSP	———	28.25 (1.7)	———
CSP no artefacts	27.17 (1.7)	27.67 (1.7)	28.50* (1.7)
CSP IC regularised	29.50* (1.7)	29.17 (1.7)	28.83* (1.7)
All channels			
CSP	———	31.75 (1.8)	———
CSP no artefacts	29.08 (1.8)	32.33 (1.8)	35.00 (1.7)
CSP IC regularised	31.42* (1.8)	33.50 (1.8)	32.00 (1.8)

Median error rates over 80 subjects from baseline CSP, CSP trained on non-artefactual activity, and CSP with artefact regularisation for three artefact groups (standard deviations in parentheses). * indicates differences from baseline CSP (p <0.05).

an example indicating that artefacts were used to control the BCI since performance decreased after artefact removal and muscular artefact contamination is seen in the last two CSP patterns before artefact removal but not after.

IV. Conclusion

We investigated the influence of different artefacts on motor-imagery based BCIs. Using data from an 80-subject BCI study and a recently proposed multi-class IC classifier, we found that muscle artefacts alone and all non-neural artefacts as a group have a small impact on the BCI system. In contrast, ocular artefacts alone had no significant influence, probably because eye artefacts mostly affect frequency ranges below those containing the motor-imagery μ-rhythm.

More specifically, we observed above-chance performance when CSP was trained on muscular or mixed artefacts, but not if trained on ocular artefacts. Up to 9 subjects used muscle artefacts to improve their BCI control. This may be problematic if healthy participants use artefacts to operate a BCI system which should be transferable to severely motor-impaired patients. However, we note that the overall contribution of muscle artefacts was significant, but small.

When removing artefacts, BCI performance did not improve, which is consistent with previously published analyses that used a different automatic classifier of ICs [2], [17]. Regularising against muscle artefacts significantly improved BCI performance when all available 119 channels were used but significantly impaired performance for the more suitable 48-channel configuration.

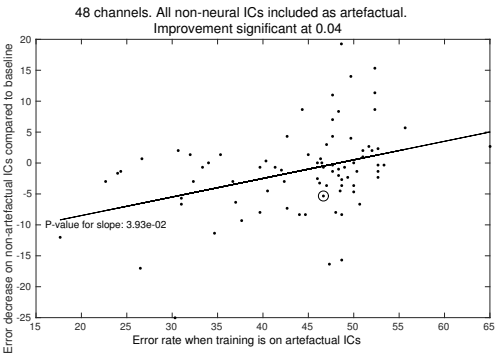


Fig. 2: Error decrease when CSP was run on neural ICs relative to the performance on artefactual ICs. The circled dot represents the subject whose CSP patterns are shown in Fig. 3.

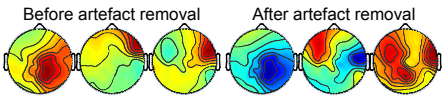


Fig. 3: The three most discriminative CSP patterns for class one from all data and neural ICs only on the all-channel configuration. Artefact removal increased the error rate from 11.67% to 17%.

We conclude that it is difficult to improve CSP performance on the 48-channel configuration by artefact processing. We conjecture that this difficulty may mainly arise from using the motor-imagery paradigm which relies on activity in the motor cortices, recorded from central scalp positions.

Acknowledgment

We thank the authors of [10] for providing data and Claudia Sannelli for advise on optimal electrode configurations.

References

[1] D. J. McFarland, W. A. Sarnacki, T. M. Vaughan, and J. R. Wolpaw, "Brain-computer interface (bci) operation: signal and noise during early training sessions," Clinical Neurophysiology, vol. 116, no. 1, pp. 56 – 62, 2005.

[2] 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.

[3] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," IEEE Reviews in Biomedical Engineering, vol. 7, pp. 50–72, 2014.

[4] M. Kawanabe, W. Samek, K.-R. Müller, and C. Vidaurre, "Robust common spatial filters with a maxmin approach," Neural Computation, vol. 26, no. 2, pp. 1–28, 2014.

[5] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," IEEE Transactions On Biomedical Engineering, vol. 58, no. 2, pp. 355–362, Feb 2011.

[6] B. Reuderink, "Robust brain-computer interfaces," Ph.D. dissertation, University of Twente, Enschede, The Netherlands, October 2011.

[7] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," Journal of Neural Engineering, vol. 9, no. 2, p. 026013, 2012.

[8] 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, Jan 2015.

[9] C. Sannelli, T. Dickhaus, S. Halder, E.-M. Hammer, K.-R. Müller, and B. Blankertz, "On optimal channel configurations for smr-based braincomputer interfaces," Brain Topography, vol. 23, no. 2, pp. 186–193, 2010.

[10] B. Blankertz, C. Sannelli, S. Halder, E. M. Hammer, A. Kübler, K. R. Müller, G. Curio, and T. Dickhaus, "Neurophysiological predictor of SMR-based BCI performance," Neuroimage, vol. 51, no. 4, pp. 1303–1309, Jul 2010.

[11] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," IEEE Signal Processing Magazine, vol. 25, no. 1, pp. 41–56, Jan 2008.

[12] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," Journal of Neuroscience Methods, vol. 134, pp. 9–21, 2004.

[13] F. C. Viola, J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener, "Semi-automatic identification of independent components representing EEG artifact." Clinical Neurophysiology, vol. 120, no. 5, pp. 868–77, 2009.

[14] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features," Psychophysiology, vol. 48, no. 2, pp. 229–240, 2010.

[15] B. Blankertz, M. K. R. Tomioka, F. U. Hohlefeld, V. Nikulin, and K.-R. Müller, "Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing," in Advances in Neural Information Processing 20. MIT Press, 2008, p. 2008.

[16] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," NeuroImage, vol. 87, no. 0, pp. 96 – 110, 2014.

[17] I. Winkler, S. Brandl, F. Horn, E. Waldburger, C. Allefeld, and M. Tangermann, "Robust artifactual independent component classification for bci practitioners," Journal of Neural Engineering, vol. 11, no. 3, p. 035013, 2014.

Appendix C

Brain-Computer Interfacing under Distraction: An Evaluation Study

Brain-Computer Interfacing under Distraction: An Evaluation Study

Stephanie Brandl[†], Laura Frølich[‡], Johannes Höhne[†], Klaus-Robert Müller[†¶], and Wojciech Samek[§]

[†]Berlin Institute of Technology, Marchstr. 23, 10587 Berlin, Germany
[‡]Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[¶]Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea
[§]Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany

E-mail: stephanie.brandl@tu-berlin.de, klaus-robert.mueller@tu-berlin.de, wojciech.samek@hhi.fraunhofer.de

Abstract. While motor-imagery based Brain-Computer Interfaces (BCIs) have been studied over many years by now, most of these studies have taken place in controlled lab settings. Bringing BCI technology into everyday life is still one of the main challenges in this research field. This paper systematically investigates BCI performance under 6 types of distractions that mimic out-of-lab environments. We report results of 16 subjects and show that the performance of the standard CSP+RLDA classification pipeline drops significantly in this "simulated" out-of-lab setting. We then investigate three methods for improving the performance: 1) artifact removal, 2) ensemble classification, and 3) a 2-step classification procedure. While artifact removal does not improve the BCI performance, both ensemble classification and the 2-step classification procedure significantly enhance the performance compared to the standard procedure.

1. Introduction

Brain-Computer Interfacing (BCI) [1] [2] allows non-muscular communication between a human being and a computer device by detecting a user's intents via brain signals, e.g. with an electroencephalogram (EEG), and translating them into control commands. This is particularly useful for people affected by diseases which lead to the loss of muscular control, such as amyotrophic lateral sclerosis (ALS), brainstem stroke, multiple sclerosis and especially those who suffer from locked-in syndrome. BCIs can be applied not only for communication but also to control external devices such as a wheelchair [3], for rehabilitation [4] and mental state monitoring [5].

Over the years, various improvements in BCI research have been presented. Integrating machine learning algorithms caused substantial reduction in calibration time [6, 7, 8, 9] which crucially enhanced usability of BCIs. Also, novel approaches in robust feature extraction

[10, 11, 12, 13, 14], artifact detection [15, 16, 17] and adaptive methods [18, 19, 20] led to great improvements especially with respect to reliability.
Those approaches already work well in controlled lab environments which are highly artificial and significantly differ from everyday life situations, where people have to handle various visual, auditory or other cognitive distractions. In order to fulfill its main purpose, to provide disabled people with a non-muscular communication pathway, BCI research has recently begun to go beyond lab environments. Since algorithms may not work in real world scenarios it becomes mandatory to investigate and enhance them.
First steps into the real world have been made [5]. Ambulatory BCIs, e.g., allow participants to walk indoors [21] or even outdoors [22] while using a P300 spelling device. Another study investigated the influence of speaking while performing motor imagery tasks [23]. Also several patient studies have been carried out on stroke and even locked-in patients [24, 25, 26]. However, there still lacks, to the best of our knowledge, a study where data is recorded in systematic out-of lab scenarios and evaluated in detail. In this paper we want to fill this gap by presenting and analyzing a motor imagery-based BCI study where 16 healthy participants were distracted in 6 different scenarios (including no distraction) while performing motor imagery tasks. With those distraction scenarios we intended to simulate a more realistic environment where participants e.g. listen to news, watch a flickering video, search the room for a particular number or handle vibro-tactile stimulation. The aim of this study was to investigate BCI performance in environments different from the training environment and analyze the problems that occur in such scenarios.
Note that this paper represents an extension of a preliminary version of this study [27]. In particular, we investigate standard machine learning techniques, commonly used in BCI research, in semi realistic scenarios. Three major problems which lead to poor performance in those out-of-lab scenarios are identified

(i) Artifact contamination
(ii) Feature Shifting
(iii) High cognitive workload

and then we discuss several new approaches which overcome those issues.
The rest of this paper is organized as follows: In the next section we present the BCI study. In the third section we evaluate BCI performance with standard machine learning techniques and discuss the problems that arose. In the fourth section we present novel approaches to tackle the identified problems before we summarize and discuss our findings in the Conclusion.

2. Experiments

This study investigates BCI performance in a semi-realistic environment where we considered everyday life situations such as watching TV or listening to news. Simulating those scenarios in-lab gives us the possibility to systematically analyze them and draw conclusions for future experiments.

Table 1: Overview over distraction tasks.

Condition	**Task**	**Purpose**
Clean	Motor imagery without distractions	Control Condition
Closed Eyes	Motor imagery with closed eyes	Investigation of α-rhythm
News	Motor imagery while listening to news sequences	Distraction + activation of auditory cortex
Numbers	Searching the room for one of the 26 letter-number combinations hanging on the wall while doing the motor imagery task	Distraction + muscular artifacts
Flicker	Motor imagery while watching a video with a flicker in gray shades at a frequency of 10Hz	Investigation of SSVEP
Stimulation	Motor imagery plus vibtro-tactile stimulation on both forearms with carrier frequencies of 50 and 100Hz, modulated at 9, 10 and 11Hz	Investigation of SSVSEP

2.1. Participants

We recorded EEG from 16 healthy participants (6 female; range: 22-30 years; mean age: 26.3 years) of which only three had previously participated in another BCI experiment. Since all the instructions were in German, a certain level of language proficiency was required. Three of the participants are members of the TU Berlin Machine Learning Group, whereas the other volunteers were paid for their participation.

2.2. Experimental Setup

The participants sat at a distance of 1m away from the 24" computer screen in an armchair. During the experiment the participants were wearing headphones to receive auditory instructions.
To record the EEG signals, we used a Fast'n Easy Cap (EasyCap GmbH) with 63 wet Ag/AgCl electrodes and placed them at symmetrical positions according to the international 10–20 system [28] with reference to the nose. We furthermore used two 32-channel amplifiers (BrainProducts) to amplify the signals, which were sampled at 1000 Hz.
Including breaks and preparation, each experimental session lasted about three hours, with signal recording taking about 90 minutes. Before starting the main experiment we conducted some baseline EEG recordings in which the participant had to alternately open and close both eyes for about 15 seconds with 4 repetitions each.
We divided the main experiment into 7 runs. Each run lasted about 10 minutes and consisted of 72 trials. Since the first run was used as a calibration phase, no distractions were added and no feedback was given. After the calibration phase, each run consisted of 12 trials per

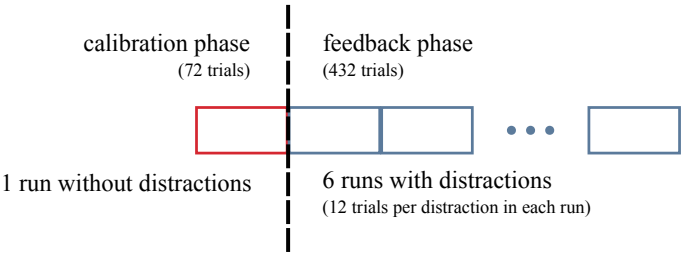


Figure 1: The experiment consists of 7 runs, each containing 72 motor imagery trails. The calibration run does not contain distractions, whereas each feedback run consists of 12 trials per distraction type (including the control condition).

distraction. Each trial lasted 4.5 seconds and included one motor imagery task. Auditory instructions in the form of *left* and *right* commands were given over the headphones at the beginning of each trial (since the experiment was conducted in German, the actual instructions were *links* and *rechts*). When the trial finished after 4.5 seconds, there was a *stop* command followed by a break of 2.5 seconds, after which the next trial started. Every three to four minutes the participant had the possibility to take a break.
To keep motivation levels high we included auditory feedback after the calibration phase. Therefore, Laplacian filters [29] of the C3 and C4 electrode were calculated and an LDA classifier (linear discriminant analysis) [9] was computed using the spectral power of the signals in a broad band (9-13Hz and 18-26Hz) as features. During the feedback phase, the classifier was applied to classify the motor imagery tasks and to provide auditory feedback. Due to the *closed eyes* task we could not give any feedback over the screen. This means that the *stop* command was followed by a *decision left* (*Entscheidung links*) or *decision right* (*Entscheidung rechts*) during the 2.5 seconds break.

2.3. Distractions

To study the effects of an increased cognitive load and additional artifacts, we included five distraction tasks in addition to the motor imagery task and a control condition in the experimental setup (see Table 1). We will now explain the design and motivation for those distractions.

(i) Clean
This condition serves as a control group. This means no distraction was added.

(ii) Closed Eyes
Participants performed the motor imagery task with closed eyes. Here, we investigate the effect of a more prominent alpha rhythm due to the closed eyes. Since the motor task related mu rhythm appears within a similar frequency band (8-13Hz), we expect an overlay with the alpha rhythm. Because of this task we gave all instructions and feedback over headphones instead of visually.

(iii) News
Sequences of a public newscast were played over the headphones containing current news and news from 1994. Each sequence was played once in each experiment, except for participant *od*, for whom some files were played twice. Here, we analyze the influence of the cognitive distraction and of an activated auditory cortex on the motor imagery performance. To make sure that the user still received the instructions, the volume was adapted for this task.

(iv) Numbers
For this task, 26 sheets of paper with a randomly mixed letter-number combination had been put up on the wall in front of the participant and also on the left and right side of the room. This means it was made necessary to turn the head in order to see the sheets. For each trial a new window appeared on the screen asking the participant to search the room for a particular letter to match with a stated number. The combinations were shown two to three times during each experiment. We counted the found letters and out of 72 trials, 59.7 combinations were found on average. This task investigates the effect of a high cognitive distraction and of additional muscular artifacts.

(v) Flicker
The participant watches a video with a flicker in gray shades alternating at a frequency of 10Hz. We included this task to analyze the influence of the *steady state visually evoked potential* (SSVEP) [30] .

(vi) Stimulation
We placed two vibration tactiles with a diameter of 3cm on the insides of both forearms, one over the wrist and another one just below the elbow. To investigate the interference of *steady state vibration somatosensory evoked potential* (SSVSEP) [31, 32] on the motor imagery task, vibratory stimulation was carried out with carrier frequencies of 50 and 100Hz, each modulated at 9, 10 and 11Hz.

2.4. Data Analysis

We downsampled the data to 100Hz and selected an individual frequency band between 5 and 35Hz for each experiment for the offline analysis. We also selected the time interval with highest discrimination individually. For feature extraction we used Common Spatial Patterns (CSP) [33] with three filters per class and trained a regularized linear discriminant analysis based classifier [34, 35, 9].

3. Evaluation of in-lab training

In this section, we present the results we obtained from classification based on the calibration data from the first run. That is, we trained the classifier on *clean* calibration data and tested it on all the distraction tasks (including the control group). Translating this to our systematic framework, we trained *in-lab* and tested in the *out-of-lab* setting. Since the

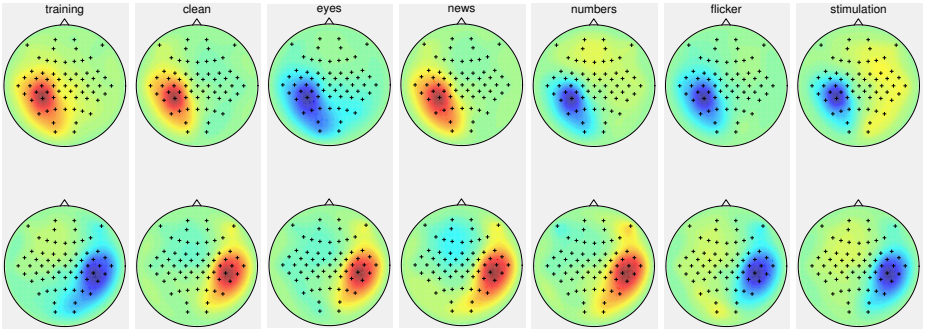


Figure 2: CSP Patterns for participant *od* for training and testing data

resulting classification rates lead to the assumption of poor out of lab BCI performance, we investigated possible reasons for this outcome.

3.1. Classification on clean training

Average classification rates are summarized in Table 2, exemplar CSP patterns for participant *od* are displayed in Figure 2. Classification accuracies vary much between participants [between 49.42% and 90.97%] but also within the experiments [*njz:* $45.83\% - 83.33\%$]. Most of the volunteers participated for the first time in a BCI experiment, so not everyone achieved classification rates significantly higher than chance level. Applying a binomial test ($\alpha = 0.05$) led to a threshold of 61.11% over which we could assume actual BCI control.
Out of the 16 participants, 3 of them did not reach that threshold in their best distraction task (*nkk, nkl, nkp*). Especially the *numbers* task which included searching the room and saying the letters out loud seem to have caused major difficulties for users to focus on the motor imagery task. Whereas most users gained their highest classification rates in the *flicker* task. So not all distraction tasks lead to lower classification rate, visual or auditory distraction seem to have less impact on BCI performance than e.g. additional muscular artifacts. Translating these findings to real-world scenarios means that it would be possible to watch TV or listen to music or news and use a BCI at the same time [36].

3.2. Why Poor Performance ?

However, in most experiments there is at least one task where the classification rate does not pass the threshold of 61.11%, so we still need to find out how to improve the overall performance. We found several possible explanations, one is that the distraction tasks influence the EEG recordings in a way that lead to major feature shifts between calibration data (*clean*) and testing data (with distractions). Some of the distraction might cause too many artifacts which could contaminate data in a way that makes it impossible to identify actual neural activity. It is also worth considering that some tasks are more cognitively demanding

Table 2: Mean classification accuracies for all distractions for subject *od*. One row represents one experiment, the first column shows the participant codes. For each experiment, the conditions with highest (bold) and lowest (red) performance rates are highlighted.

CSP	**overall**	**clean**	**eyesclosed**	**news**	**numbers**	**flicker**	**stimulation**
od	90.97	**95.83**	**95.83**	93.06	72.22	**95.83**	93.06
njy	60.42	62.50	54.17	65.28	50.00	**69.44**	61.11
njz	71.30	**83.33**	81.94	75.00	45.83	77.78	63.89
nkk	50.00	48.61	**55.56**	43.06	51.39	51.39	50.00
nkl	52.55	45.83	48.61	54.17	54.17	**61.11**	51.39
nkm	60.42	**68.06**	52.78	65.28	56.94	55.56	63.89
nkn	58.00	62.50	52.78	61.11	49.30	**65.28**	56.94
nko	82.13	93.06	83.33	80.56	62.50	**94.44**	78.87
nkp	51.62	51.39	**55.56**	50.00	50.00	52.78	50.00
nkq	63.26	**73.61**	59.72	61.97	47.89	66.67	69.44
nkr	61.11	63.89	61.11	62.50	51.39	**65.28**	62.50
nks	49.42	47.22	47.14	45.83	47.89	**54.17**	**54.17**
nkt	61.34	66.67	62.50	66.67	51.39	**70.83**	50.00
obx	82.87	88.89	87.50	81.94	70.83	**91.67**	76.39
nku	51.62	**61.11**	52.78	47.22	50.00	48.61	50.00
ma4	51.16	56.34	**58.33**	48.61	49.30	41.67	52.78
overall	62.39	**66.68**	63.10	62.64	53.81	66.41	61.53

than others such that some participants may not have been able to fully concentrate on the motor imagery task. This leads to three possible explanations for the poor BCI performance which we will tackle in the following sections.

- artifact contamination
- major feature shifts between distraction tasks
- participants are too distracted to focus on the motor imagery task, so the data is not separable

3.3. Artifacts

Since different artifact types influence data in different ways, the impact of their contamination and the methods to remove them also differ. We wanted to quantify the extent of contamination in each distraction task to investigate differences that might explain classification performances in the different conditions. Additionally, we investigated whether the removal of artifact groups could improve classification. We performed these analyses by decomposing the calibration data with Independent Component Analysis (ICA) and classifying the resulting independent components as different artifact types. We used

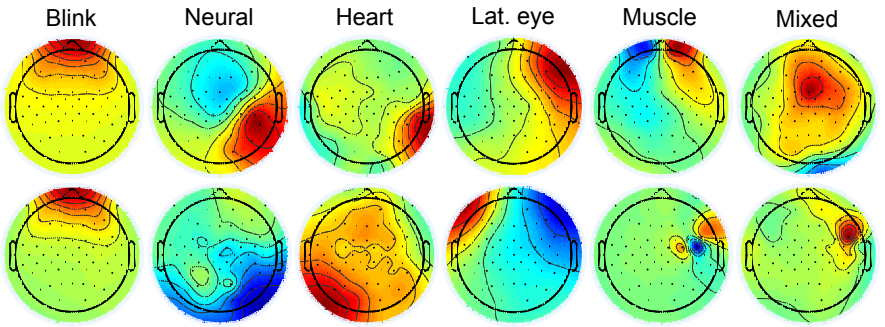


Figure 3: Examples of scalp maps of classified independent components from each class used by IC _MARC.

the implementation of Extended Infomax in EEGLab [37] to perform the ICA and an automatic classifier of independent components of EEG data, IC_MARC [38], to classify ICs. IC_MARC assigns a probability to each independent component of representing neural activity, eye blinks, heart beat artifacts, lateral eye movements, or muscle contractions. In addition to these five well defined classes, a class referred to as "mixed" is also included. This class contains artifact types other than those already mentioned and independent components that include several types of activity or are noisy. We classified independent components to the class for which the highest probability was predicted by IC_MARC. Figure 3 shows two randomly selected examples of each class from the *clean* condition. The samples from the blink, neural, heartbeat, and lateral eye movement classes are good examples of what we would expect in these classes. The top sample from the muscle class is more similar to what we would expect from the lateral eye class, while the bottom sample from the muscle class is typical of a muscular artifact. The two samples from the mixed class do not clearly belong to another class, as expected.

We grouped these artifact classes into five groups: (1) muscular artifacts, (2) ocular artifacts, (3) non-neural components, (4) muscle and mixed artifacts, and (5) non-mixed artifacts. Figure 4 shows the mean and standard deviations over subjects of percent data variance explained by each artifactual independent component group in each of the distractor conditions. The percent variance explained by the artifact groups is quite similar over the conditions, except for the *numbers* condition. In this condition, the ocular artifacts, non-neural independent components, and non-mixed artifacts explain more than twice as much data variance as in the other conditions. We might expect that a BCI system would perform similarly across distraction tasks whose artifact distributions are similar. If this is the case, one classifier should work well across the *non-number* distraction tasks. The *numbers* distraction task on the other hand requires a separate classifier. The artifact distribution of unseen data might be informative enough to distinguish between these two cases in order to select the appropriate classifier. This would make the 2-step approach described later a suitable method for out-of-lab BCI systems.

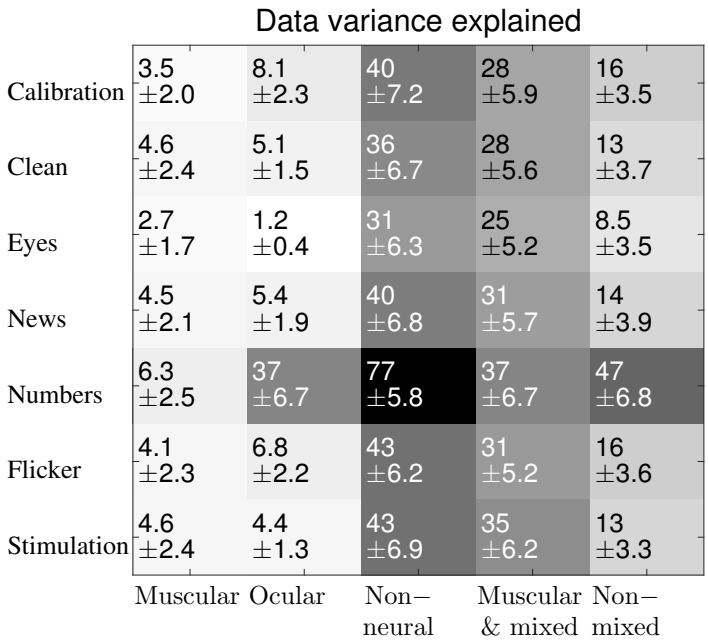


Figure 4: Percent data variance explained by each artifact group in each condition. The numbers give the mean over subjects plus/minus its standard deviation.

Figure 5 shows the median power of each independent component class as a function of frequency for the independent components from the *calibration* data. These independent components were the ones used to clean the other conditions in subsection 4.1. The power spectra were first calculated for each epoch using the default settings of the function *periodogram* in Matlab R2014b at 100 evenly spaced frequencies between 5 and 33 Hz. The median over epochs for each subject was then calculated, followed by the median over subjects for each independent component class. Since data was band-pass filtered during pre-processing, the spectra are flat below 8Hz and above 30Hz. It is reassuring that the neural independent components' power peaks at around 7-15 Hz since this band contains the motor-imagery signal's frequencies. The low amplitudes of the blink and lateral eye independent components' spectra relative to neural components is not surprising since these artifacts' activity tends to lie in frequency bands lower than 8 Hz (blinks' power peaks at 3 Hz and drops off before 10.5 Hz while lateral eye movements exhibit most power at frequencies below 6 Hz [39, p. 1237],[40]). Heart beats' activity mainly lies between about 15-32 Hz [41]. The high power seen at lower frequencies than 15 Hz for heart beats indicates that independent components classified as heart beat artifacts probably also contain other types of activity. Muscle artifacts are active at high frequencies, from about 20-300 Hz [42]. This expectation is reflected in the plot. Since mixed artifacts may contain many types of activity, we do not have any expectations for how the power curve for mixed components may look.

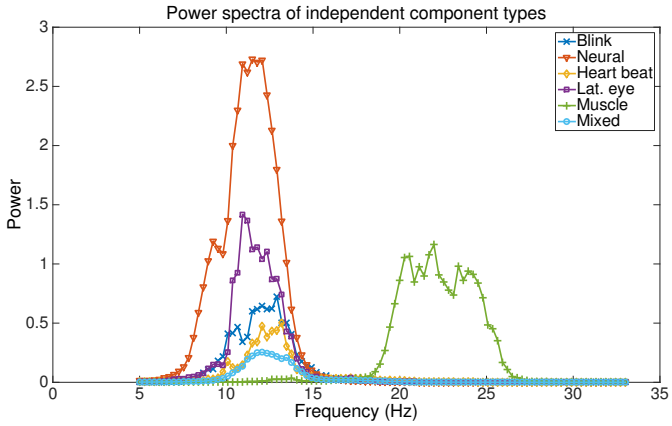


Figure 5: Power spectra of independent components from the six classes (blinks, neural, heart beats, lateral eye, muscle, and mixed). The power spectra were calculated for each epoch independently. Then the median was first taken over epochs for each subject, and then over subjects.

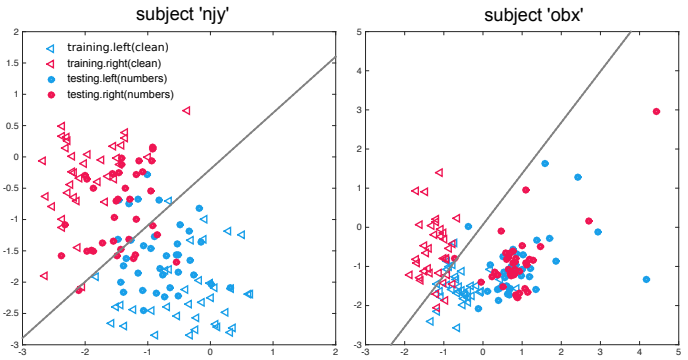


Figure 6: Features of participants *njy* and *obx* of the classifications between left and right hand motor imagery (two best CSP filters) where CSP was only trained on *clean* and tested on *numbers*.

3.4. Feature Shifts

In Figure 6 we plotted training and testing features for the *numbers* task (2 best CSP filters) for participants *njy* and *obx* respectively. Since we trained both tasks on the same calibration data, this plot shows how differently data shifts between training and testing. For participant *njy*, training features differ from testing features, but they are still separable. Whereas for participant *obx* test set features shift in a way that makes it impossible to separate them with the trained classifier. The corresponding classification rates (70.83% and 50%) support that

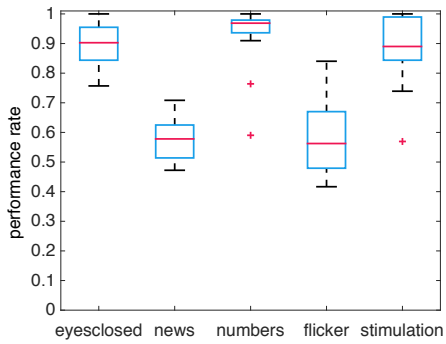


Figure 7: Mean classification accuracies across all 16 experiments under different distraction conditions against *clean* motor imagery for both hands.

finding.

We also classified the different distraction tasks against motor imagery without distractions (*clean*) with one CSP filter per condition for both hands. Average classification rates over all 16 experiments are visualized in form of boxplots in Figure 7. While classification rates for the *news* and *flicker* task against *clean* are mostly around chance level, one clearly sees that it is much easier to classify *stimulation, eyes closed* or *numbers* against *clean* where the median is around 90% and 95% accuracy, respectively. These results lead to the conclusion that task-related influences are much higher for the *stimulation* and *numbers* task than for the *news* and *flicker* task. This causes more discrimination between *numbers* and *clean* and therefore better classification rates between these classes.

This means that there are indeed major feature shifts in the data, especially in the *closed eyes*, *numbers* and *stimulation* tasks which significantly complicates classification. Including an adaptation step into the classification process could solve this problem if we assume that the data is separable at all.

3.5. Non-discriminativity

To find out whether the data is separable in general, we computed one classifier for each distraction task. This means we only tested on the same distraction task as we trained. We therefore separated the 72 trials that we had recorded into groups of 12 trials and computed classification rates via a 6-fold cross validation. So we computed a classifier for each group of 60 trials and tested on the remaining 12 trials respectively, repeating this concept for all 6 different tasks. Average classification rates are displayed in Table 3.

Comparing those with results from Table 2 where we computed one classifier for all tasks (see Section 3.1), the overall classification rates improved for most participants. While the overall classification rate for the *news* task hardly changed, the performance for the *numbers* task averaged over all participants improved by almost 7%.

Several participants (*nku, nkp, nkl, nkk*) still could not reach the threshold of 61.11% which

Table 3: Mean classification accuracies for 6 classifiers. One row represents one experiment, the first column shows the participant codes. For each experiment, the conditions with highest (bold) and lowest (red) performance rates are highlighted.

separate	**overall**	**clean**	**eyesclosed**	**news**	**numbers**	**flicker**	**stimulation**
od	95.83	98.61	**100.00**	98.61	83.33	98.61	95.83
njy	63.66	59.72	68.06	65.28	41.67	**73.61**	**73.61**
njz	72.92	75.00	**84.72**	81.94	68.06	68.06	59.72
nkk	51.62	48.61	50.00	**58.33**	55.56	50.00	47.22
nkl	49.31	**52.78**	48.61	51.39	40.28	**52.78**	50.00
nkm	58.33	63.89	52.78	56.94	**66.67**	54.17	55.56
nkn	50.13	55.56	55.56	54.17	42.42	27.78	**65.28**
nko	91.31	**95.83**	84.72	91.67	90.28	94.44	90.91
nkp	50.93	**54.17**	44.44	52.78	48.61	**54.17**	51.39
nkq	71.53	70.83	69.44	72.73	60.61	**80.56**	75.00
nkr	55.56	48.61	48.61	51.39	**66.67**	52.78	65.28
nks	53.16	**62.50**	50.00	48.61	43.94	52.78	61.11
nkt	65.97	63.89	72.22	58.33	65.28	**79.17**	56.94
obx	85.65	91.67	83.33	86.11	81.94	**95.83**	75.00
nku	52.78	52.78	54.17	54.17	48.61	50.00	**56.94**
ma4	59.24	43.94	48.61	59.72	**71.21**	68.06	63.89
overall	64.24	64.90	63.45	65.14	60.95	**65.80**	65.23

shows that their data is not even separable into left and right hand motor imagination. However, for participants *nkq* and *nko* classification rates improved by $8\% - 9\%$.
This leads to the conclusion that, applying the correct classifier, left and right hand motor imagination is indeed separable for most participants. Thinking about real-world scenarios, the problem however is, that we not always know which task a user is carrying out while controlling the BCI.

4. Evaluation of New Strategies for Out-of-Lab

In the last section we found that artifacts highly contaminate the data, especially in the numbers task. Another problem is that testing data significantly shifts from calibration data [43, 44, 10]. However, if we compute task-specific classifiers we could separate left from right hand motor imagination, for most participants. Since standard machine learning methods are not able to handle feature shifts and artifact contamination we need to further investigate other methods such as adaptation or artifact removal. In this section, we propose three strategies to improve classification.

(i) Artifact removal, since we discovered that data is highly artifact-contaminated (see

Section 3.3), we remove the classified artifacts before classification.

(ii) Classifier ensemble, instead of dividing the dataset into the different distraction tasks, we apply all six classifiers from Section 3.5 and average the classifiers output.

(iii) 2-step classification, we first identify in which distraction task the motor imagination was conducted before applying the respective classifier.

4.1. Improvement via artifact reduction

Table 4 shows the mean classification rates over subjects for each distraction condition when each artifact group is removed from data. To remove artifact groups, the independent components from each artifact group were projected out from both the calibration and test data. From the first row of Table 4, we see that the best artifact group to remove is that containing ocular artifacts. This improves the classification performance for all distractions except *stimulation* and the *clean* condition. The most difficult groups to remove are the muscular and muscular and mixed groups. Both of these groups cause decreased classification performances in four conditions. The performance in the *news* distraction is improved by removing all the artifact groups. Similarly, the *numbers* condition is also improved when any artifact group, except the muscle group, is removed. However, the improvements are not statistically significant. When testing whether the performances with artifacts removed differ from the baseline performances with a two-sided t-test, no p-value is below 0.05 (not correct for multiple hypothesis tests). These results are consistent with previous investigations in which removing artifacts did not improve BCI performance significantly [45].

Table 4: Mean classification accuracies for all distractions and removed artifact groups. For each experiment, the artifact group with highest (**bold**) and lowest (red) performance rates and performances better than baseline (blue) are highlighted. Lowest scores better than baseline are purple. The overall (rightmost) column from Table 2 is reproduced for baseline comparison.

	Muscular	Ocular	Non-neural	Muscular and mixed	Non-mixed artefacts	overall	baseline
overall	62.30	**62.69**	62.20	61.83	62.01	62.20	62.39
clean	66.46	65.67	**66.97**	65.94	64.72	65.95	66.68
eyesclosed	62.50	**63.80**	60.75	60.68	63.19	62.19	63.10
news	62.81	63.76	63.51	**63.94**	63.42	63.49	62.64
numbers	53.56	55.29	**56.86**	54.69	55.12	55.10	53.81
flicker	**67.10**	66.58	64.93	65.28	65.10	65.80	66.41
stimulation	**61.35**	61.01	60.14	60.49	60.48	60.69	61.53

Table 5: Mean classification accuracies for classifier ensembles.

od	njy	njz	nkk	nkl	nkm
97.92	64.58	75.93	49.07	50.23	65.28
nkn	nko	nkp	nkr	nkq	nks
58.92	93.66	52.78	73.24	55.56	54.69
nkt	obx	nku	ma4	overall	
71.99	88.89	52.55	55.63	66.31	

4.2. Improvement via classifier ensemble

Instead of choosing a task-specific classifier for each trial, we propose an ensemble approach, where we applied all six classifiers to all trials and averaged over the output to determine whether a left or right hand motor imagination was conducted.
Average classification rates for all 16 participants can be found in Table 5. Compared to the results from Table 3 where we applied only one of those six classifiers, classification rates increased by 1%. This means ensemble classification works even more reliably than computing task-specific classifiers.

4.3. Improvement via 2-step classification

Calculating one classifier for each distraction indeed yields higher classification rates but if we think about applying this concept to real world situations, we might not have that much prior knowledge about the scenarios the BCI is used in. Therefore, we propose a 2-step classification approach which combines classifying the respective condition and building several classifiers. In the first step, we want to find out in which condition the respective task was conducted. Here, we only distinguish between *numbers* task and *not numbers* task (*clean, closed eyes, news, flicker, stimulation*). After categorizing a task to one of these groups, we apply one of two classifiers (one for *numbers* and one for *not numbers*) to decide whether this trial consisted of a *right* or *left* hand motor imagination. For this approach, we only consider the 6 runs including distractions and conduct a 6-fold cross validation equivalent to the one in Section 3.5.
Results are summarized in Table 6 where the average classification rates for both steps are listed. The overall classification rate is the weighted average performance of the *not numbers* and *not number* task. The weighting compensates for the different number of trials in both tasks.
As already discussed in Section 3.4, the different conditions are easily separable, results of the 1st step in Table 6 show that classification rates are mostly between 94% and above 99%. Except for participant *nkm* where we are only able to classify 86% of the conditions correctly. Since only 12 out of 72 trials in each run belong to the *numbers* task, this result lead to the conclusion that conditions are not really distinguishable for this participant. However,

Table 6: Mean classification accuracies for 2-step classification.

		1st step	2nd step	
	overall	**cond**	**not numbers**	**numbers**
od	96.53	100.00	99.17	83.33
njy	66.20	96.53	70.54	46.84
njz	77.55	97.45	78.71	72.00
nkk	48.61	94.68	48.12	50.57
nkl	46.99	99.31	47.90	42.67
nkm	66.90	86.34	70.66	50.62
nkn	57.75	96.95	59.08	51.90
nko	93.19	96.71	93.82	90.00
nkp	49.07	95.37	48.56	51.19
nkq	77.93	99.53	80.28	66.20
nkr	58.80	99.07	57.26	66.22
nks	57.75	98.83	58.43	54.29
nkt	76.85	99.77	79.11	65.75
obx	90.28	99.31	91.92	82.19
nku	52.08	98.84	52.65	49.32
ma4	61.27	98.83	60.45	65.28

for most participants (except *nkk,nkl,nkp,nkr*) we reached higher classification rates for all experiments with this 2-step approach, even compared to the accuracies for 6 different classifiers (see Table 3). One reason for this may also be the amount of training data we used to train the *not numbers* classifier. Whereas we used between 60 and 72 training trials for the previous approaches, here we could use now up to 300 training trials.

5. Conclusion

In this paper we presented a motor imagery-based BCI study where participants had to handle 5 different distraction tasks in addition to the motor imagery task. The idea behind those tasks was to simulate a semi-realistic environment and to systematically analyze the influence of different scenarios on the motor imagery performance. Since CSP results only led to low classification rates we proposed three different approaches to improve performance, artifact removal, ensemble CSP and a 2-step approach. In the comparisons of performances in each task before and after removing artifacts, we did not see any significant differences. In Figure 8, we display the comparison of the results of ensemble and 2-step approaches to our original CSP approach (see Section 3.1) together with the p-values of a one sided Wilcoxon signed rank test. Each circle represents one participant. Considering a significance level of $\alpha = 0.05$ we get significant improvement for ensemble CSP and the 2-step approach. However for

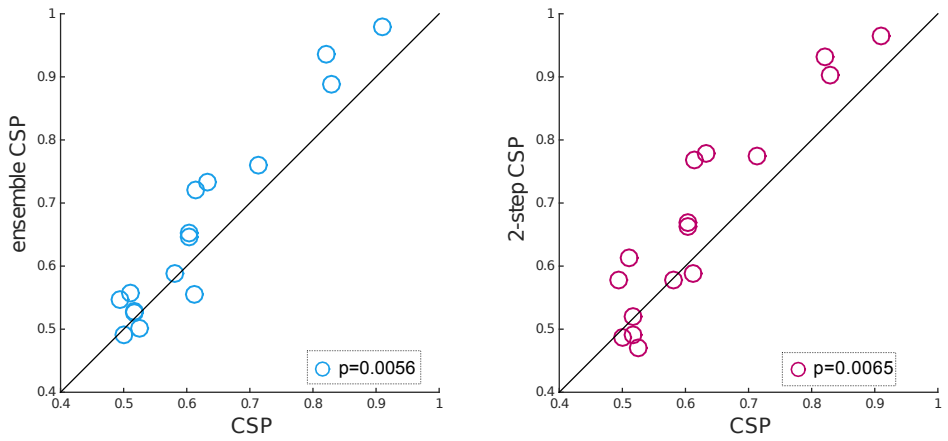


Figure 8: The three new approaches compared to CSP trained on *clean*

both, the ensemble CSP and the 2-step approach we observe that participants who already get significant BCI performance for the original CSP approach improve even more with the proposed methods.

Those findings lead to the conclusion that not every BCI user may be able to handle distraction tasks equally well to the motor imagery task. We also have to note that most of the participants were confronted for the first time with a BCI system. Imagining a movement is relatively abstract and some participants may improve by engaging in more feedback training before going "out-of-lab".

After first steps have been made to leave the controlled lab environment, this study systematically and quantitatively analyses how different scenarios influence BCI performance. The difficulties we identified, especially with the muscular artifacts in the numbers task need to be considered for future studies and are worth being further analyzed. Also training the BCI users more detailed beforehand could lead to a better understanding and higher performance rates. Future studies will explore whether harvesting a data base of significantly larger numbers of subjects and tasks may allow a invariant and subject independent decoding [46], e.g., using deep neural networks.

Acknowledgment

This work was supported in part by the Adaptive BCI (FKZ 01GQ1115) and by the Brain Korea 21 Plus Program 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. Correspondence to SB, KRM and WS.

References

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

[2] Bernhard Graimann, Brendan Z Allison, and Gert Pfurtscheller. *Brain-computer interfaces: Revolutionizing human-computer interaction*. Springer, 2010.

[3] Tobias Kaufmann, Andreas Herweg, and Andrea Kübler. Toward brain-computer interface based wheelchair control utilizing tactually-evoked event-related potentials. *Journal of neuroengineering and rehabilitation*, 11(1):7, 2014.

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

[5] Klaus-Robert Müller, Michael Tangermann, Guido Dornhege, Matthias Krauledat, Gabriel Curio, and Benjamin Blankertz. Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *Journal of neuroscience methods*, 167(1):82–90, 2008.

[6] Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Klaus-Robert Müller, and Gabriel Curio. The non-invasive berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

[7] Matthias Krauledat, Michael Tangermann, Benjamin Blankertz, and Klaus-Robert Müller. Towards zero training for brain-computer interfacing. *PLoS ONE*, 3(8):e2967, 2008.

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

[9] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[10] Wojciech Samek, Carmen Vidaurre, Klaus-Robert Müller, and Motoaki Kawanabe. Stationary common spatial patterns for brain-computer interfacing. *Journal of Neural Engineering*, 9(2):026013, 2012.

[11] M. Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst.*, 24(4):610–619, 2013.

[12] Wojciech Samek, Motoaki Kawanabe, and Klaus-Robert Müller. Divergence-based framework for common spatial patterns algorithms. *IEEE Reviews in Biomedical Engineering*, 7:50–72, 2014.

[13] Motoaki Kawanabe, Wojciech Samek, Klaus-Robert Müller, and Carmen Vidaurre. Robust common spatial filters with a maxmin approach. *Neural Computation*, 26(2):1–28, 2014.

[14] Stephanie Brandl, Klaus-Robert Müller, and Wojciech Samek. Robust common spatial patterns based on bhattacharyya distance and gamma divergence. In *Proc. of Int. Winter Workshop on Brain-Computer Interface*, pages 1–4, 2015.

[15] Mehrdad Fatourechi, Ali Bashashati, Rabab K Ward, and Gary E Birch. Emg and eog artifacts in brain computer interface systems: A survey. *Clinical neurophysiology*, 118(3):480–494, 2007.

[16] Alois Schlögl, Claudia Keinrath, Doris Zimmermann, Reinhold Scherer, Robert Leeb, and Gert Pfurtscheller. A fully automated correction method of eog artifacts in eeg recordings. *Clinical neurophysiology*, 118(1):98–104, 2007.

[17] 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.

[18] Peter Sykacek, Stephen J Roberts, and Maria Stokes. Adaptive BCI based on variational bayesian kalman filtering: an empirical evaluation. *IEEE Transactions on Biomedical Engineering*, 51(5):719–727, 2004.

[19] Pradeep Shenoy, Matthias Krauledat, Benjamin Blankertz, Rajesh PN Rao, and Klaus-Robert Müller. Towards adaptive classification for bci. *Journal of neural engineering*, 3(1):R13, 2006.

[20] Carmen Vidaurre, Claudia Sannelli, Klaus-Robert Müller, and Benjamin Blankertz. Machine-learning based co-adaptive calibration. *Neural computation*, 23(3):791–816, 2011.

[21] Fabien Lotte, Junya Fujisawa, Hideaki Touyama, Rika Ito, Michitaka Hirose, and Anatole Lécuyer. Towards ambulatory brain-computer interfaces: A pilot study with p300 signals. In *Proc. of the Int.*

Conf. on Advances in Computer Enterntainment Technology, pages 336–339, 2009.

[22] Maarten De Vos, Katharina Gandras, and Stefan Debener. Towards a truly mobile auditory brain–computer interface: exploring the p300 to take away. *International journal of psychophysiology*, 91(1):46–53, 2014.

[23] Hayrettin Gürkök, Mannes Poel, and Job Zwiers. Classifying motor imagery in presence of speech. In *Proc. of Int. Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

[24] C Neuper, GR Müller, A Kübler, N Birbaumer, and G Pfurtscheller. Clinical application of an eeg-based brain–computer interface: a case study in a patient with severe motor impairment. *Clinical neurophysiology*, 114(3):399–409, 2003.

[25] Kai Keng Ang, Cuntai Guan, Karen Sui Geok Chua, Beng Ti Ang, Christopher Wee Keong Kuah, Chuanchu Wang, Kok Soon Phua, Zheng Yang Chin, and Haihong Zhang. A large clinical study on the ability of stroke patients to use an eeg-based motor imagery brain-computer interface. *Clinical EEG and Neuroscience*, 42(4):253–258, 2011.

[26] Johannes Höhne, Elisa Holz, Pit Staiger-Sälzer, Klaus-Robert Müller, Andrea Kübler, and Michael Tangermann. Motor imagery for severely motor-impaired patients: evidence for brain-computer interfacing as superior control solution. *PLOS ONE*, 9(8):e104854, 2014.

[27] Stephanie Brandl, Johannes Höhne, Klaus-Robert Müller, and Wojciech Samek. Bringing bci into everyday life: Motor imagery in a pseudo realistic environment. In *Proc. of the Int. IEEE/EMBS Neural Engineering Conference (NER)*, pages 224–227, 2015.

[28] H.H. Jasper. The ten twenty electrode system of the international federation. *EEG Clin. Neurophysiol.*, 10:371–375, 1958.

[29] Dennis J McFarland, Lynn M McCane, Stephen V David, and Jonathan R Wolpaw. Spatial filter selection for eeg-based communication. *Electroencephalography and clinical Neurophysiology*, 103(3):386–394, 1997.

[30] Jian Ding, George Sperling, and Ramesh Srinivasan. Attentional modulation of ssvep power depends on the network tagged by the flicker frequency. *Cerebral cortex*, 16(7):1016–1029, 2006.

[31] Shozo Tobimatsu, You Min Zhang, and Motohiro Kato. Steady-state vibration somatosensory evoked potentials: physiological characteristics and tuning function. *Clinical neurophysiology*, 110(11):1953–1958, 1999.

[32] Anne-Marie Brouwer and Jan BF Van Erp. A tactile p300 brain-computer interface. *Frontiers in neuroscience*, 4:19, 2010.

[33] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE Trans. Rehab. Eng.*, 8(4):441–446, 1998.

[34] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[35] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Proc. Magazine*, 25(1):41–56, 2008.

[36] Hiroshi Morioka, Atsunori Kanemura, Jun-ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii. Learning a common dictionary for subject-transfer decoding with resting calibration. *NeuroImage*, 111:167–178, 2015.

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

[38] Laura Frølich, T. S. Andersen, and Morten Mørup. Classification of independent components of eeg into multiple artifact classes. *Psychophysiology*, 52(1):32–45, 2015.

[39] Ernst Niedermeyer and Fernando Henrique Lopes da Silva. *Electroencephalography : basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, Philadelphia, 2005.

[40] T. Gasser, L. Sroka, and J. Mocks. The transfer of EOG activity into the EEG for eyes open and closed. *Electroencephalogr Clin Neurophysiol*, 61(2):181–193, Aug 1985.

[41] Joe-Air Jiang, Chih-Feng Chao, Ming-Jang Chiu, Ren-Guey Lee, Chwan-Lu Tseng, and Robert Lin. An automatic analysis method for detecting and eliminating ECG artifacts in EEG. *Computers in Biology*

and Medicine, 37(11):1660 – 1671, 2007.

[42] S. D. Muthukumaraswamy. High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Front Hum Neurosci*, 7:138, 2013.

[43] Pradeep Shenoy, Matthias Krauledat, Benjamin Blankertz, Rajesh P. N. Rao, and Klaus-Robert Müller. Towards adaptive classification for BCI. *Journal of neural engineering*, 3(1):R13–R23, 2006.

[44] Paul von Bünau, Frank C. Meinecke, Franz Király, and Klaus-Robert Müller. Finding stationary subspaces in multivariate time series. *Physical Review Letters*, 103:214101, 2009.

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

[46] Siamac Fazli, Sven Dähne, Wojciech Samek, Felix Bießmann, and Klaus-Robert Müller. 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, 2015.

Appendix D

Removal of muscular artifacts in EEG signals: A comparison of ICA and other linear decomposition methods

Removal of muscular artifacts in EEG signals: A comparison of ICA and other linear decomposition methods

Laura Frølich*† Irene Winkler*‡

Background. The electroencephalogram (EEG) is contaminated by undesired signals of non-neural origin, such as eye and muscle movements. The most common approach for muscle artifact reduction is to linearly decompose EEG signals into source components using Independent Component Analysis (ICA), to separate artifactual and neural sources. While many different linear decomposition methods are available, only few studies compared their performance on real EEG data.

Comparison with existing methods. Here we compare three of the most commonly used ICA methods (Extended InfoMax, FastICA, TDSEP) and three other linear decomposition methods (Fourier-ICA, Spatio-Spectral-Decomposition (SSD), Parafac2). We use an automatic artifactual component classifier (IC_MARC) and EEG recordings from 18 subjects that are heavily contaminated by muscle artifacts. Subjects performed self-paced foot movements which led to expected event-locked neural activity (event-related desynchronization (ERD)) as well as a clearly visible event-locked muscle artifact. This allows us to evaluate the methods ability to remove the event-locked muscle artifact while maintaining ERD.

Results. As expected, we find that it is in general not possible to completely remove the artifact while retaining all neural activity. Nevertheless, the three analyzed ICA methods drastically reduce the muscle artifact and perform as well or better than the three other linear decomposition methods - but only if the data are adequately high-pass filtered. We observed no consistent performance differences between the three compared ICA methods. This indicates that for ICA-based artifact removal, data-prepocessing choices have a more pronounced effect than the choice of the ICA method.

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.

In this paper, we are concerned with the removal of muscle artifacts, caused by e.g. chewing, swallowing, head or tongue movements. Muscle activity is usually a high-frequency activity (> 20 Hz) (Goncharova et al., 2003), but may present a wide spectral band distribution which pertubes all classic EEG bands. Because muscle activity arises from different type of muscle groups, muscle artifacts are harder to stereotype than eye artifacts (cf. McMenamin et al. (2010); Muthukumaraswamy (2013)).

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) (Makeig et al., 1996; Vigário, 1997; Jung et al., 2000; Vigário et al., 2000). ICA linearly transforms EEG signals into independent source components (ICs). If artifactual and neural activity are contained

*Authors contributed equally. Correspondence to: lffr@dtu.dk; irene.winkler@tu-berlin.de
†Technical University of Denmark
‡Technische Universität Berlin

in separate components, artifactual components can be identified and a cleaner EEG can be reconstructed without them.

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 neural signals). Nevertheless, their application usually leads to a good, albeit not perfect separation of artifactual and neural sources. Typically, a number of mixed components still 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).

Even though many different BSS/ICA methods are available, not many validation and comparison studies exist which compare the performance of BSS algorithms on real EEG data. Most studies focused on simulated data, in which real or simulated 'artifact-free' and 'artifactual' data are linearly mixed at some known ratio (Kierkels et al., 2006; Fitzgibbon et al., 2007; Romero et al., 2008; Crespo-Garcia et al., 2008; Klemm et al., 2009; Olbrich et al., 2011; Albera et al., 2012; Safieddine et al., 2012; Vazquez et al., 2012). In that way, the ground truth is available and the results can be quantified. However, the simulated contamination may not reflect true muscle contamination. For example, muscle activity does not always occur independently from the neural signals of interest.

Validation of ICA-based artifact reduction on real data is scarce and often focuses on eye artifacts (Wallstrom et al., 2004; Hoffmann & Falkenstein, 2008). For muscle artifacts, McMenamin et al. (2010) conducted a validation study using a data set in which muscle activity and neural activity in the alpha band were independently varied (by instructing participants to close or open their eyes, and to tense or relax their cranial muscles). They found that muscle artifact removal using the extended InfoMax algorithm (Bell & Sejnowski, 1995; Lee et al., 1999) and manually selecting artifactual components was not perfect, but superior to previously validated regression-based techniques. However, different ICA algorithms were not compared.

Such a comparison was conducted by Delorme et al. (2012). They compared decompositions of 22 different BSS algorithms by evaluating measures of independence (based on mutual information) as well as the 'dipolarity' of the resulting components. Here 'dipolarity' is defined as the number of component whose scalp maps can be explained by a single equivalent dipole with less than a specified error variance. They find that mutual information based ICA methods such as InfoMax result in the highest number of near-dipolar components.

In this paper, we compare the overall artifact reduction performance of several linear decomposition methods using a data set which contains both event locked neural activity as well as an event locked muscle artifact. More specifically, 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 (alpha band) and 15-30 Hz (beta band) rhythms over corresponding sensorimotor areas (Neuper & Pfurtscheller, 2001). The data set also contains a clearly visible event-locked muscle artifact, which probably stems from subjects moving their head rhythmically along with the foot movement. This allows us to evaluate the methods' ability to remove the event-locked muscle artifact while maintaining ERD.

We compare the three most common ICA/BSS methods for EEG data, namely Extended Infomax (Bell & Sejnowski, 1995; Lee et al., 1999), FastICA (Hyvärinen & Oja, 1997; Hyvärinen, 1999) and SOBI/TDSEP (Belouchrani et al., 1997; Ziehe & Müller, 1998) with three other linear decomposition methods, namely Fourier-ICA (Hyvärinen et al., 2010), Spatio-Spectral Decomposition (SSD) (Nikulin et al., 2011), and Parafac2 (Kiers et al., 1999; Harshman, 1972). These three methods have not been proposed for artifact removal, but they might be well suited for our data set because we are interested in oscillatory activity. To select the artifactual components, we use a previously validated automatic artifactual component classifier (IC_MARC, (Frølich et al., 2015)).

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 the ability of 6 linear decomposition methods to separate artifactual from neural activity. In some sense, all try to solve the blind source separation (BSS) problem, which is the task of recovering underlying source 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 then estimated such that the estimated sources

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

best fulfill pre-defined assumptions.

For artifact reduction, our hope is that artifactual and neural 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. Several algorithms are available to solve this task, and we focus here on three of the most commonly used methods: Extended Infomax (Bell & Sejnowski, 1995; Lee et al., 1999) as implemented in EEGLab (Delorme & Makeig, 2004), FastICA (Hyvärinen & Oja, 1997; Hyvärinen, 1999) and SOBI/TDSEP (Belouchrani et al., 1997; Ziehe & Müller, 1998).

Extended Infomax and FastICA are classical ICA methods which rely on higher-order statistics to define independence. Infomax was derived from a neural network viewpoint, while FastICA maximizes the negentropy of the component distributions. Second-order methods take the temporal structure of the time series into account and enforce decorrelation over time. Here we use TDSEP (Temporal Decorrelation source SEParation) (Ziehe & Müller, 1998; Ziehe et al., 2004), which is equivalent to SOBI (Second Order Blind Identification) (Belouchrani et al., 1997). TDSEP/SOBI amounts to finding a demixing $\hat{W}$ which leads to minimal cross-covariances over several time-lags between all pairs of components of $\hat{S}$.

Running ICA We used Extended Infomax, which finds both sub- and super-Gaussian sources, with the default settings in EEGLab for our analyses. We ran FastICA with the *symmetric* approach and all other options at default EEGLab values. We used code from A. Ziehe in the estimation of the TDSEP model, available at `www.user.tu-berlin.de/aziehe/code/ffdiag_pack.zip`, setting the number of time lags, τ, to 99. We extracted as many components as there were channels for all three methods.

2.2.2 Fourier-ICA

Hyvärinen et al. (2010) recently proposed to apply ICA on short-time Fourier transforms of EEG signals, in order to find more 'interesting' oscillatory sources than with time-domain ICA. ICA optimization then translates into optimizing the sparseness of the Fourier coefficients, which should separate oscillatory signals at different frequencies.

Fourier-ICA has not been specifically designed to extract artifacts. In fact, the authors point out that time-domain ICA can be interpreted as maximizing non-Gaussianity. ICA may therefore be very well suited to find artifacts, which often are very non-Gaussian due to outliers in their time courses. Rather, the hope is that Fourier-ICA is better able to extract relevant oscillatory sources. In our setting, we evaluate to which extent clean oscillatory activity can be obtained. Fourier-ICA might therefore be a promising method.

Running Fourier-ICA We used the implementation described in Hyvärinen et al. (2010) to run FourierICA with the default parameters. The minimum and maximum frequencies to be analysed by FourierICA were 8 and 14 Hz for our alpha band analyses and 15 and 30 Hz for our beta band analyses. We extracted as many components as there were channels.

2.2.3 SSD

Another recently proposed method for the extraction of oscillations is Spatio-Spectral Decomposition (SSD) (Nikulin et al., 2011). The purpose of SSD is to extract oscillations in a frequency band of interest at maximal signal-to-noise ratio (SNR). More specifically, SSD maximizes the signal power in the frequency band of interest while simultaneously minimizing it at the neighboring frequency bins. SSD seeks spatial filters $\mathbf{w} \in \mathbb{R}^M$ which maximize

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

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). Preliminary results for SSD on our data set were described in Winkler et al. (2015b)

Running SSD We set the frequency bands of interest to 8 - 14 Hz for the alpha band analyses and 15-30 Hz for the beta band analyses. The sidebands were 2 Hz long. We extracted as many components as there were channels and ordered them according to their SNR.

2.2.4 Parafac2

Often in analyses of EEG data, data is averaged or concatenated across trials. However, this disregards the variation in the $channel \times time$ structure across trials. Tensor methods instead exploit the multi-dimensional structure of EEG data to infer the factors, e.g. spatial and temporal patterns along with their degrees of expression in each trial, that best explain the observed data (Deburchgraeve et al., 2009; Acar et al., 2007; De Vos et al., 2007; Vanderperren et al., 2010; Paulick et al., 2014). Here we discuss the 3D structure consisting of $channel \times time \times epoch$, represented as multi-linear data matrix (tensor) $\mathcal{X} \in \mathbb{R}^{M \times T_e \times N}$, where M is the number of channels, T_e is the number of samples recorded in one epoch, and N is the number of epochs. The Parafac model (also known as the CanDecomp model) (Harshman, 1970; Carroll & Chang, 1970; Kiers et al., 1999), is an unsupervised tensor decomposition method in which each spatial pattern interacts with only one temporal pattern and one trial strength factor, each temporal pattern interacts with only one spatial pattern and trial strength factor, etc. The formal expression of the Parafac model for each epoch, $\mathcal{X}_n \in \mathbb{R}^{M \times T_e}$ $n \in 1, 2, \ldots, N$, is:

$$\mathcal{X}_n = A(FD_n)^\top + R_n,$$

where $R_n \in \mathbb{R}^{M \times T_e}$ contains the differences between the model and the data. The matrix $A \in \mathbb{R}^{M \times K}$ holds the spatial patterns common to all epochs while the matrix $F \in \mathbb{R}^{T_e \times K}$ contains the temporal patterns common to all epochs. The matrices $D_n \in \mathbb{R}^{K \times K}$ are diagonal matrices, ensuring that the j^{th} column of A only interacts with the j^{th} column of F. The magnitude of this interaction for the j^{th} factor pair in trial n is determined by the value of the j^{th} diagonal element of D_n. Although the true number of generating components is unknown, as for all the other methods, component numbers much lower than the number of channels are usually used in tensor decompositions of EEG data (Acar et al., 2007; Weis et al., 2010; Vanderperren et al., 2010; Paulick et al., 2014)

Parafac2 is an extension of the Parafac model. While Parafac extracts spatial and temporal factors that are the same across epochs, Parafac2 allows for some variation in the temporal factors across epochs, which is reasonable for artifacts' time courses. Parafac2 has previously been shown to explain EEG data better than Parafac (Weis et al., 2010). The model for each epoch is:

$$\mathcal{X}_n = A(F_n D_n)^\top + R_n, \tag{3}$$

By factorising F_n as $P_n F$, where $P_n \in \mathbb{R}^{T_e \times K}$ is orthonormal, a matrix, $F \in \mathbb{R}^{K \times K}$, containing a temporal structure common to all epochs can be obtained Kiers et al. (1999). Since the matrices P_n are orthonormal, they represent epoch-specific rotations and flips of the temporal profiles in F.

Running Parafac2 We used the *nway331* toolbox (Andersson & Bro, 2000) to run Parafac2. In order to avoid implicitly supervising Parafac2, we split the data into one second epochs instead of using the brake presses to define trials. We required the trial-strength factors to be orthogonal, which prevents the Parafac2 solution from degenerating. We also ran the analyses with the constraint on the spatial mode. However, this can affect the estimated scalp maps strongly, which could be a problem for the classification relying on spatial features. Indeed, the results stemming from constraining the spatial mode amplified the artifact instead of reducing it. For clarity of exposition, we do not include these results. We used the default initialisation, which initialises with the best run of 10 preliminary short runs. This ensures that the initial point is not a local minimum. Since Parafac2 estimates the matrices A, D_n, F, and P_n, there are more free parameters compared to other BSS methods that just estimate A. This limited the number of identifiable components, so we only estimated up to 20 components for each subject with Parafac2. However, this should not be a problem for the model since the number of components in tensor models of EEG data is typically quite low, as mentioned above.

2.2.5 High-pass filtering

It is well known that high-pass filtering EEG data prior to ICA may improve the quality of the artifact separation (Hyvärinen et al., 2001; Pignat et al., 2013). In fact, it is a fairly standard procedure to remove drifts prior to ICA-based artifact removal, and the benefit has been demonstrated in several studies (Groppe et al., 2009; Zakeri et al.; Winkler et al., 2015a). Our data was already subjected to standard EEG processing, and on our band-pass filtered data drifts are not a problem (cf. Section 2.1).

However, filtering at higher frequencies might also be beneficial when oscillatory processes are of interest. For example, 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 (Scheeringa et al., 2011). We might therefore benefit from a high cut-off frequency also in our study. Furthermore, we use information on the frequency band of interest for both FourierICA and SSD. In order to obtain a fairer comparison to SSD and FourierICA, we ran the other four decomposition methods (InfoMax, FastICA, TDSEP, Parafac2) both on the broad-band filtered data and on the data after a high-pass filter with a high cut-off frequency had been applied. In the analyses of the movement artifact in the alpha band, the cut-off frequency was set at 7 Hz. The cut-off for the beta band analyses was set at 14 Hz.

2.3 Automatic classification of estimated sources

Successful artifact removal relies on the correct identification of artifactual and non-artifactual components. This identification of artifactual component is a non trivial task and requires time and expert knowledge. For a description of typical artifact components we refer the reader to Chaumon et al. (2015). Here we use a previously validated automatic classifier of artifactual components, IC_MARC, to classify the sources estimated by each method (Frølich et al., 2015). IC_MARC was developed for sources derived by ICA, but may also be used to classify sources obtained from other methods.

IC_MARC assigns probabilities to ICs of belonging to each of six classes (blinks, lateral eye movements, electrical heart beat artifact, muscle artifact, neural, or mixed artifact) and relies on multinomial regression to predict class probabilities for each IC. We use these probabilities in two ways in this paper: 1) by classifying all components to the class for which the highest probability was predicted, we clean the data by removing all ICs not classified as neural and 2) we use the probabilities of the ICs being neural to determine the order of IC removal. We use a version of IC_MARC which is based on a feature set containing only spatial features that we have seen to work well previously. IC_MARC tends to have a high specificity and sensitivity for the neural class with a balanced accuracy of 88% for 8023 independent components when training on one study and testing on another (Frølich et al., 2015).

2.4 Evaluation: Event-Related Desynchronization (ERD)

Each method was independently applied to the continuous EEG data. To compare the methods, we plot grand-average Event-Related (De-)Synchronization (ERD/ERS) in the alpha (8-14 Hz) and 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):

$$\text{ERD}(t) := \frac{\text{Power}(t) - \text{Reference power}}{\text{Reference power}} \tag{4}$$

where 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.

From the literature, we expect ERD in both frequency ranges to be most prominent over central sensorimotor areas, and to start prior to the voluntary foot movement (cf. Neuper & Pfurtscheller (2001)). In our data, however, we additionally see a contamination of the ERD in the form of a swift, strong peak at movement onset (cf. Figures 1 and 2). This is probably due to subjects moving their heads along with the fairly rhythmical foot movement once per second.

The goal of artifact removal is to remove this muscle artifact while retaining the neural activity. Hence we aim to obtain a cleaner signal such that obtained ERD resembles the uncleaned data before and after the event-locked muscle artifact while exhibiting low ERD throughout the foot movement. To quantify how well each method obtains this goal, we define the following heuristic ERD quality measure:

$$\text{Quality}(ERD) := \left(\max_{t \in [-50,\, 50]} \{ERD(t)\} - priorERD \right) \cdot \left(\min_{t \in [-50,\, 50]} \{ERD(t)\} - priorERD \right) \cdot 100\,, \tag{5}$$

which we compute separately for each subject and preprocessing method. Here $priorERD$ denotes the mean of the 30% lowest ERD values prior to foot movement in the uncleaned data (computed between -300 ms and -50 ms relative to EMG peak activity). For each preprocessing variant, $\max\{ERD(t)\}$ and $\min\{ERD(t)\}$ are computed as the maximum and minimum values of ERD on the cleaned data between -50 ms and 50 ms relative to EMG peak activity. An effective artifact removal method will reduce both $\max\{ERD(t)\}$ and $\min\{ERD(t)\}$ to be close to the ERD before the event-locked muscle artifact, such that it is similar to $priorERD$. An effective artifact removal method will therefore be reflected by a low quality score.

We use this ERD quality measure to evaluate the methods' dependence on the number of source components retained and the variance these explain. For each method, except SSD, we rank the obtained components by the probability of being an artifact as determined by IC_MARC. For SSD, we rank the components according to SNR. Retaining a smaller or larger number of sources corresponds to either a strict or soft policy for the removal of potential artifactual sources. Therefore, we vary the number of retained components from 1 to the number of channels (except for Parafac2, for which only 20 components were estimated), and we report the average ERD quality measure over subjects.

3 Results

Figures 1 and 2 show the grand-average ERD data with no cleaning and the same data cleaned by removing all non-neural sources for each method, except SSD for which we retained the 10 components with highest SNR as in (Dähne et al., 2014; Winkler et al., 2015c). This choice of 10 SSD components was based on prior experience. We also looked at the results for SSD with components chosen according to IC_MARC, as for the other methods. However, the performance resulting from this component selection was lower than that using the SNR. For SSD, we therefore only present the results using SNR.

Figure 1 shows the data for the alpha band while Figure 2 shows the the same data, but band-pass filtered for the beta band. The top of each figure contains the ERD time course at channel Cz, while the scalp maps corresponding to the intervals marked in light and dark gray are depicted for some of the best performing methods in the bottom part. The results from applying the decomposition methods after high-pass filtering data at a high cut-off frequency are shown in dashed lines. Prior to foot movement, we see a typical foot ERD over central sensorimotor areas as expected. During movement, the ERD is contaminated by a muscular artifact which spans the whole scalp. The compared methods are able to reduce this artifact to varying degrees.

From the top parts of the figures, for both bands, we see that all decomposition methods improve if the data is high-pass filtered at a high cut-off frequency before being decomposed. We also see that high-pass filtered Extended Infomax obtains the best ERD quality measure, which is in accordance with this method achieving the lowest band power during the movement artifact. In the beta band, high-pass filtered Extended Infomax is able to

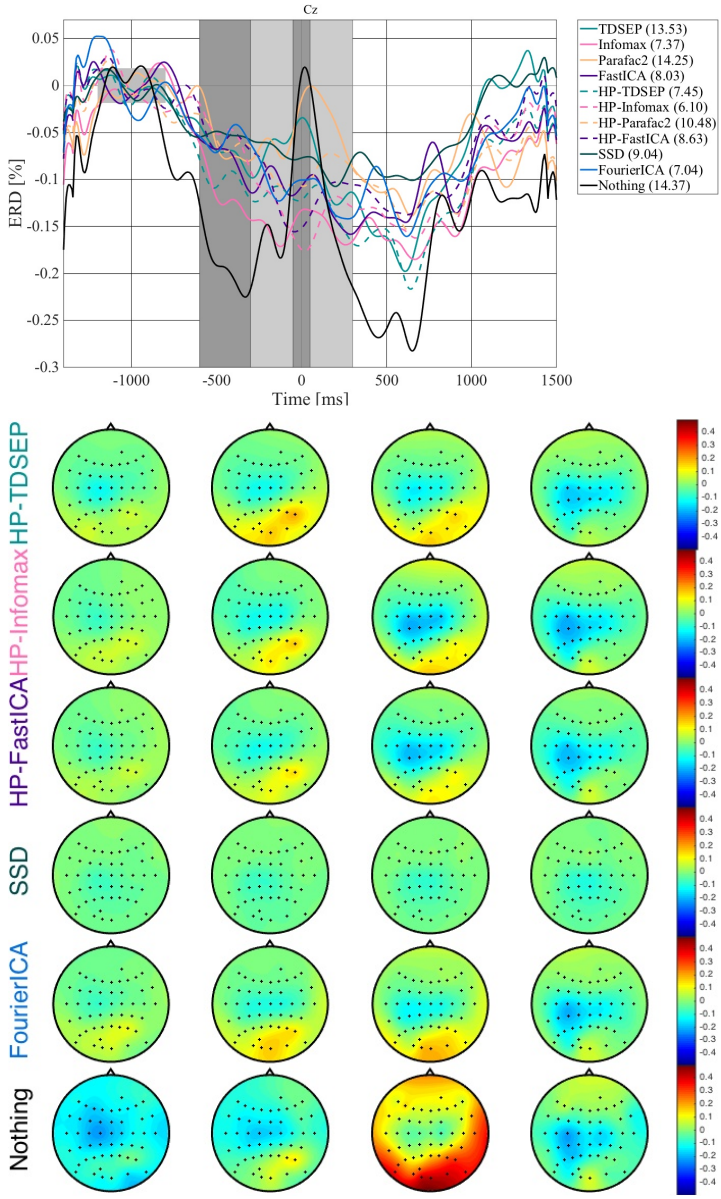


Figure 1: Grand-average ERD/ERS for 18 subjects recorded during self-paced foot movements in the alpha band (7-14 Hz), aligned to EMG peak activity. (Top) Time courses of data reconstructed from neural ICs (and for SSD with the ten components with highest SNR) at channel Cz. The legend contains the ERD quality measure for each method, lower is better. (Bottom) Series of ERD maps in the marked intervals ([-600 -300], [-300 -50], [-50 50], [50 300]) for selected methods. The maps represent a top view on the head with nose pointing upwards, + indicate electrodes. The average ERD quality measures over subjects are shown in the legend.

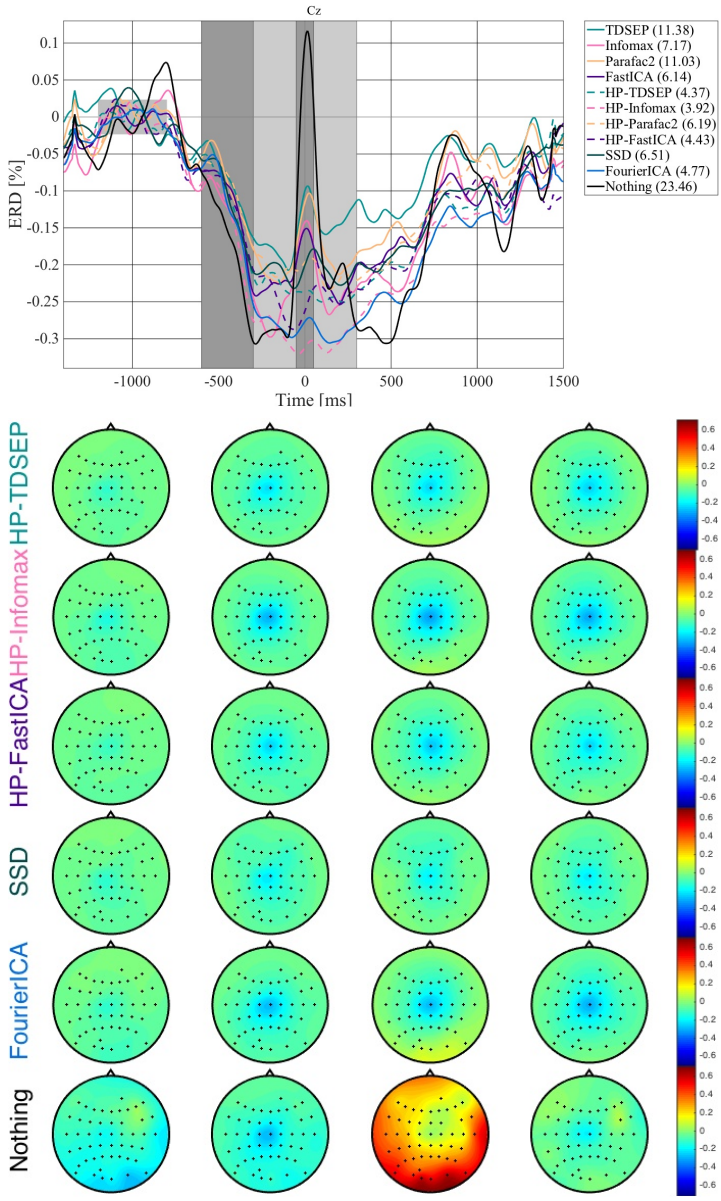


Figure 2: Grand-average ERD/ERS for 18 subjects recorded during self-paced foot movements in the beta band (15-30 Hz), aligned to EMG peak activity. The plots show time courses of data reconstructed from neural ICs (and for SSD with the ten components with highest SNR) at channel Cz and series of ERD maps in the marked intervals, as in Figure 1. The average ERD quality measures over subjects are shown in the legend.

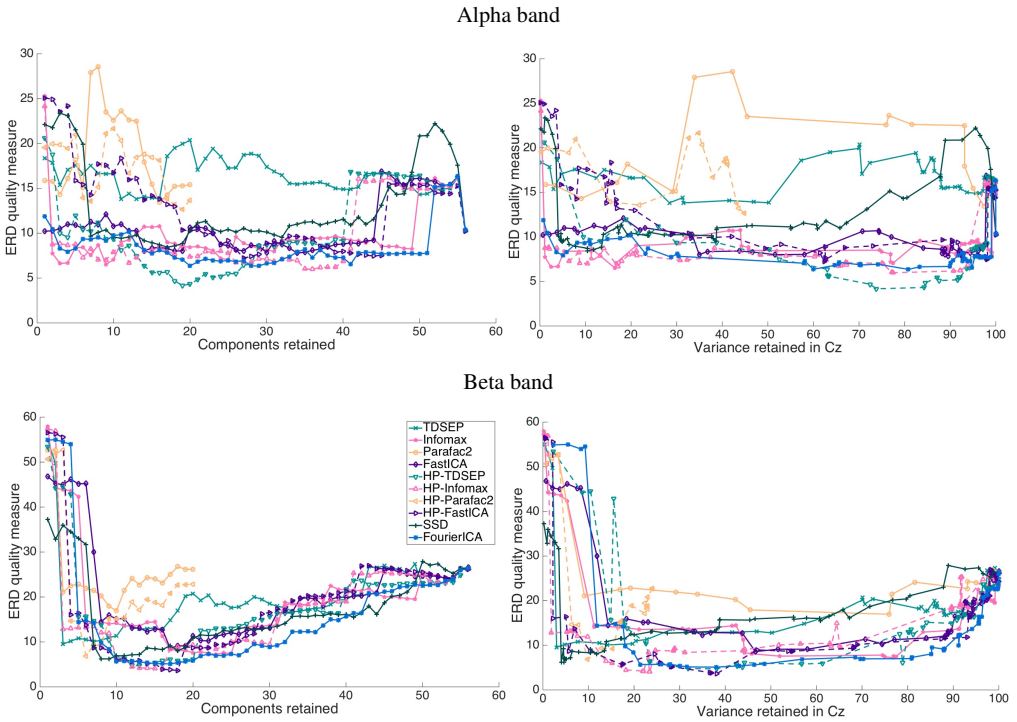


Figure 3: ERD quality measure in dependence of the number of components retained (left) and the variance retained (right), for the alpha and beta bands. Lower is better.

almost completely eliminate the artifact while maintaining the ERD. However, artifact removal seems to be more difficult in the alpha band, where the cleaned ERD is considerably smaller than the ERD obtained on the raw data.

In the bottom parts of the figures, scalp maps are shown for some of the best performing methods. We see that all these methods result in scalp maps similar to the 'Nothing' condition at times with no movement artifact, and do not change during the movement artifact.

Figure 3 shows the ERD quality measure as a function of the number of components retained (left column), the percent variance of data retained in the Cz channel (right column) for both the alpha band (top row) and the beta band (bottom row). For SSD, components were removed in order of decreasing SNR while components were removed in order of decreasing probability of being neural as determined by IC_MARC for the other methods. As indicated by Figures 1 and 2 which show the case of retaining all ICs whose highest probability was for the neural class, Figure 3 shows that high-pass filtering the data at a high cut-off frequency improves the ERD quality measure for all methods (their dashed lines lie below their solid lines). Parafac2 obtains poor (high) ERD quality measures in both bands. For the alpha band, high-pass filtered TDSEP obtains the best (lowest) ERD quality measure, and even manages to do so while retaining a large proportion of data variance (top right plot). The performances of high-pass filtered Extended Infomax and FourierICA are very similar to that of high-pass filtered TDSEP. For the beta band, SSD, FourierICA, and the high-pass filtered ICA methods obtain similar performances.

4 Discussion

In this paper, we analysed and compared the artifact reduction capabilities of the three most common time-domain ICA methods (Extended InfoMax, FastICA, TDSEP) with three other linear decomposition methods (Fourier-ICA, SSD and Parafac2). We used an automatic artifact classifier and a data set from 18 subjects who performed self-

paced foot movements. Movements are well-known to be preceded by an ERD of alpha and beta band rhythms over sensorimotor areas, and we evaluated the ability of the compared methods to remove a clearly visible event-locked muscle artifact while maintaining ERD.

We found that several methods, including the three ICA methods, were able to remove much of the movement artifact, but, as we might expect, not without losing at least some of the neural signal as well. In the beta band, the ERD contamination by the movement artifact is manifested in a narrower time window than in the alpha band, and the artifact also seems easier to correct.

We evaluated the methods' dependence on the number of source components retained and the variance these explain. It is reassuring that the performances of the methods, relative to each other, remain at about the same level for all numbers of components retained and explained variances. This indicates that there are indeed true differences between the methods that do not strongly depend on whether a strict or mild cleaning policy is used. Also, the best performing methods yielded good ERD quality measures over a long range of retained components or retained data variance. This means that the quality of data cleaning is robust to the choice of cleaning policy, as long as it is not too extreme.

With respect to each of the methods' performances, let us first note the importance of adequate filtering. All three ICA methods and Parafac2 consistently achieved better artifact reduction performance when the data had been high-pass filtered at the cut-off frequency just below the frequency band of interest before decomposition. Filtering might guide the decomposition towards extracting the components that explain the activity we are interested in. That is, if we are not interested in low frequencies in further analysis, we may benefit from removing them *before* ICA decomposition. This effect seems to be relevant, probably because the low-frequency parts of an EEG signal contain a large portion of its variance.

In both the alpha and beta bands, the three ICA methods (with high-pass filtering) and Fourier-ICA performed best. This is especially interesting since the observed muscle artifacts are not occurring independently from motor planning neural activity –which clearly violates ICA's assumptions. While a co-activiation of artifacts and neural activity is quite common in practice, our results suggest that ICA may still be a sensible, albeit not perfect choice, even in those settings. This is in line with the findings from McMenamin et al. (2010).

In contrast to Delorme et al. (2012), we observed no consistent performance differences between the three compared ICA methods. This is probably because our performance criterion, which quantifies a neural phenomenon in the cleaned data, is not as sensitive as their dipolarity measure. Our results suggest that the choice of the ICA method may often not result in strong differences in data quality (which is probably why different EEG researchers use different ICA methods with similar success). In our data set, high-pass filtering was far more important than the choice of the ICA method.

However, the tensor decomposition method Parafac2 cannot be recommended for artifact removal. All results showed that Parafac2 performs worse than the other methods. Since data was epoched in one-second epochs before running Parafac2, some epochs would likely have contained only part of the motor artifact while others could have contained two motor artifacts. Even though the spatial pattern for the artifact would be the same across epochs, the temporal pattern would vary to such an extent that it would be unlikely to be found as a consistent component across epochs by Parafac2. Hence Parafac2 is probably better suited to other purposes, for example for the extraction of event-related potentials (ERPs) in settings where either none or one ERP is expected in each epoch (Weis et al., 2010).

On the other hand, ICA, Fourier-ICA, and SSD all had some success in removing the artefact. While the ICA methods and Fourier-ICA perform slighly better than SSD, both in the alpha and beta bands, SSD achieves decent results. SSD is designed to increase the signal-to-noise ratio of oscillatory sources, and it is therefore not surprising that it can be suitable to separate artifacts (=noise) from oscillatory neural signals. Because SSD is faster to evaluate, it may be a good compromise between the time it takes to decompose the data and the quality of artifact separation.

References

Acar, Evrim, Aykut-Bingol, Canan, Bingol, Haluk, Bro, Rasmus, and Yener, Bülent. Multiway analysis of epilepsy tensors. *Bioinformatics*, 23(13):i10–i18, 2007.

Albera, L., Kachenoura, A., Comon, P., Karfoul, A., Wendling, F., Senhadji, L., and Merlet, I. ICA-based EEG

denoising: a comparative analysis of fifteen methods. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 60(3):407 – 418, 2012.

Andersson, C. A. and Bro, R. The n-way toolbox for matlab. *Chemom.Intell.Lab.Syst.*, 1(52):1–4, 2000.

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

Belouchrani, Adel, Abed-Meraim, Karim, Cardoso, Jean-François, and Moulines, Eric. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

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.

Carroll, J Douglas and Chang, Jih-Jie. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 35(3):283–319, 1970.

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.

Crespo-Garcia, Maite, Atienza, Mercedes, and Cantero, Jose L. Muscle artifact removal from human sleep EEG by using independent component analysis. *Ann. Biomed. Eng.*, 36:467–475, 2008.

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.

De Vos, Maarten, De Lathauwer, Lieven, Vanrumste, Bart, Van Huffel, Sabine, and Van Paesschen, Wim. Canonical decomposition of ictal scalp eeg and accurate source localisation: principles and simulation study. *Computational intelligence and neuroscience*, 2007, 2007.

Deburchgraeve, Wouter, Cherian, PJ, De Vos, Maarten, Swarte, RM, Blok, JH, Visser, Gerhard Henk, Govaert, Paul, and Van Huffel, Sabine. Neonatal seizure localization using parafac decomposition. *Clinical Neurophysiology*, 120(10):1787–1796, 2009.

Delorme, A. and Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*, 134(1):9–21, Mar 2004.

Delorme, Arnaud, Palmer, Jason, Onton, Julie, Oostenveld, Robert, and Makeig, Scott. Independent EEG sources are dipolar. *PloS One*, 7(2):e30135, 2012.

Fitzgibbon, Sean P., Powers, David M. W., Pope, Kenneth J., and Clark, C. Richard. Removal of EEG noise and artifact using blind source separation. *Clinical Neurophysiology*, 24(3):232–243, 2007.

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

Goncharova, I. I., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, 114:1580–1593, 2003.

Groppe, David M., Makeig, Scott, and Kutas, Marta. Identifying reliable independent components via split-half comparisons. *NeuroImage*, 45(4):1199 – 1211, 2009.

Harshman, Richard. Parafac2: Extensions of a procedure for explanatory factor analysis and multidimensional scaling. *The Journal of the Acoustical Society of America*, 51(1A):111–111, 1972. doi: http://dx.doi.org/10.1121/1.1981298.

Harshman, Richard A. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 1970.

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.

Hoffmann, Sven and Falkenstein, Michael. The correction of eye blink artefacts in the EEG: a comparison of two prominent methods. *PLoS One*, 3(8):e3004, 2008.

Hyvärinen, Aapo. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

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

Hyvärinen, Aapo, Karhunen, Juka, and Oja, Erkki. *Independent Component Analysis*. John Wiley & Sons, New York, 2001.

Hyvärinen, Aapo, Ramkumar, Pavan, Parkkonen, Lauri, and Hari, Riitta. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. *Neuroimage*, 49:257–271, 2010.

Jung, Tzyy-Ping, Makeig, Scott, Humphries, Colin, Lee, Te-Won, Mckeown, Martin J., Iragui, Vicente, and Sejnowski, Terrence J. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37:163–178, 2000.

Kierkels, J.J.M., van Boxtel, G.J.M., and Vogten, L.L.M. A model-based objective evaluation of eye movement correction in eeg recordings. *Biomedical Engineering, IEEE Transactions on*, 53(2):246–253, 2006.

Kiers, Henk AL, Ten Berge, Jos MF, and Bro, Rasmus. Parafac2-part i. a direct fitting algorithm for the parafac2 model. *Journal of Chemometrics*, 13(3-4):275–294, 1999.

Klemm, Matthias, Haueisen, Jens, and Ivanova, Galina. Independent component analysis: comparison of algorithms for the investigation of surface electrical brain activity. *Medical & biological engineering & computing*, 47(4):413–423, 2009.

Lee, Te-Won, Girolami, Mark, and Sejnowski, Terrence J. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11(2):417–441, 1999.

Makeig, Scott, Bell, Anthony J., Jung, Tzyy-Ping, and Sejnowski, Terrence J. Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, 8:145–151, 1996.

McMenamin, Brenton W., Shackman, Alexander J., Maxwell, Jeffrey S., Bachhuber, David R. W., Koppenhaver, Adam M., Greischar, Lawrence L., and Davidson, Richard J. Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG. *NeuroImage*, 49:2416–2432, 2010.

Muthukumaraswamy, Suresh D. High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Frontiers in human neuroscience*, 7, 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.

Olbrich, Sebastian, Jödicke, Johannes, Sander, Christian, Himmerich, Hubertus, and Hegerl, Ulrich. Ica-based muscle artefact correction of EEG data: What is muscle and what is brain?: Comment on McMenamin et al. *NeuroImage*, 54(1):1 – 3, 2011.

Paulick, Claudia, Wright, Marvin N, Verleger, Rolf, and Keller, Karsten. Decomposition of 3-way arrays: A comparison of different parafac algorithms. *Chemometrics and Intelligent Laboratory Systems*, 137:97–109, 2014.

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

Pignat, Jean Michel, Koval, Oleksiy, Ville, Dimitri Van De, Voloshynovskiy, Sviatoslav, Michel, Christoph, and Pun, Thierry. The impact of denoising on independent component analysis of functional magnetic resonance imaging data. *Journal of Neuroscience Methods*, 213(1):105 – 122, 2013.

Romero, Sergio, Mañanas, Miguel A., and Barbanoj, Manel J. 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, 2008.

Safieddine, Doha, Kachenoura, Amar, Albera, Laurent, Birot, Gwénaël, Karfoul, Ahmad, Pasnicu, Anca, Biraben, Arnaud, Wendling, Fabrice, Senhadji, Lotfi, and Merlet, Isabelle. Removal of muscle artifact from EEG data: comparison between stochastic (ICA and CCA) and deterministic (EMD and wavelet-based) approaches. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–15, 2012.

Scheeringa, René, Fries, Pascal, Petersson, Karl-Magnus, Oostenveld, Robert, Grothe, Iris, Norris, David G., Hagoort, Peter, and Bastiaansen, Marcel C.M. Neuronal dynamics underlying high- and low-frequency EEG oscillations contribute independently to the human BOLD signal. *Neuron*, 69(3):572 – 583, 2011.

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

Vanderperren, Katrien, De Vos, Maarten, Mijović, B, Ramautar, JR, Novitskiy, Nikolay, Vanrumste, Bart, Stiers, Peter, Van den Bergh, BRH, Wagemans, Johan, Lagae, Lieven, et al. Parafac on erp data from a visual detection task during simultaneous fmri acquisition. In *Proc. of the International Biosignal Processing Conference., Berlin, Germany*, volume 103, pp. 1–4, 2010.

Vazquez, R. Romo, Velez-Perez, H., Ranta, R., Dorr, V. Louis, Maquin, D., and Maillard, L. Blind source separation, wavelet denoising and discriminant analysis for EEG artefacts and noise cancelling. *Biomedical Signal Processing and Control*, 7(4):389 – 400, 2012.

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.

Vigário, Ricardo, Särelä, Jaakko, Jousmiki, V, Hämäläinen, Matti, and Oja, Erkki. Independent component approach to the analysis of EEG and MEG recordings. *Biomedical Engineering, IEEE Transactions on*, 47(5): 589–593, 2000.

Vigário, Ricardo Nuno. Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalography and clinical neurophysiology*, 103(3):395–404, 1997.

Wallstrom, Garrick L., Kass, Robert E., Miller, Anita, Cohn, Jeffrey F., and Fox, Nathan A. Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods. *Psychophysiology*, 53:105–19, 2004.

Weis, Martin, Jannek, Dunja, Roemer, Florian, Guenther, Thomas, Haardt, Martin, and Husar, Peter. Multidimensional parafac2 component analysis of multi-channel eeg data including temporal tracking. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 5375–5378. IEEE, 2010.

Winkler, Irene, Debener, Stefan, Müller, Klaus-Robert, and Tangermann, Michael. On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In *IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4101–4105, 2015a.

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

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, 2015c.

Zakeri, Z., Assecondi, S., Bagshaw, A.P., and Arvanitis, T.N. Influence of signal preprocessing on ICA-based EEG decomposition. In *XIII MEDICON 2013*, pp. 734–737.

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

Ziehe, Andreas, Laskov, Pavel, Nolte, Guido, and Müller, Klaus-Robert. 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, 2004.

Appendix E

Multi-linear Discriminant Analysis with Tucker and PARAFAC Structures optimised on the Stiefel Manifold

Multi-linear Discriminant Analysis with Tucker and PARAFAC Structures optimised on the Stiefel Manifold

Laura Frølich, Tobias S. Andersen, and Morten Mørup,

***Abstract*—The primary objective of this paper is to propose new discriminant feature extraction methods for multi-way/tensor data, referred to as Multi-linear Discriminant Analysis (MDA) methods. Whereas existing MDA methods use heuristic optimisation procedures based on an ambiguous Tucker structure, we propose to optimise MDA rigorously, i.e. monotonously with convergence guarantees, using manifold optimisation. We also endow the MDA methods with the PARAFAC structure. We contrast the proposed MDA methods to conventional MDA methods and unsupervised multi-way feature extraction approaches, comparing the methods on binary single-trial classification on two electroencephalography data sets, one taken from a BCI competition. We find that Manifold optimisation substantially improves the MDA objective functions when compared to the existing MDA procedures. However, when inspecting classification performances we do not find substantial differences between the supervised methods, but observe substantially better performances compared to unsupervised feature extraction, even when unsupervised models have many components. Notably, even though the MDA procedures were applied to raw BCI data, their performances were on par with competition participants' results based on ample pre-processing. Finally, we empirically find that the PARAFAC structure is less susceptible to ambiguous representations, thereby providing more easily interpretable results.**

***Index Terms*—Linear Discriminant Analysis, LDA, Multi-linear Discriminant Analysis, MDA, Electroencephalography, EEG, Tensor, Manifold optimisation.**

I. Introduction

LINEAR Discriminant Analysis (LDA) is a widely used method for feature extraction/dimensionality reduction and classification [1], [2]. When observations are arranged as vectors that are not too high-dimensional relative to the number of observations, LDA often obtains high classification rates [2, p. 111], especially taking its relatively simple formulation and estimation into account. Conversely, when data is high-dimensional, standard LDA runs into singularity problems. When data is multi-way, i.e. consisting of observations that have more than one mode the simplest way to handle such multi-way data is to vectorise it in order for standard methods to process it. However this leads to high-dimensional observations. Instead, the intrinsic multi-way structure of such data can be retained throughout analyses. Research on supervised multi-linear (sometimes referred to as tensor and multi-way) methods that exploit the intrinsic multi-way structure of data has proliferated within the last decade. The individual dimensions of data can be referred to as "modes". Multi-linear methods both ameliorate the challenge of high-dimensional observations and allows interactions between modes to be taken into account throughout analyses.

Several strategies have been pursued using multi-linear methods for classification. Some have used unsupervised tensor methods for feature extraction followed by a supervised classification step [3]–[11]. Others have employed unsupervised tensor decomposition methods were separately for each class making the procedure as a whole supervised [12]–[16]. Additionally, the loss function optimised by unsupervised tensor decomposition methods has been combined with a supervised loss function and a regularisation term, respectively [17], [18]. However, a more direct approach to learning the most discriminative projections of tensor data has been to incorporate the assumed tensor structure of data into the formulation of a supervised loss function [19]–[24]. In particular, this has been explored in the context of Multi-linear Discriminant Analysis (MDA) that quantifies the degree of discrimination between classes achieved by a set of mode-specific projection matrices [25]–[31], thus generalizing LDA to multi-way data. As MDA exploits the multi-linear structure of data instead of vectorising observations it addresses the problem of high-dimensional observations that LDA suffers from when multi-linear data is vectorised. Existing MDA methods assume a Tucker structure of data and find the projection matrices using heuristic optimisation procedures in which the projection matrix for each mode is estimated by the singular value decomposition [29] or as a standard [31] generalized [26] eigenvalue problem.

L. Frølich was with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark e-mail: (lffr@dtu.dk).

T. S. Andersen and M. Mørup are with the Department of Applied Mathematics and Computer Science, Technical University of Denmark.

A. Contributions

We aimed to investigate the following:

1) What are the gains from optimising MDA rigorously instead of via previously proposed heuristic methods?
2) Is the more flexible Tucker structure necessary in MDA or can a PARAFAC type of structure suffice?
3) How do the classification performances using features extracted by MDA compare to features extracted using standard unsupervised multi-linear decomposition approaches?

To investigate 1), we propose to optimise the MDA objective functions rigorously using the conjugate gradient method on a manifold specified jointly for all modes. We compare the objective function values and classification rates of the existing MDA methods to those obtained from our rigorous optimisation.

The Tucker structure permits all interactions between factors from different modes of the tensor while the PARAFAC structure is more restrictive by only considering interactions within the same factors in each mode. While the Tucker models are subject to rotational invariance, the PARAFAC structure is more constrained and may thereby provide unique representations. These characteristics make the PARAFAC model more easily interpretable and thus attractive if the PARAFAC structure is sufficient for modeling data. We investigate 2) by extending MDA to also have PARAFAC structure. For completion we also consider the logistic regression framework proposed in [20] having both PARAFAC and Tucker structure.

To investigate 3), i.e. the utility of MDA over existing unsupervised multi-linear feature extraction approaches we compared the performance of MDA to the classification rates obtained when features were extracted using the following unsupervised multi-way decomposition approaches; PARAFAC [32], [33] and the PARAFAC2 [34], [35] model as well as the Tucker and Tucker2 [36] model.

We compared the methods in their performance for single-trial classification of electroencephalography (EEG) data. EEG data measures the electrical potential over time from multiple electrodes placed at the scalp and has a natural multi-dimensional structure, consisting of e.g. channels, time, trials, and subjects. However, these multi-way structures are lost when data is averaged or concatenated across dimensions, as is common in analyses of EEG data [37]. Tensor methods, on the other hand, are able to exploit the inherent multi-dimensional structure. By retaining the the multi-way structure, it may be possible to learn classifiers of EEG single trials that benefit from the structure of space and time. Single-trial classification of event-related potentials (ERPs) in EEG data is an important problem as several Brain-Computer Interface (BCI) systems rely on ERP classification (e.g. the P300 speller [38], Steady State Visually Evoked Potential (SSVEP) [39], Steady State Vibration Somatosensory Evoked Potential [40] paradigms, and the Predictive Auditory Spatial Speller with two-dimensional stimuli [41]). Although an SSVEP-based BCI system that greatly improves upon previously established information transfer rates has recently been demonstrated in [42], improved single-trial classification rates could increase this even further.

Work using supervised tensor methods to classify EEG data has appeared since at least a decade ago [19]. Some methods propose a first step in which the original data dimensionality is reduced through feature extraction using an unsupervised tensor decomposition method such as PARAFAC or Tucker, followed by a supervised step in which the extracted features are given as input to a standard classifier, for example a support vector machine (SVM) or the K-nearest neighbor (KNN) method [43]–[52]. Others have taken a more direct approach, directly supervising the tensor decomposition step itself [19], [20], [29], [53]–[59]. All these proposals have employed standard pre-processing of the EEG data such as low-pass filtering at a low cut-off frequency (<50 Hz) and noise removal, some even using independent component analysis followed by manual selection of artefactual components [55], [57]. Others have performed a spectral decomposition of the data, adding frequency as a new dimension to the data tensor [29], [43]–[45], [47], [48], [50]–[53], [55], [60], [61]. We extensively compare existing unsupervised and supervised approaches to our proposed extensions of MDA.

II. Methods

Linear Discriminant Analysis (LDA) aims to maximise the between-class scatter while minimising the within-class scatter. Assume there are N observations of J dimensional vectors and refer to the n^{th} observation as $\boldsymbol{x}_n$. Let C denote the number of classes and N_c be the number of observations in class c. Also, denote the set of indices of observations belonging to class c by $\mathcal{C}_c$. Let $\bar{\boldsymbol{x}}$ be the mean of all N observations and $\bar{\boldsymbol{x}}_c$ be the mean of observations from class c. Finally, let the matrix $\mathbf{U}_{LDA}$, with no specific structure, contain projection vectors in its columns. Hence, a K-factor model would use a $J \times K$ dimensional $\mathbf{U}_{LDA}$. Define the within-class scatter matrix ($\mathbf{W}_{LDA}$) and between-class scatter matrix ($\mathbf{B}_{LDA}$) as:

$$\begin{aligned}\mathbf{W}_{LDA} &= \sum_{c=1}^{C} \sum_{n \in \mathcal{C}_c} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_c)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_c)^\top \\ \mathbf{B}_{LDA} &= \sum_{c=1}^{C} N_c (\bar{\boldsymbol{x}}_c - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_c - \bar{\boldsymbol{x}})^\top, \end{aligned} \tag{1}$$

To find the most discriminative projection vectors, an objective function should be maximised as a function of $\mathbf{U}_{LDA}$. Different functions have been used to quantify the within-class scatter relative to the between-class scatter. The scatter-ratio objective function can be interpreted as maximising the sum of squared differences between the class means while minimising the sum of squared differences to the mean within each class and has been used in MDA methods previously [26], [29]. The formulation of the scatter-ratio objective function is:

$$\frac{Tr(\mathbf{U}_{LDA}^{\top}\mathbf{B}_{LDA}\mathbf{U}_{LDA})}{Tr(\mathbf{U}_{LDA}^{\top}\mathbf{W}_{LDA}\mathbf{U}_{LDA})} \tag{2}$$

The scatter-difference objective function has also been considered [3], [29]. The rationale for this objective is that that this objective is equivalent to the solution of (2) when ζ is set as the Lagrange multiplier [3], [29]. This objective is defined as:

$$\begin{aligned} &Tr(\mathbf{U}_{LDA}^{\top}\mathbf{B}_{LDA}\mathbf{U}_{LDA}) - \\ &\zeta Tr(\mathbf{U}_{LDA}^{\top}\mathbf{W}_{LDA}\mathbf{U}_{LDA}). \end{aligned} \tag{3}$$

The trace of matrix ratios objective is appealing since it is optimised by the eigenvectors corresponding to the largest eigenvalues of $\mathbf{W}_{LDA}^{-1}\mathbf{B}_{LDA}$ [1], and is formulated as:

$$\begin{aligned} &Tr\Big(\big(\mathbf{U}_{LDA}^{\top}\mathbf{W}_{LDA}\mathbf{U}_{LDA}\big)^{-1} \\ &\mathbf{U}_{LDA}^{\top}\mathbf{B}_{LDA}\mathbf{U}_{LDA}\Big). \end{aligned} \tag{4}$$

The ratio of determinants objective maximises the volume between the class means to the volume spanned by the difference within the classes between observations and their class mean has also been used previously [31], [62]. It is defined as:

$$\frac{det(\mathbf{U}_{LDA}^{\top}\mathbf{B}_{LDA}\mathbf{U}_{LDA})}{det(\mathbf{U}_{LDA}^{\top}\mathbf{W}_{LDA}\mathbf{U}_{LDA})}. \tag{5}$$

Although, this objective differs from (4) the solution has the same stationary points as (4) as shown in Appendix A. We therefore presently consider the objective functions for LDA given by (2) and (4).

A. Multi-linear methods

For clarity of exposition, we limit our presentation to matrix observations. Assume there are N $J_1 \times J_2$ matrix observations. We refer to the n^{th} observation as $\mathbf{X}_n$. Let $\bar{\mathbf{X}}$ be the mean of all N observations and $\bar{\mathbf{X}}_c$ be the mean of observations from class c. The operator $vec(\mathbf{X})$ vectorises the matrix $\mathbf{X}$ column-wise. We will use the Kronecker product and the Khatri-Rao product. The Kronecker product of an $m \times \ell$ matrix $\mathbf{A}$ with elements $a_{i,j}$ and a $p \times q$ matrix $\mathbf{C}$ is defined as [63]:

$$\mathbf{A} \otimes \mathbf{C} = \begin{pmatrix} a_{1,1}\mathbf{C} & a_{1,2}\mathbf{C} & \cdots & a_{1,\ell}\mathbf{C} \\ a_{2,1}\mathbf{C} & a_{2,2}\mathbf{C} & \cdots & a_{2,\ell}\mathbf{C} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}\mathbf{C} & a_{m,2}\mathbf{C} & \cdots & a_{m,\ell}\mathbf{C} \end{pmatrix}.$$

The Khatri-Rao product is the column-wise Kronecker product and is denoted by $\odot$. That is, column j of the matrix $\mathbf{D} = \mathbf{A} \odot \mathbf{C}$ is the Kronecker product of the j^{th} columns of $\mathbf{A}$ and $\mathbf{C}$: $\mathbf{D}_{:,j} = \mathbf{A}_{:,j} \otimes \mathbf{C}_{:,j}$ [63]. Hence the matrices $\mathbf{A}$ and $\mathbf{C}$ must have the same number of columns for their Khatri-Rao product to be defined. Finally, we define the matricizing operation which rearranges the elements of a tensor into matrix form. We define this operation as:

$$\mathcal{X} \in \mathbb{R}^{J_1 \times \ldots \times J_P} \rightarrow \mathbf{X}_{(p)} \in \mathbb{R}^{J_p \times J_1 \ldots J_{p-1} J_{p+1} \ldots J_P} \tag{6}$$

when matricising along the p^{th} mode [63]. For matrices, matricizing along the first mode does not alter the matrix while matricizing along the second mode transposes the matrix.

In the following, we denote the projection matrix for mode p by $\mathbf{U}^{(p)}$.

1) Unsupervised feature extraction methods: Just as the unsupervised method Principal Component Analysis can be used to decompose data when observations are vectors, the multi-linear methods Tucker, Tucker2, PARAFAC, and PARAFAC2 are multi-way versions of unsupervised decomposition methods.

Using the matricizing operation and Kronecker product, the Tucker model for a 3D tensor, $\mathcal{X}$ is [36], [63]:

$$\begin{aligned} \mathbf{X}_{(1)} &\approx \mathbf{U}^{(1)}\mathbf{G}_{(1)}(\mathbf{U}^{(3)} \otimes \mathbf{U}^{(2)})^{\top} \\ \mathbf{X}_{(2)} &\approx \mathbf{U}^{(2)}\mathbf{G}_{(2)}(\mathbf{U}^{(3)} \otimes \mathbf{U}^{(1)})^{\top} \\ \mathbf{X}_{(3)} &\approx \mathbf{U}^{(3)}\mathbf{G}_{(3)}(\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^{\top}, \end{aligned} \tag{7}$$

where $\mathcal{G}$ is a tensor, usually referred to as the core array, that gives the strengths of interactions between each combination of factors from the three projection matrices. The matrix $\mathbf{G}_{(p)}$ is the p-mode matricization of $\mathcal{G}$. A Tucker model that leaves one of the modes uncompressed is referred to as the Tucker2 model [63]. For example, if the identity matrix of appropriate size is used as the projection matrix in mode 3, this would be a Tucker2 model. The model formulation for the n^{th} observation when the third mode is left uncompressed is given by:

$$\mathbf{X}_n \approx \mathbf{U}^{(1)}\mathbf{G}_n{\mathbf{U}^{(2)}}^{\top}, \tag{8}$$

where the $K_1 \times K_2$ matrix $\mathbf{G}_n$ is a compressed representation of the n^{th} observation, $\mathbf{X}_n$.

The PARAFAC model is a more restricted version of the Tucker model, where all off-diagonal elements in the core array are zero. The effect of this is that the only possible interactions are between columns of the same index in the projection matrices. That is, column $\mathbf{U}^{(1)}_{:,j}$ can only interact with $\mathbf{U}^{(2)}_{:,j}$ and $\mathbf{U}^{(3)}_{:,j}$. Hence, the strength of the interaction can be absorbed into any one of the projection matrices, eliminating the need for a core array. The PARAFAC model can thus be written as follows [32], [33], [63]:

$$\begin{aligned}\mathbf{X}_{(1)} &\approx \mathbf{U}^{(1)}(\mathbf{U}^{(3)} \odot \mathbf{U}^{(2)})^\top \\ \mathbf{X}_{(2)} &\approx \mathbf{U}^{(2)}(\mathbf{U}^{(3)} \odot \mathbf{U}^{(1)})^\top \\ \mathbf{X}_{(3)} &\approx \mathbf{U}^{(3)}(\mathbf{U}^{(2)} \odot \mathbf{U}^{(1)})^\top. \end{aligned} \tag{9}$$

While both the Tucker and PARAFAC models assume the same factors for all observations, only interacting at different strengths, a more flexible model is obtained by allowing some variation in the factors for one mode over observations. This is done in the PARAFAC2 model, with the constraint that the cross-product of each factor must be the same across observations. Assume the variation is allowed over the second mode. Then the PARAFAC2 model for observation n is given by [34], [35]:

$$\mathbf{X}_n \approx \mathbf{U}^{(1)} diag(\mathbf{U}^{(3)}_{(n,:)})(\mathbf{F}\mathbf{H}_n), \tag{10}$$

where $\mathbf{H}_n\mathbf{H}_n^\top = \mathbf{I} \forall n$. As such the extracted second mode loading (i.e., $\mathbf{F}\mathbf{H}_n$) has the same covariance structure $\mathbf{F}\mathbf{H}_n\mathbf{H}_n^\top\mathbf{F}^\top = \mathbf{F}\mathbf{F}^\top$ for all the observations.

Notably, the Tucker and Tucker2 models are not unique since any invertible matrix $\mathbf{Q}$ can be multiplied with $\mathbf{U}^{(p)}$ and the core array $\mathbf{G}_{(p)}$ multiplied by its inverse. In contrast, no such ambiguity exists for the PARAFAC model where uniqueness has been established under mild conditions [64]. Uniqueness results have also been established for the PARAFAC2 model [35].

a) Classification: If observations are stored along the third mode, each column in the projection matrix $\mathbf{U}^{(3)}$ (third-mode factor), found by Tucker, PARAFAC, and PARAFAC2 will contain one value per observation. These can be used as features for classification. For Tucker2, the projection matrices for the first and second modes ($\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$) can be used to project trials into lower dimensional spaces. The core array ($\mathcal{G}_n$) of these lower-dimensional representations of observations can be used as classification features.

2) Supervised feature extraction methods: Multilinear discriminant analysis (MDA) aims to find projection matrices that project tensor observations ($\mathcal{X}_n \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_P}$) into a maximally discriminative lower dimensional space, $\mathbb{R}^{K_1 \times K_2 \times \ldots \times K_P}$ with $K_p \leq J_p$, $p = 1, 2, \ldots, P$. The projection matrix for mode p thus has the dimensions $J_p \times K_p$.

Define the tensor-generalisations of $\mathbf{W}_{LDA}$ and $\mathbf{B}_{LDA}$ as:

$$\begin{aligned}\mathbf{W} &= \sum_{c=1}^{C} \sum_{n \in \mathcal{C}_c} vec(\mathbf{X}_n - \bar{\mathbf{X}}_c) vec(\mathbf{X}_n - \bar{\mathbf{X}}_c)^\top \\ \mathbf{B} &= \sum_{c=1}^{C} N_c vec(\bar{\mathbf{X}}_c - \bar{\mathbf{X}}) vec(\bar{\mathbf{X}}_c - \bar{\mathbf{X}})^\top. \end{aligned} \tag{11}$$

These can be generalised to general tensors, $\mathcal{X}_n$, by substituting all occurrences of the matrices $\mathbf{X}_n$, $\bar{\mathbf{X}}_c$, and $\bar{\mathbf{X}}$ by their tensor counterparts $\mathcal{X}_n$, $\bar{\mathcal{X}}_c$, and $\bar{\mathcal{X}}$.

By setting $\mathbf{U} = \mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)}$ and substituting this for $\mathbf{U}_{LDA}$, the objective functions (2)-(5) become directly applicable to matrix observations. Their further generalisation to observations with P modes is straight-forward by defining $\mathbf{U} = \mathbf{U}^{(P)} \otimes \mathbf{U}^{(P-1)} \ldots \mathbf{U}^{(1)}$.

The methods Discriminant Analysisis with TEnsor Representation (DATER) [26] and Constrained Multilinear Discriminant Analysis (CMDA) [29] aim to optimise the scatter ratio objective function (2), substituting $\mathbf{W}$ and $\mathbf{B}$ for $\mathbf{W}_{LDA}$ and $\mathbf{U}_{LDA}$ with $\mathbf{U}$. Another existing MDA method that optimises the ratio of determinants has also been proposed [31]. We refer to this method as DATEReig. All three methods are based on an alternating optimisation procedure estimating each mode iteratively one at a time. When updating mode p, they project $\mathbf{W}$ and $\mathbf{B}$ unto all modes except mode p:

$$\begin{aligned}\mathbf{W}^{\tilde{p}}_{proj} &= \sum_{c=1}^{C} \sum_{n \in \mathcal{C}_c} (\mathbf{X}_n - \bar{\mathbf{X}}_c)_{(p)} \mathbf{U}^{\tilde{p}\top} \mathbf{U}^{\tilde{p}} (\mathbf{X}_n - \bar{\mathbf{X}}_c)^\top_{(p)} \\ \mathbf{B}^{\tilde{p}}_{proj} &= \sum_{c=1}^{C} N_c (\bar{\mathbf{X}}_c - \bar{\mathbf{X}})_{(p)} \mathbf{U}^{\tilde{p}\top} \mathbf{U}^{\tilde{p}} (\bar{\mathbf{X}}_c - \bar{\mathbf{X}})^\top_{(p)}, \end{aligned} \tag{12}$$

where

$$\mathbf{U}^{\tilde{p}} = \mathbf{U}^{(P)} \otimes \ldots \mathbf{U}^{(p+1)} \otimes \mathbf{U}^{(p-1)} \ldots \mathbf{U}^{(1)}. \tag{13}$$

CMDA then updates $\mathbf{U}^{(p)}$ by setting it equal to the first K_p singular vectors of $\left(\mathbf{W}^{\tilde{p}}_{proj}\right)^{-1} \mathbf{B}^{\tilde{p}}_{proj}$ which was proven in [29] to result in the objective function forming an asymptotically bounded sequence. Since a matrix defined by singular vectors is orthonormal, CMDA in effect uses the orthonormality constraint. DATER instead uses the first K_p generalised eigenvectors of the Generalised Eigenvalue Problem: $\mathbf{B}^{\tilde{p}}_{proj}\mathbf{U}^{(p)} = \mathbf{W}^{\tilde{p}}_{proj}\mathbf{U}^{(p)}\Lambda_k$. Solving the Generalised Eigenvalue Problem leads to $\mathbf{W}^{\tilde{p}}_{proj}$-orthogonality (${\mathbf{U}^{(p)}}^\top \mathbf{W}^{\tilde{p}}_{proj} \mathbf{U}^{(p)} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix [65]). Since the matrix $\mathbf{W}^{\tilde{p}}_{proj}$ is different

for each mode, this means that the projection matrices for the different modes are constrained differently by DATER. Similarly, DATEReig solves the standard eigenvalue problem defined as: $\left(\mathbf{W}_{proj}^{\tilde{p}}\right)^{-1}\mathbf{B}_{proj}^{\tilde{p}}\mathbf{U}^{(p)} = \mathbf{D}\mathbf{U}^{(p)}$, where $\mathbf{D}$ is a diagonal matrix. Hence DATEReig in effect is also subject to orthonormal constraints on the projection matrices. Finally, the method Direct General Tensor Discriminant Analysis (DGTDA) [29] optimises the scatter difference objective function in (3). It does this by iterating over each mode once, setting ζ equal to the largest singular value of $\left(\mathbf{W}_{proj}^{\tilde{p}}\right)^{-1}\mathbf{B}_{proj}^{\tilde{p}}$ when solving for mode p. The projection matrix for mode p is then set equal to the first K_p singular vectors of $\mathbf{B}_{proj}^{\tilde{p}} - \zeta\mathbf{W}_{proj}^{\tilde{p}}$.

Rather than optimising criteria based on multi-linear extensions of LDA it may be advantageous to optimise a measure of classification directly. This was first proposed using logistic regression and a one component PARAFAC structured model [19] and later extended to an arbitrary number of factors [20]:

$$\sum_{n=1}^{N} y_n(w_0 + \psi_{PARAFAC}(\mathbf{X}_n)) - \log(1+\exp(w_0+\psi_{PARAFAC}(\mathbf{X}_n)), \quad (14)$$

such that the probability that observation $\mathbf{X}_n$ belongs to class one is:

$$\frac{1}{1+\exp\left(-\left(w_0+\psi_{PARAFAC}\left(\mathbf{X}_n\right)\right)\right)}, \quad (15)$$

where

$$\psi_{PARAFAC}(\mathbf{X}_n) = Tr({\mathbf{U}^{(1)}}^{\top}\mathbf{X}_n\mathbf{U}^{(2)}) \quad (16)$$

$$= \sum_{k=1}^{K_1}[(\mathbf{U}^{(1)} \odot \mathbf{U}^{(2)})^{\top} vec(\mathbf{X}_n)]_k \quad (17)$$

Thus, the number of factors is the same for both modes ($K_1 = K_2$) whereas the approach does not rely on orthonormality imposed on the projection matrices. Notably, despite the PARAFAC type of structure imposed, the model is not unique. For any two square matrices $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ satisfying $\mathbf{Q}^{(2)}{\mathbf{Q}^{(1)}}^{\top} = \mathbf{I}$, we have that

$$\begin{aligned} &Tr((\mathbf{U}^{(1)}\mathbf{Q}^{(1)})^{\top}\mathbf{X}_n(\mathbf{U}^{(2)}\mathbf{Q}^{(2)})) \\ &= Tr(\mathbf{Q}^{(2)}{\mathbf{Q}^{(1)}}^{\top}{\mathbf{U}^{(1)}}^{\top}\mathbf{X}_n\mathbf{U}^{(2)}) \\ &= Tr({\mathbf{U}^{(1)}}^{\top}\mathbf{X}_n\mathbf{U}^{(2)}), \end{aligned}$$

hampering model interpretation unless additional constraints are imposed.

For comparison, we also considered the following extension of the above logistic regression model to the more flexible Tucker structure that the existing MDA approaches are based on. This allows for interactions between all factors from the different modes. For simplicity, we assume that data observations are matrices, resulting in the following log-likelihood:

$$\sum_{n=1}^{N} y_n(w_0 + \psi_{Tucker}(\mathbf{X}_n)) - \log(1+\exp(w_0+\psi_{Tucker}(\mathbf{X}_n)), \quad (18)$$

where

$$\psi_{Tucker}(\mathbf{X}_n) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}[{\mathbf{U}^{(1)}}^{\top}\mathbf{X}_n\mathbf{U}^{(2)}]_{k_1,k_2}\mathbf{V}_{k_1,k_2} \quad (19)$$

with $\mathbf{V}_{k_1,k_2} = 1$ for $k_1 = k_2$ to remove scaling ambiguities between the projection matrices and the matrix of interaction coefficients, $\mathbf{V}$. As for BDCA, there are no constraints on $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$.

B. MDA based on manifold optimisation and PARAFAC structure

The existing MDA approaches rely on heuristic optimisation procedures employing either eigenvalue decomposition or the singular value decomposition. We presently exploit manifold optimisation as implemented in the recently released *ManOpt* toolbox [66]. This framework allows rigorous optimisation of arbitrary objective functions on a variety of manifolds, as long as their gradients are known. Amongst others, the toolbox has implementations of optimisation over the Stiefel manifold, which consists of orthonormal matrices [67]. By optimising over a cross product of Stiefel manifolds, one for each mode, it is possible to optimise all projection matrices at once while imposing orthonormality constraints. Other constraints can easily be enforced on some or all modes by changing the manifolds in the cross product that the optimisation is performed over.

We propose four new MDA methods by optimising one existing and three new MDA objective functions rigorously on a manifold. We use the cross product manifold of Stiefel manifolds to ensure that the resulting projection matrices are orthonormal and simultaneously optimized all the projection matrices using the conjugate gradient method. The three new objectives are a PARAFAC version of an existing Tucker-structure MDA objective (the scatter-ratio (2)), and a PARAFAC and Tucker version of the trace-ratio objective (4).

We now define orthonormal projection matrices with the Tucker and PARAFAC structures, respectively:

$$\begin{aligned} \mathbf{U}_{Tucker} &= \mathbf{U}^{(P)} \otimes \mathbf{U}^{(P-1)} \ldots \mathbf{U}^{(1)} \\ \mathbf{U}_{PARAFAC} &= \mathbf{U}^{(P)} \odot \mathbf{U}^{(P-1)} \ldots \mathbf{U}^{(1)}. \end{aligned} \quad (20)$$

The objective functions and the names we refer to the methods by are:

Manifold Tucker/PARAFAC Discriminant Analysis with the trace of ratios objective function (ManTDA/ManPDA):

$$Tr\left((\mathbf{U}_s^\top \mathbf{W}\mathbf{U}_s)^{-1}\mathbf{U}_s^\top \mathbf{B}\mathbf{U}_s\right), \tag{21}$$

where the structure variable s is either $Tucker$ or $PARAFAC$.

Manifold Tucker/PARAFAC Discriminant Analysis with the scatter ratio objective function (ManTDA_sr/ManPDA_sr):

$$\frac{tr\left(\mathbf{U}_s^\top \mathbf{B}\mathbf{U}_s\right)}{tr\left(\mathbf{U}_s^\top \mathbf{W}\mathbf{U}_s\right)}. \tag{22}$$

Again, the structure variable s is either $Tucker$ or $PARAFAC$.

C. Uniqueness of MDA

MDA based on the Tucker structure is not unique when considering the objective functions given above. In fact, the projection matrix of each mode can separately be multiplied any orthonormal matrix $\boldsymbol{R}$ without changing the value of the objective function.

To see this consider the scatter ratio objective (22), where we have numerator and denominator terms of the following structure $Tr\left(\mathbf{U}_s^\top \mathbf{M}\mathbf{U}_s\right) = Tr\left(\mathbf{U}_s\mathbf{U}_s^\top \mathbf{M}\right)$. Without loss of generality we let $P = 2$. Let $\mathbf{U} = \mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)}$ and $\tilde{\mathbf{U}} = (\mathbf{U}^{(2)}\mathbf{R}) \otimes \mathbf{U}^{(1)}$. We then obtain:

$$\begin{aligned}\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top =& (\mathbf{U}^{(2)}\mathbf{R}\mathbf{R}^\top \mathbf{U}^{(2)^\top}) \otimes (\mathbf{U}^{(1)}\mathbf{U}^{(1)^\top}) \\ =& \mathbf{U}\mathbf{U}^\top,\end{aligned}$$

where we have made use of the facts that $(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top$ and $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD})$ [68].

Similarly, again considering $P = 2$ without loss of generality, we obtain for the trace of matrix ratio objective (21):

$$\begin{aligned}&Tr\left((\tilde{\mathbf{U}}^\top \mathbf{W}\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^\top \mathbf{B}\tilde{\mathbf{U}}\right) \\ &= Tr\left(((\mathbf{R}^\top \otimes \mathbf{I}_{K_1})\mathbf{U}^\top \mathbf{W}\mathbf{U}(\mathbf{R} \otimes \mathbf{I}_{K_1}))^{-1}\right. \\ &\quad \left.((\mathbf{R}^\top \otimes \mathbf{I}_{K_1})\mathbf{U}^\top \mathbf{B}\mathbf{U}(\mathbf{R} \otimes \mathbf{I}_{K_1}))\right) \\ &= Tr\left(((\mathbf{R} \otimes \mathbf{I}_{K_1})^{-1}(\mathbf{U}^\top \mathbf{W}\mathbf{U})^{-1}(\mathbf{R}^\top \otimes \mathbf{I}_{K_1})^{-1})\right. \\ &\quad \left.(\mathbf{R}^\top \otimes \mathbf{I}_{K_1})\mathbf{U}^\top \mathbf{B}\mathbf{U}(\mathbf{R} \otimes \mathbf{I}_{K_1})\right) \\ &= Tr\left((\mathbf{U}^\top \mathbf{W}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{B}\mathbf{U}\right),\end{aligned}$$

where $\mathbf{I}_{K_1}$ is the $K_1 \times K_1$ identity matrix. Here, we have also made use of the fact that $(\mathbf{A} \otimes \mathbf{B})^{-1} = (\mathbf{A}^{-1} \otimes \mathbf{B}^{-1})$ [68].

For the PARAFAC version of MDA (for P=2) we can consider alternative representations of $\mathbf{U} = \mathbf{U}^{(2)} \odot \mathbf{U}^{(1)}$ by multiplying two orthonormal matrices $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ to form $\tilde{\mathbf{U}} = (\mathbf{U}^{(2)}\mathbf{R}^{(2)}) \odot (\mathbf{U}^{(1)}\mathbf{R}^{(1)})$. Exploiting the property [69]:

$$(\mathbf{U}^{(2)}\mathbf{R}^{(2)}) \odot (\mathbf{U}^{(1)}\mathbf{R}^{(1)}) = (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})(\mathbf{R}^{(2)} \odot \mathbf{R}^{(1)}),$$

we obtain for the term used separately in the numerator and denominator of the scatter ratio objective function (22) :

$$\begin{aligned}&\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top = \\ &(\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})(\mathbf{R}^{(2)} \odot \mathbf{R}^{(1)})(\mathbf{R}^{(2)} \odot \mathbf{R}^{(1)})^\top (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^\top,\end{aligned}$$

and for the objective function in (21):

$$\begin{aligned}&Tr\left((\tilde{\mathbf{U}}^\top \mathbf{W}\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^\top \mathbf{B}\tilde{\mathbf{U}}\right) \\ &= Tr\left(\left((\mathbf{R}^{(2)} \odot \mathbf{R}^{(1)})^\top (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^\top \mathbf{W}\right.\right. \\ &\quad \left.\left.(\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})(\mathbf{R}^{(2)} \odot \mathbf{R}^{(1)})\right)^{-1}\right. \\ &\quad (\mathbf{R}^{(2)} \odot \mathbf{R}^{(1)})^\top (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^\top \mathbf{B} \\ &\quad \left.(\mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})(\mathbf{R}^{(2)} \odot \mathbf{R}^{(1)})\right).\end{aligned}$$

Due to the Khatri-Rao product structure it is no longer clear that the above objective functions for $\tilde{\mathbf{U}}$ can be reduce to the objective functions based on $\mathbf{U}$ except for the trivial situation in which $\mathbf{R}^{(2)}$ and $\mathbf{R}^{(1)}$ are identical permutation matrices. We empirically tested the objective functions where $\mathbf{R}^{(2)} = \mathbf{R}^{(1)}$, $\mathbf{R}^{(2)} = \mathbf{R}^{(1)^\top}$, and $\mathbf{R}^{(2)} \neq \mathbf{R}^{(1)^\top}$ and found that the random orthonormal matrices we generated indeed did not provide equivalent objective function values. Note that the case $\mathbf{R}^{(2)} = \mathbf{R}^{(1)}$ would result in the same objective function value for $BDCA$.

III. Data

In our analysis we considered the following two EEG datasets.

a) Stekelenburg&Vroomen data:: This data set consists of data from Experiment 2 in a set of three experiments performed and described by Stekelenburg and Vroomen [70] containing data from 16 subjects. For our analyses, we used control trials (gray box shown on computer, no sound) and non-verbal auditory trials (clapping (103-107 ms) and tapping of spoon on cup (292-305 ms), gray box on screen). Trials containing values exceeding 150 μV or lower than -150 μV 200 ms prior to or 800 ms after stimulus onset were removed. The baseline of trials, defined as the mean of the 200 ms before stimulus onset, were subtracted. Trials were defined as lasting from stimulus onset until 500ms after stimulus onset. These data were recorded at 512 Hz. We balanced the trials so that there were equally many from each class (2604 trials in total over all subjects and both classes). To make leave-one-subject-out cross-validation possible, we used 50 electrodes common to all subjects.

b) BCI Competition data:: We also compared the methods on dataset II [71] from BCI competition III [72] [1] from a P300 speller paradigm. These data were recorded from two subjects at 240 Hz from 64 electrodes and band-pass filtered during recording between 0.1-60 Hz. We extracted trials from stimulus onset until 667 ms after stimulus onset. For each subject, a training data set containing single-trial labels was available. The test data consisted of EEG recordings and the true spelled letters, but not single-trial labels.

These two data sets represent different challenges. While there are many trials in the BCI data set (61,200 per subject), this data set is unbalanced, with one target trial for every five non-target trials. On the other hand, we balanced the Stekelenburg&Vroomen data set but have much fewer trials for this data set. We performed the analyses for the BCI data for each subject using 5-fold cross-validation (CV) while the Stekelenburg&Vroomen data was analysed by concatenating data from all subjects and performing leave-one-subject-out CV thereby quantifying model generalization to new subjects.

Since compression of the temporal mode by the considered models should be enough to extract the temporal signature relevant to classification, we avoid pre-processing steps such as down-sampling and band-pass filtering. Likewise, we do not perform a spectral decomposition. This also has the benefit of not adding a mode to the data representation.

IV. Empirical analyses

We compared the classification performance of logistic regression using features extracted by four existing supervised tensor methods (Discriminant Analysis with TEnsor Representation (DATER) [26], DATEReig [31], Constrained Multilinear Discriminant Analysis (CMDA) [29], and Direct General Tensor Discriminant Analysis (DGTDA) [29]) and the proposed manifold MDA approaces (ManTDA, ManPDA, ManTDA_sr, and ManPDA_sr). We compared the performances of these supervised approaches to logistic regression using features extracted by Tucker, Tucker2 [36], PARAFAC [32], [33], and PARAFAC2 [34], [35]). For comparison, we further included the Bilinear Discriminant Component Analysis (BDCA) [20] as well as our extension of BDCA to the Tucker representation (BDCA_Tucker) both of which combine feature extraction and classification by logistic regression.

1) Classification: For the Stekelenburg&Vroomen data, we used leave-one-subject-out cross-validation (CV) to estimate the between-subject performances of the models. Since there were 16 subjects, we had 16 CV folds. Each subject was left out in turn, and the models were trained on the remaining 15 subjects. To estimate the models' performances on unseen subjects, data from the left-out CV fold was then used as test data. In order to see how well each model fits the training data, we also inspected classification performances when the models were used to classify trials from the 15 CV folds that they were trained on.

For each of the two subjects from the BCI data, we performed 5-fold CV. Again, we inspected the models' performances both on training data (classifying trials form the four CV folds used for training) and on validation data (classifying the trials from the CV fold left out during training). We used the performances on the validation data to choose the best number of components for each model. Each model was then trained again using this number of components on all the CV folds. These models were then applied to the test data for which single-trial labels were not available. In a final step, these single-trial classifications were used to predict the letters spelled and these were compared to the correct letters.

All classification was performed within the logistic regression framework and the area under the Receiver Operating Curve (AUC) was used to quantify the classification performances. To calculate the AUC, the probabilities predicted by the logistic regression models were compared to the true single trial labels.

We extracted features in an unsupervised manner by decomposing data with the Tucker, Tucker2, PARAFAC, and PARAFAC2 models. For the Tucker, PARAFAC, and PARAFAC2 models, we used the estimated mode-3 (over which trials vary) factors as features in the input to logistic regression. For Tucker2, we used the derived projection matrices to project observations into the lower dimensional space. We then used all the scalar elements of the lower-dimensional representation as features in logistic regression. As for the Tucker2 model, we trained logistic regression models with the scalar elements of the low-dimensional representations of observations found by the MDA methods as classification features. The probabilities given by the BDCA methods were used directly when calculating the AUCs.

A. Component numbers

The supervised tensor classification methods find projection matrices that compress multi-way observations into lower-dimensional representations. With K components in each mode the dimensions of the lower dimensional space become $K \times K$ for matrix observations given by $\mathbf{U}^{(1)^\top}\mathbf{X}_n\mathbf{U}^{(2)}$, as for our data sets. Hence each observation leads to K^2 features in the lower-dimensional discriminative space. We investigated performances for one, three, and five components for the Tucker-structure projection methods (Tucker2, CMDA, DATER, DATEReig, DGTDA, ManTDA, ManTDA_sr,

[1] http://www.bbci.de/competition/iii/

and BDCA_Tucker). For the PARAFAC variants of the projection methods only the diagonal elements are used, i.e. $\mathrm{diag}(\mathbf{U}^{(1)^\top}\mathbf{X}_\mathrm{n}\mathbf{U}^{(2)})$. Hence, to get the same number of features as input to logistic regression for all methods, we also included 9 and 25 components for these methods. Likewise, we use the estimated trial strengths as features when classifying based on the unsupervised PARAFAC, PARAFAC2 and Tucker methods, giving only one feature per component for each trial. Hence we also estimated these models with 9 and 25 components.

B. Model implementations

We used the *nway* [73] toolbox to estimate the PARAFAC, PARAFAC2, Tucker, and Tucker2 models. These models were initialised with the best of 10 short runs, which were themselves initialised with random matrices. The BDCA methods were initialised with random normal values. The factors for the trial mode were constrained to be orthogonal for PARAFAC and PARAFAC2. For Tucker and Tucker2, all projection matrices were constrained to be orthogonal.

The existing MDA methods (DATER, DATEReig, CMDA, and DGTDA) were optimised by Matlab code that we wrote based on the pseudo-code in the papers describing these methods [26], [29]. CMDA and DATER were initialised with random orthogonal matrices while DGTDA does not need initialisation.

The BDCA methods were initialised with normally distributed random values. To avoid the log-likelihood from overflowing in the first iteration, the standard deviation of the initial random values for the Stekelenburg&Vroomen data was set to 0.01 while a lower value, 10^{-5}, was necessary to avoid overflow for the BCI data.

The proposed MDA methods based on manifold optimisation were optimised using the *ManOpt* [66] toolbox for Matlab. The models were initialised both with random orthonormal matrices and with projection matrices obtained from short runs of CMDA. However, the results from the two initialisation methods did not differ. For clarity of exposition, we only show the results from random initialisation.

For BDCA it was originally recommended to use the Damped Newton procedure in the *immoptibox* [74] to optimise the BDCA log-likelihood objective [20]. We optimised BDCA using both the suggested Damped Newton method and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) implementation, also available in the *immoptibox*. These two optimisation methods turned out to achieve very similar classification rates. The BFGS method was found to be slightly faster despite it only requiring gradients. We therefore used BFGS optimisation to optimise the BDCA_Tucker model and only show results from the BFGS optimisation.

All iterative methods were started three times and run for up to 5000 iterations or until convergence for the Stekelenburg&Vroomen data and for 1000 iterations for the BCI data. The best of the three solutions was chosen for further analysis in order to minimise the risk of analysing solutions from local minima. The convergence criteria used for CMDA, DATER and DATEReig were those originally proposed for CMDA and DATER [26], [29].

C. Visualisation

The projection matrices found by the supervised methods function as dimension-reducing filters that maximise the class-discriminative information in the filtered data. However, such filters are not suited for visualisation in order to interpret a model [75]. Instead, the interesting spatial properties of the estimated sources consist of how their activity is expressed on the scalp. This can be derived from the filters by pre-multiplying the data covariance matrix of electrodes onto the filter (projection) matrix if sources can be assumed to be uncorrelated. Similarly, we pre-multiplied the data covariance of temporal samples onto the temporal filter matrices to visualise the time courses of the sources. Since the MDA models with Tucker structure and BDCA are rotationally invariant, they do not have straight-forward interpretations, except in the one-component case.

On the other hand, each column in a projection matrix can only interact with one column from projection matrices for the other modes when using the PARAFAC structure. Also, we empirically observed that the PARAFAC formulations of MDA objectives were not invariant to rotations via random orthogonal matrices, making their interpretation more intuitive.For these reasons, we limit visualisations to one-component Tucker models and PARAFAC-structure MDA models when visualising more than one component.

V. Results

A. Objective function values

Figure 1 shows the objective function values obtained by CMDA, DATER, DATEReig, and our proposed manifold optimisation of the scatter-ratio objective with Tucker structure, which is also the objective function considered in the proof of the iterates of CMDA forming an asymptotically bounded sequence [29]. Objective function values for the trace of matrix ratio objective are also shown since this is optimised by DATEReig [31].CMDA, DATEReig, and the manifold methods share the same constraints on the projection matrices and are hence directly comparable. Each iteration for DATER, DATEReig, and CMDA corresponds to an update of the projection matrix for one of the modes.

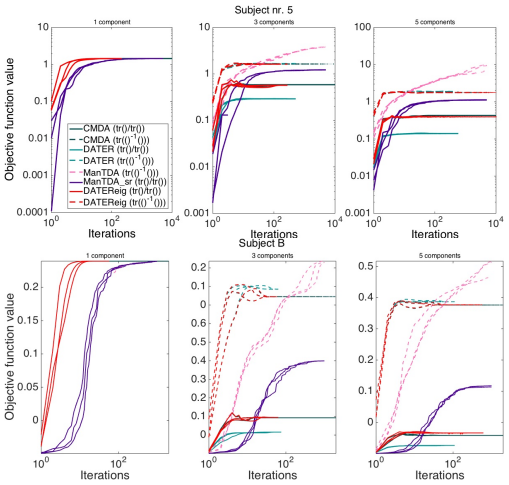


Figure 1. Objective function values for the methods which aim to optimise this. Scatter ratio objective function (2) values are shown as full lines while the matrix ratio objective (4) is shown as dashed lines. Top: Stekelenburg&Vroomen data for the CV fold with subject 5 left out. Bottom: subject B from the BCI data.

Each iteration for the manifold optimisation corresponds to one update in all modes since all modes are optimised at once in this approach.

The top of Figure 1 shows the optimisation for the CV fold with subject 5 left out in the Stekelenburg&Vroomen data. Although the evolution of objective function values over iterations differs to some extent between the leave-one-subject-out CV folds, they all look similar to the example shown here, which was chosen randomly. The bottom part of the figure shows the optimisation for CV fold number 1 for subject B. This is similar to the other CV folds, including those for subject A.

For one component, CMDA, DATEReig, and DATER obtain their optimum almost instantly, while the manifold methods are slower to reach this value. For the other component numbers, ManTDA and ManTDA_sr attain the highest objective function values, followed by CMDA and DATEReig. However, we see that the convergence of CMDA and DATEReig is not monotone as it increases rapidly to begin with, followed by a decline before stabilising. The alternation between optimising the two modes is also seen between each iteration in the beginning of the optimisation. Although more difficult to see, DATER also exhibits these characteristics, first reaching one objective function value, followed by a decline as well between each iteration. This shows that CMDA, DATEReig, and DATER do not optimise the scatter-ratio objective consistently when alternating between modes.

B. Cross-validated classification performance

In this section, we show the classification performance quantified by the area under receiver operating curve (AUC). Figure 2 shows the AUC when evaluating on training data for Stekelenburg&Vroomen data (top) and for the two BCI subjects (A in the middle and B at the bottom). When evaluating on training data, we expect all methods to improve when more components are added. This is indeed the case for all methods. On all the training data, ManTDA is the best-performing MDA method, followed by CMDA and DATER. The three other manifold approaches have slightly lower performances than CMDA, DATEReig, and DATER while the two BDCA methods outperform all the MDA methods. The two manifold methods with the PARAFAC structure (ManPDA and ManPDA_sr) do not show any substantial performance improvements for component numbers higher than one.

On the Stekelenburg&Vroomen training data, ManTDA, BDCA and BDCA_Tucker outperform the other methods, even obtaining perfect classification performances (AUC value of one) whereas the other MDA methods, except DGTDA, are very close to these best performances. On the Stekelenburg&Vroomen data, the PARAFAC-structure and Tucker-structure formulations of the objective functions have very similar performances but the PARAFAC-structure versions of MDA do not improve to perfection, as BDCA does for the largest component numbers. The performances are nearly identical, and low, for the unsupervised PARAFAC and Tucker models, even when allowed a large number of components. the Tucker2 method, which projects each trial into a lower dimensional space analogously to the MDA methods, performs substantially better than the other unsupervised methods, even outperforming DGTDA.

On the BCI training data, the two BDCA methods also outperform ManTDA. Here, the performance of BDCA is substantially higher than all other methods. With 25 components, BDCA again obtains AUC values of one, for both BCI subjects. On the BCI data, we observe some performance differences between ManPDA and ManTDA, with ManTDA performing best. For subject A, Tucker2 again outperforms DGTDA while it is on the same (low) level as PARAFAC and Tucker for subject B.

Figure 3 shows the classification performances obtained when evaluating on test data. Again, the results from the Stekelenburg&Vroomen data are shown in the top of the figure, with BCI subjects A and B in the middle and bottom, respectively. Again, ManPDA and ManPDA_sr do not show any substantial performance improvements for component numbers higher than one. On the test data, the dominance of the BDCA methods observed on the training data disappears.

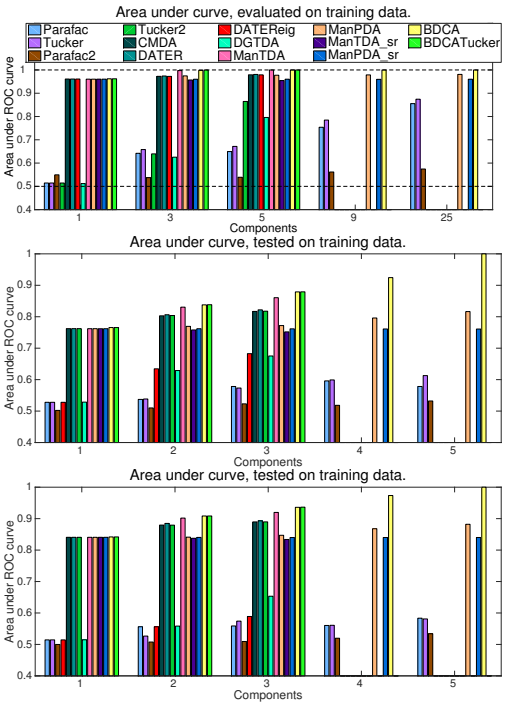


Figure 2. Testing on training data (data from each CV fold that was also used to train on). Top: Stekelenburg&Vroomen data. Middle: BCI data, subjet A. Bottom: BCI data, subjet B.

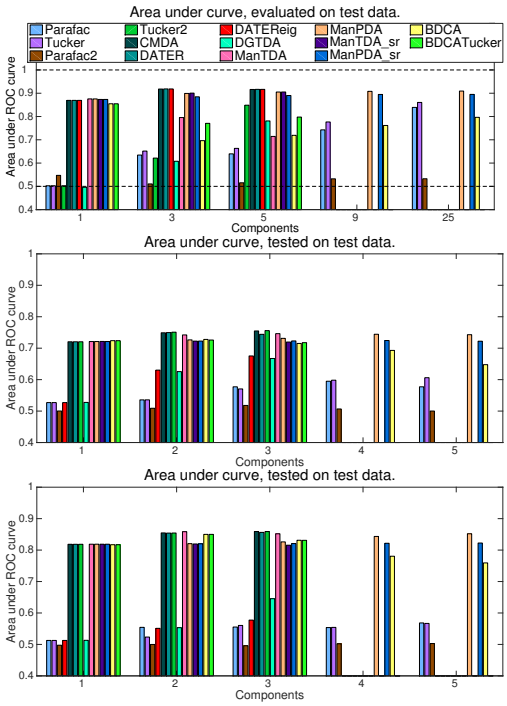


Figure 3. Testing on validation data (data left out from each CV fold). Top: Stekelenburg&Vroomen data. Middle: BCI data, subjet A. Bottom: BCI data, subjet B.

When evaluating on Stekelenburg&Vroomen test data, ManTDA and the BDCA methods perform worse than the other supervised methods, especially for high component numbers. With five components, they and DGTDA are even outperformed by Tucker2. The other MDA methods still obtain the highest performances, with Tucker, PARAFAC, and PARAFAC2 only obtaining low AUCs until 25 components. At this point, Tucker and PARAFAC approach the MDA performances, but are still outperformed by ManPDA and ManPDA_sr. ManTDA_sr and ManPDA_sr.

On the BCI data, ManTDA and the BDCA methods perform at the same level as all the MDA methods while Tucker and PARAFAC do not reach this level, with any component number. With four and five components (also with three for subject A), DGTDA is somewhat better than the unsupervised methods without coming close to the other supervised methods. While the performances of CMDA, DATER, and ManTDA are slightly better, all the MDA methods perform at similar levels.

C. BCI data letter classification performance

Table I shows average classification rates of letters across the two subjects in the BCI data. The first column gives the classification rates when each row/column was flashed 15 times to spell a character. The second column shows the results for 5 flashes. The average classification rates obtained by the five teams with highest performances in the competition are also shown, reproduced from the competition website[2]. We see that DATEReig obtains the best performance, closely followed by CMDA, DATER and ManTDA. We observe only small differences between PARAFAC and Tucker versions of otherwise same models.

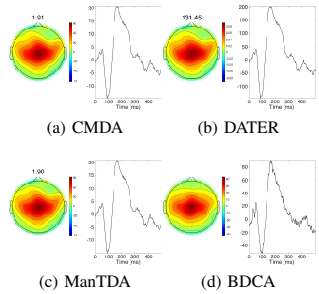
D. Model interpretation

We now show the temporal and spatial patterns of several of the fitted models. The components were derived and arranged in no particular order. Since the performances of the unsupervised methods are very low, we focus on visualisations of the supervised methods.

[2]http://www.bbci.de/competition/iii/results/index.html

	15 flashes	5 flashes
DATEReig	0.930	0.695
CMDA	0.925	0.695
DATER	0.925	0.670
ManTDA	0.915	0.660
ManPDA	0.910	0.645
BDCA	0.895	0.645
ManPDA_sr	0.890	0.555
BDCATucker	0.890	0.655
ManTDA_sr	0.880	0.555
Tucker2	0.605	0.260
DGTDA	0.595	0.320
Parafac	0.315	0.105
Tucker	0.300	0.090
Parafac2	0.025	0.005
Alain Rakotomamonjy	0.965	0.735
Li Yandong	0.905	0.550
Zhou Zongtan	0.900	0.595
Ulrich Hoffmann	0.895	0.530
Lin Zhonglin	0.875	0.575

Table I
MEAN LETTER CLASSIFICATION RATES FOR DATASET II FROM BCI COMPETITION III FROM THE COMPARED METHODS (TOP) AND BEST FIVE COMPETITION PARTICIPANTS (BOTTOM), COPIED FROM HTTP://WWW.BBCI.DE/COMPETITION/III/RESULTS/INDEX.HTML.



(a) CMDA (b) DATER

(c) ManTDA (d) BDCA

Figure 4. Spatial and temporal patterns corresponding to the extracted spatial and temporal filters found from the training data without subject 5 in the Stekelenburg&Vroomen data by the following (from top to bottom) one component models: CMDA, DATER, ManTDA, BDCA. Logistic regression coefficients are shown above the spatial patterns.

Figure 4 shows the scalp maps and corresponding temporal signatures extracted by one-component models trained on the same CV folds from the Stekelenburg&Vroomen data as in Figure 1. In the one-component case, the PARAFAC and Tucker versions of the same objective function are identical, making BDCA and BDCA_Tucker equivalent. Also, the trace of the matrix ratio is the same as the scatter ratio in this case, making all the methods optimised on manifolds equivalent. We included one-component models from each of the set of equivalent models in Figure 4.Except for different scaling in DATER, the components fitted by CMDA, DATER, and ManTDA are identical. This is reflected in the nearly identical logistic regression coefficients (shown above the spatial patterns) found for CMDA and ManTDA. The magnitude of the temporal pattern found by DATER is lower than that in CMDA and ManTDA, which is accounted for by the higher logistic regression coefficient. Since the BDCA model uses the projection into a lower dimensional space directly in the logistic regression model, no extra coefficient is included in this model. Although the spatial and temporal patterns found by BDCA are not identical to those found by the other methods, they are very similar.

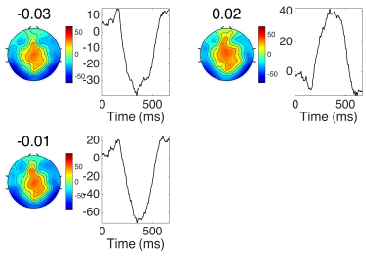
The temporal pattern of the component in the one-component model is very similar to the difference wave found by Stekelenburg and Vroomen between the two conditions that we classify (control and non-speech auditory) [70]. The centrally located scalp map is also in good accordance with their analysis of the central Cz electrode [70]. The logistic regression model was trained to predict probabilities for the auditory class. All shown components are well in line with this training since the positive logistic regression coefficients means that centrally located scalp activity with temporal activity like the difference wave in [70] indicates that a trial is from the auditory class.

Figure 5 shows the spatial and temporal patterns corresponding to the extracted filters for supervised MDA PARAFAC-structure models with three components, trained on four of the five CV folds from subject A's training data. All models extract a waveform similar to the P300 ERP, which is the theoretical foundation of P300 BCI systems. All the three components extracted by ManPDA look almost identical. The component on the right for ManPDA_sr has the same characteristics as the ManPDA components. The logistic regression model for the BCI data was trained to predict the probability of the target class, i.e. the class that should contain the P300 response. The estimated components described above and their logistic regression coefficients are in line with this since the estimates mean that central scalp activity exhibiting the P300-like waveform increase the probability of an observation being from the target class.

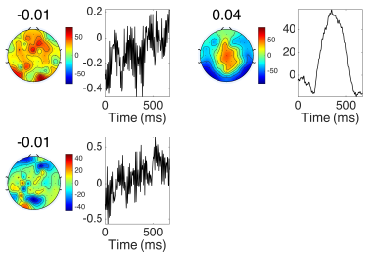
The two components shown on the left for ManPDA_sr are more difficult to interpret since their spatial patterns are not smooth and their temporal patterns are very high frequent.

VI. DISCUSSION

We saw that supervising the feature extraction step resulted in better classification rates. When feature extraction is not supervised, some directions of the data space that contain class-discriminative information but have low variance, and so explain only a small data proportion, may be lost since unsupervised feature extraction includes the data directions that best explain data variance. Even when including a large number of components, the unsupervised methods did not obtain



(a) ManPDA



(b) ManPDA_sr

Figure 5. Spatial and temporal patterns corresponding to the extracted spatial and temporal filters from the PARAFAC models (ManPDA, ManPDA_sr, BDCA from top to bottom) trained on four of five CV folds from subject A's BCI data. Fitted logistic regression coefficients are shown above the spatial patterns.

competitive classification performances, emphasising the need for supervised feature extraction methods.

Although the manifold optimisation approach obtained substantially higher objective function values than existing heuristic optimisation provides, we did not observe large classification performance differences between the supervised methods. With the same number of components, the Tucker and PARAFAC versions of the methods also performed similarly. These results indicate that it is important to use class information when finding the subspaces on which classification is performed, but the exact procedure is not as important. However, the PARAFAC-versions proposed for MDA are attractive due to their interpretability.

Combining feature extraction and learning the classifier in one step by BDCA led to the best performance on training data. However, as was also a problem for the ManTDA method, the performance dropped on Stekelenburg&Vroomen test data, especially with many components. This pattern is a sign of overfitting, both for ManTDA and the BDCA methods. On BCI data, the performance of ManTDA and BDCA did not drop on the test data as these data sets had substantially more trials. The PARAFAC structure constrains the model structure, enforcing natural regularisation. This trait of the PARAFAC structure can also be advantageous. As was originally recommended, regularising these methods would probably improve their performance [20]. Regularisation could be done in an unsupervised manner by using a Tucker2 compression of the temporal and spatial modes before applying the supervised methods. The regularisation originally recommended was a smoothing function [20], making the estimated spatial and temporal filters smoother. Alternatively, such a smoothing constraint could be applied to the patterns to make them resemble expressions of neural activity more. Other regularisation options are also possible. For example, L1 or L2 regularisation could be incorporated in the logistic regression model in the BDCA methods.

For our manifold optimisation, we used conjugate gradient as provided in the *ManOpt* toolbox [66]. However, more efficient optimisation using newer, more advanced manifold optimisation methods [76], [77] might be beneficial. In order to minimise the amount of pre-processing, we used the raw EEG trial data as input to the compared methods. In view of the lack of pre-processing, the high classification rates are surprising and indicates that the tensor methods are able to extract the temporal, as well as spatial, characteristics of data.Hence these methods might also be useful for extracting neural phenomena without prior knowledge.

VII. Conclusion

We set out to investigate whether the performance of Multi-linear Discriminant Analysis (MDA) methods could be improved through rigorous optimisation instead of existing optimisation heuristics. We found that rigorous optimisation does obtain higher objective function values than the existing procedures based on eigenvalue decomposition and the singular value decomposition. This, however, did not lead to better classification performance. Additionally, we wanted to inspect whether it is necessary to use supervised methods when searching for subspaces suitable for classification. Our results showed that supervised feature extraction methods perform substantially better than unsupervised methods. Finally, we also compared PARAFAC- and Tucker formulations of the otherwise same models. We did not observe large differences between these formulations. Hence we can conclude that it is necessary to use available observation labels when performing feature extraction, but the exact optimisation approach is not as important as long as it is supervised. For model interpretation we found that the proposed PARAFAC MDA models are attractive. Our Matlab implementations are available at http://www2.compute.dtu.dk/~lffr/publications/indexpub.php.

Acknowledgment

The authors would like to thank Jeroen J. Stekelenburg and Jean Vroomen for kindly letting us analyse their

data [70]. Morten Mørup was supported by the Lundbeck Foundation (grant nr. R105-9813).

REFERENCES

[1] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[2] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. Springer, 2011.

[3] D. Tao, X. Li, W. Hu, S. Maybank, and X. Wu, "Supervised tensor learning," in *Data Mining, Fifth IEEE International Conference on*. IEEE, 2005, pp. 8–pp.

[4] R. Ridgway, O. Irfanoglu, R. Machiraju, and K. Huang, "Image segmentation with tensor-based classification of n-point correlation functions," in *MICCAI Workshop on Medical Image Analysis with Applications in Biology*, vol. 1, 2006.

[5] S. Bourennane and C. Fossati, "About classification methods based on tensor modelling for hyperspectral images," in *Signal Processing, Image Processing and Pattern Recognition*. Springer, 2009, pp. 282–296.

[6] A. Smalter, J. Huan, and G. Lushington, "Feature selection in the tensor product feature space," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 1004–1009.

[7] S. Velasco-Forero and J. Angulo, "Classification of hyperspectral images by tensor modeling and additive morphological decomposition," *Pattern Recognition*, vol. 46, no. 2, pp. 566–577, 2013.

[8] B. Cao, L. He, X. Kong, P. S. Yu, Z. Hao, and A. B. Ragin, "Tensor-based multi-view feature selection with applications to brain diseases," in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 40–49.

[9] L. He, X. Kong, S. Y. Philip, A. B. Ragin, Z. Hao, and X. Yang, "Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages," *matrix*, vol. 3, no. 1, p. 2, 2014.

[10] X. Song, L. Meng, Q. Shi, and H. Lu, "Learning tensor-based features for whole-brain fmri classification," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer, 2015, pp. 613–620.

[11] T. Vo, D. Tran, and W. Ma, "Tensor decomposition and application in image classification with histogram of oriented gradients," *Neurocomputing*, 2015.

[12] E. Benetos and C. Kotropoulos, "A tensor-based approach for automatic music genre classification," in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–4.

[13] Y. Fu and T. S. Huang, "Image classification using correlation tensor analysis," *Image Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 226–234, 2008.

[14] E. Benetos and C. Kotropoulos, "Non-negative tensor factorization applied to music genre classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1955–1967, 2010.

[15] C. Durante, R. Bro, and M. Cocchi, "A classification tool for n-way array based on simca methodology," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 73–85, 2011.

[16] Y. Zhang, Z. Jiang, and L. S. Davis, "Discriminative tensor sparse coding for image classification," in *Proc. Brit. Mach. Vis. Conf*, 2013, pp. 83–1.

[17] F. Marini and R. Bro, "Scream: A novel method for multi-way regression problems with shifts and shape changes in one mode," *Chemometrics and Intelligent Laboratory Systems*, vol. 129, pp. 64–75, 2013.

[18] K. Wimalawarne, R. Tomioka, and M. Sugiyama, "Theoretical and experimental analyses of tensor-based regression and classification," *arXiv preprint arXiv:1509.01770*, 2015.

[19] M. Dyrholm and L. C. Parra, "Smooth bilinear classification of eeg," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 2006, pp. 4249–4252.

[20] M. Dyrholm, C. Christoforou, and L. C. Parra, "Bilinear discriminant component analysis," *The Journal of Machine Learning Research*, vol. 8, pp. 1097–1111, 2007.

[21] W. Guo, I. Kotsia, and I. Patras, "Tensor learning for regression," *Image Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 816–827, 2012.

[22] X. Li, H. Zhou, and L. Li, "Tucker tensor regression and neuroimaging analysis," *arXiv preprint arXiv:1304.5637*, 2013.

[23] W. Liu, J. Chan, J. Bailey, C. Leckie, F. Chen, and K. Ramamohanarao, "A bayesian classifier for learning from tensorial data," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 483–498.

[24] X. Tan, Y. Zhang, S. Tang, J. Shao, F. Wu, and Y. Zhuang, "Logistic tensor regression for classification," in *Intelligent Science and Intelligent Data Engineering*. Springer, 2013, pp. 573–581.

[25] M. Li and B. Yuan, "2d-lda: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527–532, 2005.

[26] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Discriminant analysis with tensor representation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 526–532.

[27] D. Tao, X. Li, X. Wu, and S. Maybank, "Tensor rank one discriminant analysis, a convergent method for discriminative multilinear subspace selection," *Neurocomputing*, vol. 71, no. 10, pp. 1866–1882, 2008.

[28] Y. Fu, J. Gao, X. Hong, and D. Tien, "Tensor regression based on linked multiway parameter analysis," in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 821–826.

[29] Q. Li and D. Schonfeld, "Multilinear discriminant analysis for higher-order tensor data classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 12, pp. 2524–2537, 2014.

[30] K. Huang and L. Zhang, "Cardiology knowledge free ecg feature extraction using generalized tensor rank one discriminant analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–15, 2014.

[31] M. Visani, C. Garcia, and J.-M. Jolion, "Normalized radial basis function networks and bilinear discriminant analysis for face recognition," in *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*. IEEE, 2005, pp. 342–347.

[32] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[33] R. A. Harshman, "Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, 1970.

[34] R. Harshman, "Parafac2: Extensions of a procedure for explanatory factor analysis and multidimensional scaling," *The Journal of the Acoustical Society of America*, vol. 51, no. 1A, pp. 111–111, 1972.

[35] H. A. Kiers, J. M. Ten Berge, and R. Bro, "Parafac2-part i. a direct fitting algorithm for the parafac2 model," *Journal of Chemometrics*, vol. 13, no. 3-4, pp. 275–294, 1999.

[36] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[37] S. J. Luck, *An Introduction to Event-Related Potentials and Their Neural Origins (Chapter 1)*. Cambridge: MIT Press, 2005.

[38] C. Guan, M. Thulasidas, and J. Wu, "High performance p300 speller for brain-computer interface," in *Biomedical Circuits and Systems, 2004 IEEE International Workshop on*. IEEE, 2004, pp. S3–5.

[39] D. Zhu, J. Bieger, G. G. Molina, and R. M. Aarts, "A survey of stimulation methods used in ssvep-based bcis," *Computational intelligence and neuroscience*, vol. 2010, p. 1, 2010.

[40] S. Tobimatsu, Y. M. Zhang, and M. Kato, "Steady-state vibration somatosensory evoked potentials: physiological characteristics and tuning function," *Clinical neurophysiology*, vol. 110, no. 11, pp. 1953–1958, 1999.

[41] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann, "A novel 9-class auditory erp paradigm driving a predictive text entry system," *Frontiers in neuroscience*, vol. 5, 2011.

[42] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao, "High-speed spelling with a noninvasive brain–computer interface," *Proceedings of the National Academy of Sciences*, vol. 112, no. 44, pp. E6058–E6067, 2015.

[43] E. Acar, C. Aykut-Bingol, H. Bingol, R. Bro, and B. Yener, "Multiway analysis of epilepsy tensors," *Bioinformatics*, vol. 23, no. 13, pp. i10–i18, 2007.

[44] M. De Vos, L. De Lathauwer, B. Vanrumste, S. Van Huffel, and W. Van Paesschen, "Canonical decomposition of ictal scalp eeg and accurate source localisation: principles and simulation study," *Computational intelligence and neuroscience*, vol. 2007, 2007.

[45] H. Lee, Y.-D. Kim, A. Cichocki, and S. Choi, "Nonnegative tensor factorization for continuous eeg classification," *International journal of neural systems*, vol. 17, no. 04, pp. 305–317, 2007.

[46] Z. Wang, A. Maier, N. K. Logothetis, and H. Liang, "Single-trial decoding of bistable perception based on sparse nonnegative tensor decomposition," *Computational intelligence and neuroscience*, vol. 2008, 2008.

[47] W. Deburchgraeve, P. Cherian, M. De Vos, R. Swarte, J. Blok, G. H. Visser, P. Govaert, and S. Van Huffel, "Neonatal seizure localization using parafac decomposition," *Clinical Neurophysiology*, vol. 120, no. 10, pp. 1787–1796, 2009.

[48] C. Nagendhiran, M. A. Kumar, S. Kharthigeyan, L. Naveen, and S. S. Prasanna, "Tensor scheme using gtda for eeg mental task classification," in *Proceedings of the 10th WSEAS international conference on Wavelet analysis and multirate systems*. World Scientific and Engineering Academy and Society (WSEAS), 2010, pp. 83–88.

[49] K. Vanderperren, M. De Vos, B. Mijović, J. Ramautar, N. Novitskiy, B. Vanrumste, P. Stiers, B. Van den Bergh, J. Wagemans, L. Lagae *et al.*, "Parafac on erp data from a visual detection task during simultaneous fmri acquisition," in *Proc. of the International Biosignal Processing Conference., Berlin, Germany*, vol. 103, 2010, pp. 1–4.

[50] A. H. Phan, A. Cichocki, and T. Vu-Dinh, "A tensorial approach to single trial recognition for brain computer interface," in *Advanced Technologies for Communications (ATC), 2010 International Conference on*. IEEE, 2010, pp. 138–141.

[51] F. Cong, A. H. Phan, Q. Zhao, A. K. Nandi, V. Alluri, P. Toiviainen, H. Poikonen, M. Huotilainen, A. Cichocki, and T. Ristaniemi, "Analysis of ongoing eeg elicited by natural music stimuli using nonnegative tensor factorization," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 494–498.

[52] T. D. Nguyen, T. Tran, D. Phung, and S. Venkatesh, "Tensor-variate restricted boltzmann machines," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[53] J. Li and L. Zhang, "Regularized tensor discriminant analysis for single trial eeg classification in bci," *Pattern Recognition Letters*, vol. 31, no. 7, pp. 619–628, 2010.

[54] A. Onishi, A. H. Phan, K. Matsuoka, and A. Cichocki, "Tensor classification for p300-based brain computer interface," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 581–584.

[55] Y. Liu, M. Li, H. Zhang, J. Li, J. Jia, Y. Wu, J. Cao, and L. Zhang, "Single-trial discrimination of eeg signals for stroke patients: a general multi-way analysis," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 2204–2207.

[56] B. Hunyadi, M. Signoretto, S. Debener, S. Van Huffel, and M. De Vos, "Classification of structured eeg tensors using nuclear norm regularization: improving p300 classification," in *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*. IEEE, 2013, pp. 98–101.

[57] A. Eliseyev and T. Aksenova, "Recursive n-way partial least squares for brain-computer interface," *PloS one*, vol. 8, no. 7, p. e69962, 2013.

[58] X. Li, C. Guan, H. Zhang, K. K. Ang, and S. H. Ong, "Adaptation of motor imagery eeg classification model based on tensor decomposition," *Journal of neural engineering*, vol. 11, no. 5, p. 056020, 2014.

[59] L. Billiet, B. Hunyadi, V. Matic, S. V. Huffel, M. Verleysen, and M. D. Vos, "Single trial classification for mobile bci - a multiway kernel approach." in *BIOSIGNALS*, H. Loose, A. L. N. Fred, H. Gamboa, and D. Elias, Eds. SciTePress, 2015, pp. 5–11.

[60] Q. Zhao, C. F. Caiafa, A. Cichocki, L. Zhang, and A. H. Phan, "Slice oriented tensor decomposition of eeg data for feature extraction in space, frequency and time domains," in *Neural Information Processing*. Springer, 2009, pp. 221–228.

[61] H. Ji, J. Li, R. Lu, R. Gu, L. Cao, and X. Gong, "Eeg classification for hybrid brain-computer interface using a tensor based multi-class multimodal analysis scheme," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.

[62] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[63] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 24–40, 2011.

[64] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, vol. 18, no. 2, pp. 95–138, 1977.

[65] M. Petschow, E. Peise, and P. Bientinesi, "High-performance solvers for dense hermitian eigenproblems," *SIAM Journal on Scientific Computing*, vol. 35, no. 1, pp. C1–C22, 2013.

[66] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a matlab toolbox for optimization on manifolds," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1455–1459, 2014.

[67] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[68] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, p. 15, 2008.

[69] S. Liu and G. Trenkler, "Hadamard, khatri-rao, kronecker and other matrix products," *Int. J. Inform. Syst. Sci*, vol. 4, no. 1, pp. 160–177, 2008.

[70] J. J. Stekelenburg and J. Vroomen, "Neural correlates of multisensory integration of ecologically valid audiovisual events," *Journal of Cognitive Neuroscience*, vol. 19, no. 12, pp. 1964–1973, 2007.

[71] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1034–1043, 2004.

[72] B. Blankertz, K.-R. Müller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlögl, G. Pfurtscheller, J. R. Millan, M. Schröder, and N. Birbaumer, "The bci competition iii: Validating alternative approaches to actual bci problems," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 14, no. 2, pp. 153–159, 2006.

[73] C. A. Andersson and R. Bro, "The n-way toolbox for matlab," *Chemom.Intell.Lab.Syst.*, vol. 1, no. 52, pp. 1–4, 2000.

[74] H. Nielsen, "immoptibox–a matlab toolbox for optimization," 2006.

[75] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96 – 110, 2014.

[76] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1-2, pp. 397–434, 2013.

[77] B. Jiang and Y.-H. Dai, "A framework of constraint preserving update schemes for optimization on stiefel manifold," *Mathematical Programming*, vol. 153, no. 2, pp. 535–575, 2015.

APPENDIX A
STATIONARY POINTS

We now show that the stationary points of the trace of matrix ratio objective and ratio of deteriminants objectives are the same. We do this by finding the derivatives of both objective functions, setting them equal to zero, and solving for the projection matrix $\mathbf{U}_{LDA}$.

We find the derivative of the trace of matrix ratios as:

$$\begin{aligned}&\frac{\partial Tr((\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})}{\partial \mathbf{U}_{LDA}} = \\ &-2\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &\times \mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &+2\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}.\end{aligned}$$

Setting this equal to zero, we find:

$$\begin{aligned}0=&-2\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &+2\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ \Leftrightarrow&\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &\times \mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA} \\ =&\mathbf{B}_{LDA}\mathbf{U}_{LDA}.\end{aligned}$$

Set $g_w(\mathbf{U}_{LDA}) = \det(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})$ and $g_b(\mathbf{U}_{LDA}) = \det(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})$. From Equation (53), p. 9 in [1], we have:

$$\begin{aligned}\frac{\partial g_w(\mathbf{U}_{LDA})}{\partial \mathbf{U}_{LDA}} =& 2g_w(\mathbf{U}_{LDA})\mathbf{W}_{LDA}\mathbf{U}_{LDA} \\ &\times(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}.\end{aligned}$$

For the ratio matrix determinants objective, we then find:

$$\begin{aligned}&\frac{\partial \frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})}}{\partial \mathbf{U}_{LDA}} = \\ &2g_b(\mathbf{U}_{LDA})\mathbf{B}_{LDA}\mathbf{U}_{LDA} \\ &\times(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1}\frac{1}{g_w(\mathbf{U}_{LDA})} \\ &-\frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})^2}2g_w(\mathbf{U}_{LDA})\mathbf{W}_{LDA}\mathbf{U}_{LDA} \\ &\times(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &= 2\frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})}(\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &-\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}).\end{aligned}$$

We set this equal to zero and solving for $\mathbf{U}_{LDA}$:

$$\begin{aligned}0=&2\frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})}(\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &-\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}) \\ \Leftrightarrow&\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ =&\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1} \\ \Leftrightarrow&\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &\times(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA}) \\ =&\mathbf{B}_{LDA}\mathbf{U}_{LDA}.\end{aligned}$$

This is the same equality as before, thus the stationary points are equivalent.

REFERENCES

[1] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," nov 2012, version 20121115.
[2] M. Petschow, E. Peise, and P. Bientinesi, "High-performance solvers for dense hermitian eigenproblems," *SIAM Journal on Scientific Computing*, vol. 35, no. 1, pp. C1–C22, 2013.

Part VII

Appendices

Appendix F

Mode multiplication examples

To concretise the model (3.1), we show a small example where $\mathcal{X}$ and $\mathcal{G}$ have dimensions $2 \times 2 \times 2$ and $U^{(1)}$, $U^{(2)}$, and $U^{(3)}$ have dimensions 2×2. From the example in Figure 3.1 we already have $\mathcal{G}^{(1)}_{(1)} = \left(\mathcal{G} \times_1 U^{(1)}\right)_{(1)}$. Unmatricising this, we get

$$\mathcal{G}^{(1)}_{:,:,1} = \begin{pmatrix} u^{(1)}_{11} g_{111} + u^{(1)}_{12} g_{211} & u^{(1)}_{11} g_{121} + u^{(1)}_{12} g_{221} \\ u^{(1)}_{21} g_{111} + u^{(1)}_{22} g_{211} & u^{(1)}_{21} g_{121} + u^{(1)}_{22} g_{221} \end{pmatrix}$$

$$\mathcal{G}^{(1)}_{:,:,2} = \begin{pmatrix} u^{(1)}_{11} g_{112} + u^{(1)}_{12} g_{212} & u^{(1)}_{11} g_{122} + u^{(1)}_{12} g_{222} \\ u^{(1)}_{21} g_{112} + u^{(1)}_{22} g_{212} & u^{(1)}_{21} g_{122} + u^{(1)}_{22} g_{222} \end{pmatrix}.$$

The mode-2 multiplication with $U^{(2)}$ results in $\mathcal{G} \times_2 U^{(2)} = \mathcal{G}^{(2)}$ with

$$\mathcal{G}^{(2)}_{:,:,1} = \begin{pmatrix} u^{(2)}_{11} g_{111} + u^{(2)}_{12} g_{121} & u^{(2)}_{21} g_{111} + u^{(2)}_{22} g_{121} \\ u^{(2)}_{11} g_{211} + u^{(2)}_{12} g_{221} & u^{(2)}_{21} g_{211} + u^{(2)}_{22} g_{221} \end{pmatrix}$$

$$\mathcal{G}^{(2)}_{:,:,2} = \begin{pmatrix} u^{(2)}_{11} g_{112} + u^{(2)}_{12} g_{122} & u^{(2)}_{21} g_{112} + u^{(2)}_{22} g_{122} \\ u^{(2)}_{11} g_{212} + u^{(2)}_{12} g_{222} & u^{(2)}_{21} g_{212} + u^{(2)}_{22} g_{222} \end{pmatrix}.$$

The mode-3 multiplication with $U^{(3)}$ results in $\mathcal{G}^{(3)}$ with

$$\mathcal{G}^{(3)}_{:,:,1} = \begin{pmatrix} u^{(3)}_{11} g_{111} + u^{(3)}_{12} g_{112} & u^{(3)}_{11} g_{121} + u^{(3)}_{12} g_{122} \\ u^{(3)}_{11} g_{211} + u^{(3)}_{12} g_{212} & u^{(3)}_{11} g_{221} + u^{(3)}_{12} g_{222} \end{pmatrix}$$

$$\mathcal{G}^{(3)}_{:,:,2} = \begin{pmatrix} u^{(3)}_{21} g_{111} + u^{(3)}_{22} g_{112} & u^{(3)}_{21} g_{121} + u^{(3)}_{22} g_{122} \\ u^{(3)}_{21} g_{211} + u^{(3)}_{22} g_{212} & u^{(3)}_{21} g_{221} + u^{(3)}_{22} g_{222} \end{pmatrix}.$$

Appendix G

Derivatives

Using derivatives of the LDA objectives and the Kronecker and Khatri-Rao producs, the corresponding MDA objectives' derivatives can be found using the chain rule. Here, we give the derivatives of some LDA objective functions and the Khatri-Rao and Kronecker products' derivatives. Additionally, we write up the derivative of the Tucker formulation of the BDCA method. We limit ourselves to the two-matrix case.

For the Kronecker and Khatri-Rao products, we also limit ourselves to giving the derivatives with respect to the first matrix in the products. By changing the order of modes, these derivatives can be applied with respect to any mode. If the derivative with respect to the p^{th} mode projection matrix ($\mathbf{U}^{(p)}$) is desired, the order of projection matrices and modes can be changed by switching places of $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(p)}$ and correspondingly changing the order of modes in observations X (e.g. using *permute* in Matlab (*permute(X, [p, 2, ..., p-1, 1, p+1, ..., P])*)). For the Kronecker and Khatri-Rao products, the cases with matrices can also easily be extended to more matrices by setting the second projection matrix equal to the product of all the projection matrices that are constant with respect to the one for which the derivative is taken.

G.1 General derivative rules

We will need the following identities in the following:

$$\begin{aligned}\frac{\partial Tr(\mathbf{U}^\top \mathbf{A}\mathbf{U})}{\partial \mathbf{U}} &= \mathbf{A}\mathbf{U} + \mathbf{A}^\top \mathbf{U} \\ &= (\mathbf{A} + \mathbf{A}^\top)\mathbf{U} \text{ Equation (108), p. 13 in Petersen and Pedersen [2012].}\end{aligned} \tag{G.1}$$

For symmetric A and C:

$$\begin{aligned}\frac{\partial Tr((\mathbf{U}^\top \mathbf{C}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{A}\mathbf{U})}{\partial \mathbf{U}} &= -2\mathbf{C}\mathbf{U}(\mathbf{U}^\top \mathbf{C}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{A}\mathbf{U}(\mathbf{U}^\top \mathbf{C}\mathbf{U})^{-1} \\ &+ 2\mathbf{A}\mathbf{U}(\mathbf{U}^\top \mathbf{C}\mathbf{U})^{-1} \text{ Equation (126), p. 14 in Petersen and Pedersen [2012].}\end{aligned} \tag{G.2}$$

We also need the quotient rule which holds since the product and chain rules hold (Equations (37) and (38), p. 8 Petersen and Pedersen [2012]), from which the quotient rule can be derived. The quotient rule states that, if $h(X) \neq 0$, then

$$\frac{\partial \frac{g(\mathbf{X})}{h(\mathbf{X})}}{\partial \mathbf{X}_{ij}} = \frac{\frac{g(\mathbf{X})}{\partial \mathbf{X}_{ij}} h(\mathbf{X}) - g(\mathbf{X}) \frac{h(\mathbf{X})}{\partial \mathbf{X}_{ij}}}{(h(\mathbf{X}))^2} = \frac{g(\mathbf{X})}{\partial \mathbf{X}_{ij}} \frac{1}{h(\mathbf{X})} - \frac{g(\mathbf{X})}{h(\mathbf{X})^2} \frac{h(\mathbf{X})}{\partial \mathbf{X}_{ij}}. \tag{G.3}$$

Kronecker The derivative of the Kronecker product is Fackler [2005], Vetter [1973]:

$$\frac{\partial \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)}}{\partial \mathbf{U}^{(1)}} = (I_{K_1,K_2} \otimes T_{J_1,p})(I_{K_1} \otimes vec(\mathbf{U}^{(2)}2) \otimes I_{J_1}),$$

where I_{J_1} is the $J_1 \times J_1$ identity matrix, I_{K_1,K_2} is the identity matrix restricted to K_1 rows and K_2 columns and T_{J_1,J_2} is the permutation matrix such that $T_{J_1,J_2} vec(\mathbf{X}) = vec(\mathbf{X}^\top)$ and the matrices $\mathbf{U}^{(p)}$ have dimensions $J_p \times K_p$ for $p = 1, 2$.

Khatri-Rao We give this as the vectorised version Kolda [2006]:

$$\frac{\partial vec(\mathbf{U}^{(1)} \odot \mathbf{U}^{(2)})}{\partial vec(\mathbf{U}^{(1)})} = \mathbf{U}^{(2)} \otimes \mathbf{I},$$

is the $J_1 \times J_1$ identity matrix and $\mathbf{U}^{(1)}$ has dimensions $J_1 \times K_1$.

G.2 Derivatives of LDA objectives

We get the derivative of the trace of matrix ratios directly from (G.2):

$$\begin{aligned}\frac{\partial Tr((\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA} \mathbf{U}_{LDA})^{-1} \mathbf{U}_{LDA}^\top \mathbf{B}_{LDA} \mathbf{U}_{LDA})}{\partial \mathbf{U}_{LDA}} &= -2\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &\times \mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &+ 2\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}.\end{aligned}$$

Setting this equal to zero, we find:

$$\begin{aligned}0 &= -2\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &+ 2\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\ &\Leftrightarrow \mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA} = \mathbf{B}_{LDA}\mathbf{U}_{LDA}.\end{aligned}$$

Using that the matrices $\mathbf{W}_{LDA}$ and $\mathbf{B}_{LDA}$ are real and symmetric, and assume that $\mathbf{W}_{LDA}$ is positive-definite, a solution to the Symmetric Generalised Eigenvalue Problem $\mathbf{B}_{LDA}\mathbf{Q} = \mathbf{W}_{LDA}\mathbf{Q}\mathbf{\Lambda}$ exists such that $\mathbf{Q}^\top \mathbf{B}_{LDA}\mathbf{Q} = \mathbf{\Lambda}$ and $\mathbf{Q}^\top \mathbf{W}_{LDA}\mathbf{Q} = \mathbf{I}$ [Petschow et al., 2013], where the columns of $\mathbf{Q}$ contain generalised eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix with the generalised eigenvectors contained in the diagonal. These generalised eigenvectors solve the above equality, as seen by substituting $\mathbf{U}_{LDA}$ with $\mathbf{Q}$:

$$\begin{aligned}\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA} &= \mathbf{B}_{LDA}\mathbf{U}_{LDA} \Leftrightarrow \\ \mathbf{W}_{LDA}\mathbf{Q}(\mathbf{Q}^\top \mathbf{W}_{LDA}\mathbf{Q})^{-1}\mathbf{Q}^\top \mathbf{B}_{LDA}\mathbf{Q} &= \mathbf{B}_{LDA}\mathbf{Q} \Leftrightarrow \\ \mathbf{W}_{LDA}\mathbf{Q}\mathbf{I}^{-1}\mathbf{\Lambda} &= \mathbf{B}_{LDA}\mathbf{Q} \Leftrightarrow \\ \mathbf{W}_{LDA}\mathbf{Q}\mathbf{\Lambda} &= \mathbf{B}_{LDA}\mathbf{Q},\end{aligned}$$

where the above equality is true by the definition of $\mathbf{Q}$ and $\mathbf{\Lambda}$.

Set $g_w(\mathbf{U}_{LDA}) = \det(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})$ and $g_b(\mathbf{U}_{LDA}) = \det(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})$. From Equation (53), p. 9 in Petersen and Pedersen [2012], we have:

$$\frac{\partial g_w(\mathbf{U}_{LDA})}{\partial \mathbf{U}_{LDA}} = 2g_w(\mathbf{U}_{LDA})\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}.$$

Then the derivative of the Fisherface criterion, the ratio of determinants of the within- and between-class scatters, in Equation (3.4) is

$$\begin{aligned}
\frac{\partial \frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})}}{\partial \mathbf{U}_{LDA}} \overset{\text{(G.3)}}{=} & \; 2g_b(\mathbf{U}_{LDA})\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1}\frac{1}{g_w(\mathbf{U}_{LDA})} \\
- & \; \frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})^2} 2g_w(\mathbf{U}_{LDA})\mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} \\
= & \; 2\frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})}(\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1} \\
- & \; \mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}).
\end{aligned}$$

Setting this equal to zero and solving for $\mathbf{U}_{LDA}$, we get

$$\begin{aligned}
0 = & \; 2\frac{g_b(\mathbf{U}_{LDA})}{g_w(\mathbf{U}_{LDA})}(\mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1} \\
- & \; \mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}) \\
\Leftrightarrow & \; \mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1} = \mathbf{B}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA})^{-1} \\
\Leftrightarrow & \; \mathbf{W}_{LDA}\mathbf{U}_{LDA}(\mathbf{U}_{LDA}^\top \mathbf{W}_{LDA}\mathbf{U}_{LDA})^{-1}(\mathbf{U}_{LDA}^\top \mathbf{B}_{LDA}\mathbf{U}_{LDA}) = \mathbf{B}_{LDA}\mathbf{U}_{LDA}.
\end{aligned}$$

This is the same equality as before, hence this is also solved as the Generalised Eigenvalue Problem $\mathbf{B}_{LDA}\mathbf{U}_{LDA} = \mathbf{\Lambda}\mathbf{W}_{LDA}\mathbf{U}_{LDA}$.

To calculate the derivative of the scatter ratio objective, we use the quotient rule (G.3) and (G.1):

$$\frac{\partial \frac{tr(\mathbf{U}^\top \mathbf{B}_{LDA}\mathbf{U})}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})}}{\partial \mathbf{U}} = 2\mathbf{B}_{LDA}\mathbf{U}\frac{1}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})} - \frac{tr(\mathbf{U}^\top \mathbf{B}_{LDA}\mathbf{U})}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})^2} \times 2\mathbf{W}_{LDA}\mathbf{U}.$$

Setting this equal to zero, we find the optimum as:

$$\begin{aligned}
0 = & \; 2\mathbf{B}_{LDA}\mathbf{U}\frac{1}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})} - \frac{tr(\mathbf{U}^\top \mathbf{B}_{LDA}\mathbf{U})}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})^2} \times 2\mathbf{W}_{LDA}\mathbf{U} \\
\Leftrightarrow & \; \frac{tr(\mathbf{U}^\top \mathbf{B}_{LDA}\mathbf{U})}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})^2}\mathbf{W}_{LDA}\mathbf{U} = \mathbf{B}_{LDA}\mathbf{U}\frac{1}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})} \\
\Leftrightarrow & \; \frac{tr(\mathbf{U}^\top \mathbf{B}_{LDA}\mathbf{U})}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})}\mathbf{W}_{LDA}\mathbf{U} = \mathbf{B}_{LDA}\mathbf{U}.
\end{aligned}$$

Due to the term $\mathbf{U}$, $\frac{tr(\mathbf{U}^\top \mathbf{B}_{LDA}\mathbf{U})}{tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})}$, this cannot be solved as a Generalised Eigenvalue Problem.

For the scatter difference objective, we get the following derivative, using (G.1) and that the matrices $\mathbf{W}_{LDA}$ and $\mathbf{B}_{LDA}$ are symmetric:

$$\frac{\partial tr(\mathbf{U}^\top \mathbf{B}_{LDA}\mathbf{U}) - \zeta tr(\mathbf{U}^\top \mathbf{W}_{LDA}\mathbf{U})}{\partial \mathbf{U}} = 2\mathbf{B}_{LDA}\mathbf{U} - 2\zeta \mathbf{W}_{LDA}\mathbf{U}.$$

Setting the above derivative equal to zero gives:

$$\begin{aligned} 0 &= 2\mathbf{B}_{LDA}\mathbf{U} - 2\zeta \mathbf{W}_{LDA}\mathbf{U} \\ &\Leftrightarrow \zeta \mathbf{W}_{LDA}\mathbf{U} = \mathbf{B}_{LDA}\mathbf{U}, \end{aligned}$$

which is a Generalised Eigenvalue Problem if ζ is a constant.

G.3 BDCA_Tucker derivatives

The BDCA_Tucker model is the following formulation of the logistic regression log-likelihood:

$$f_{BDCA\_Tucker}(w_0, \mathbf{V}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{n=1}^{N} \left[y_n(w_0 + \psi(\mathbf{X}_n)) - \log(1 + \exp(w_0 + \psi(\mathbf{X}_n))) \right],$$

where $\psi(\mathbf{X}_n) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} [\mathbf{U}^{(1)\top} \mathbf{X}_n \mathbf{U}^{(2)}]_{k_1,k_2} \mathbf{V}_{k_1,k_2}$. We set $\mathbf{V}_{k_1,k_2} = 1$ for $k_1 = k_2$ to remove scaling ambiguities between the projection matrices and the matrix of interaction coefficients, $\mathbf{V}$.

Following the definitions in Dyrholm et al. [2007], we define:

$$\pi(\mathbf{X}_n) = \mathbb{E}(y_n = 1) = \frac{1}{1 + \exp(-(w_0 + \psi(\mathbf{X}_n)))}, \tag{G.4}$$

where $y_n \in \{0, 1\}$ is the class of the n^{th} trial. The gradient of the log-likelihood is then given by:

$$\frac{\partial f_{BDCA\_Tucker}(w_0, \mathbf{V}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)})}{\partial w_0} = \sum_{n=1}^{N}(y_n - \pi(\mathbf{X}_n)),$$

$$\frac{\partial f_{BDCA\_Tucker}(w_0, \mathbf{V}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)})}{\partial \mathbf{V}_{k_1,k_2}} = \sum_{n=1}^{N}\left[y_n[\mathbf{U}^{(1)\top}\mathbf{X}_n\mathbf{U}^{(2)}]_{k_1,k_2} - [\mathbf{U}^{(1)\top}\mathbf{X}_n\mathbf{U}^{(2)}]_{k_1,k_2}\pi(\mathbf{X}_n)\right],$$

$$\frac{\partial f_{BDCA\_Tucker}(w_0, \mathbf{V}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)})}{\partial \mathbf{U}^{(1)}_{:,k_1}} = \sum_{n=1}^{N}\left[(y_n - \pi(\mathbf{X}_n))\sum_{k_2=1}^{K_2}[\mathbf{X}_n\mathbf{U}^{(2)}]_{:,k_2}\mathbf{V}_{k_1,k_2}\right],$$

and

$$\frac{\partial f_{BDCA\_Tucker}(w_0, \mathbf{V}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)})}{\partial \mathbf{U}^{(2)\top}_{:,k_2}} = \sum_{n=1}^{N}\left[(y_n - \pi(\mathbf{X}_n))\sum_{k_1=1}^{K_1}[\mathbf{U}^{(1)\top}\mathbf{X}_n]_{k_1,:}\mathbf{V}_{k_1,k_2}\right].$$

Appendix H

CSP filter regularisation matrices based on artefactual ICs

Before deciding on the final focus in the paper "Investigating effects of different artefact types on Motor Imagery BCI", we made some preliminary studies of various regularisation approaches for spatial filters in the CSP objective function. None of these improved significantly on the CSP performance compared to regularisation with the identity matrix. We briefly describe the considered regularisation approaches here.

H.1 Normalisation of scalp maps

There are two obvious ways to determine the normalisation of the scalp maps used for regularisation, and each of these imply different relative weights for the directions of the ICs regularised against.

H.2 Normalisation options

The scalp maps used to regularise could be normalised either by requiring their norm to be one or by scaling them such that their corresponding time series would have norm (or standard deviation) one. Since the data was band-pass filtered each EEG channel had mean zero. This meant that each IC time series also had mean zero since an IC is a linear combination of channels. Hence the norm of an IC time series is equivalent to its standard deviation.

H.3 Implications of normalisation for regularisation

Requiring that all scalp maps have norm one implies that all artefacts are weighted equally in the regularisation if no other weighting scheme of ICs is employed. Conversely, the standardisation of time series to have norm (or standard deviation) one, implies a weighting of ICs according to how active they were during the calibration session from which the ICA decomposition was calculated. Then the directions of more active ICs would be exposed to stronger regularisation. It is not, however, certain that this is desirable since it is only class-discriminative artefact directions that we want to avoid in the spatial filters found by CSP. Artefacts that are very active during the calibration session are likely to be present to a similar extent in both classes, and will thus automatically be ignored in the CSP spatial filters. On the other hand, it is likely that artefacts that are only active infrequently overall are more present in one class compared to the other class. Regularisation is used to help CSP spatial filters avoid such directions of artifactual ICs. Hence it is not clear that the overall activity of an IC, as quantified by the norm or standard deviation of its time series, should be a useful normalisation factor.

We chose to normalise scalp maps to have norm one.

H.3.1 Weighted regularisation against artifactual ICs

By using the probabilities of class membership to weight the regularisation of individual ICs, ICs with more artifactual characteristics should be exposed to stronger regularisation than ICs for which the probability of being neural is higher. Let $\boldsymbol{q_{art}}$ be the vector of probabilities that each IC considered artifactual belongs to one of the classes defined as artifactual. Let $diag(\boldsymbol{q_{art}})$ be a diagonal

matrix with the elements in the diagonal equal to the elements of $\boldsymbol{q_{art}}$. Again, let $\mathbf{A}_{art}$ be a matrix containing the normalised scalp maps of ICs considered artifactual in its columns. The penalty matrix is then set equal to $(\mathbf{A}_{art} \times diag(\boldsymbol{q_{art}})) \times (\mathbf{A}_{art} \times diag(\boldsymbol{q_{art}}))^{\top}$.

H.3.2 Weighted regularisation against all ICs

This scheme is similar to the above, but all ICs are regularised against, weighted by their probability of being artifactual. Hence even ICs that are most likely from the neural class are regularised against, but with a smaller weight. In this case, the penalty matrix is $(\mathbf{A} \times diag(\boldsymbol{q})) \times (\mathbf{A} \times diag(\boldsymbol{q}))^{\top}$, where the matrix $\mathbf{A}$ contains the normalised scalp maps of all ICs and the vector $\boldsymbol{q}$ contains the probabilities of being artifactual for all ICs.

H.4 Subject-independent regularisation schemes

In the following regularisation schemes, the penalty matrix for a subject is calculated without using the calibration data of that subject. In some of the following regularisation schemes, data from other subjects is used to calculate the penalty matrix. All these regularisation schemes focus on moving away from the space of muscular ICs since the first experiments on the subject-dependent level showed that muscle artefacts are the most problematic artefact class for CSP on motor-imagery data.

H.4.1 Distances of electrodes from centre

Figurw H.1 shows the magnitude of penalty on each electrode when using squared Gaussian bell distances from the scalp centre, defined by $\left(1 - \exp(-(x^2 + y^2)/r)\right)^2$, where r is the radius of the scalp and x and y are the 2-D coordinates of an electrode. By setting the diagonal elements of a diagonal matrix equal to these distances, a subject-independent regularisation scheme was obtained.

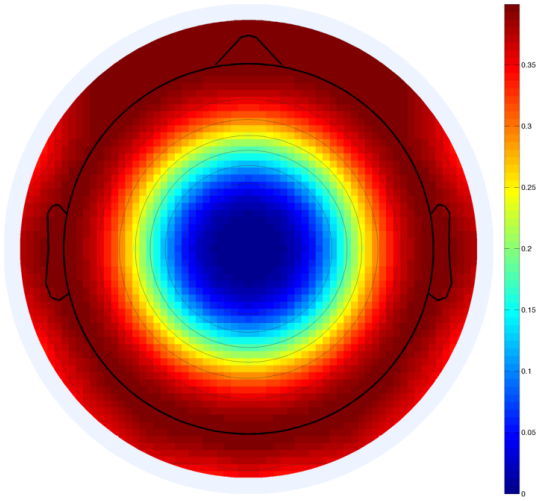


Figure H.1: Penalty of electrodes when penalised by squared Gaussian distances from the scalp centre.

H.4.2 Average electrode weights in muscle ICs of other subjects

Figure H.2 shows the magnitude of penalty on each electrode for subject one when penalising each electrode by its average weight in muscle ICs times the probability of that IC being muscular for all other subjects with the same EEG channels. The electrode penalties obtained in this manner were used as the diagonal elements in a diagonal matrix to regularise CSP. Formally, the weights were obtained as shown in Algorithm 2.

H.4.3 Concatenated muscle ICs from other subjects

We also attempted to learn directions of muscular artefacts for one subject from other subjects by constructing the penalty matrix by concatenating column vectors of muscular IC patterns from other subjects with the same EEG channel configuration as the one subject.

Algorithm 2 Find electrode penalties using electrodes' average weights in muscle IC patterns and return the penalty matrix V.

$muscle\_ics \leftarrow ()$
for $isubj =$ subjects with same channels ▷ each other subject with the same EEG channels **do**
  for $iic = nmuscle\_ics$ ▷ each muscle IC of subject $isubj$ **do**
    $current\_pattern \leftarrow A_{isubj}(:, iic)$
    $muscle\_ics \leftarrow (muscle\_ics,\ current\_pattern)$ ▷ concatenate column vectors of patterns of muscle ICs from subject $isubj$
  end for
end for
$\boldsymbol{w} \leftarrow rowMeans(muscle\_ics)$ ▷ take the row means of the concatened columns of muscle IC patterns to get the average weights of electrodes in muscle IC patterns.
$V \leftarrow diag(\boldsymbol{w})$ **return** V

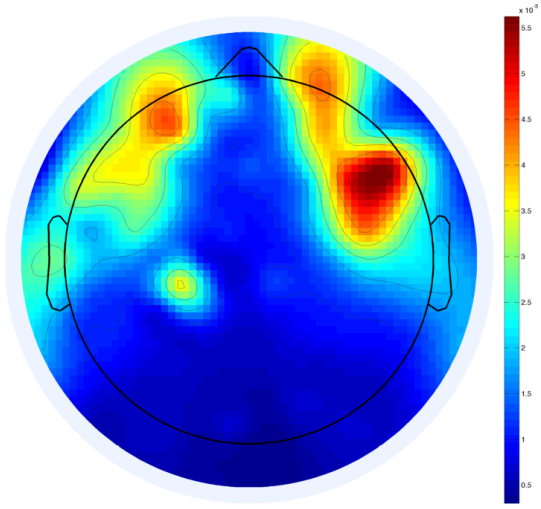


Figure H.2: Average electrode weights in muscle ICs of all but subject one.

Bibliography

Carsten Allefeld, Peter Beim Graben, and Jürgen Kurths. Advanced methods of electrophysiological signal analysis and symbol grounding?: dynamical systems approaches to language. Nova Publishers, 2008.

C. A. Andersson and R. Bro. The n-way toolbox for matlab. Chemom.Intell.Lab.Syst., 1(52):1–4, 2000.

Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19(7):711–720, 1997.

Christopher M Bishop. Pattern recognition and machine learning. springer, 2006.

Benjamin Blankertz, Klaus-Robert Müller, Dean J Krusienski, Gerwin Schalk, Jonathan R Wolpaw, Alois Schlögl, Gert Pfurtscheller, Jd R Millan, Michael Schröder, and Niels Birbaumer. The bci competition iii: Validating alternative approaches to actual bci problems. Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 14(2):153–159, 2006.

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

Benjamin Blankertz, Claudia Sannelli, Sebastian Halder, Eva M Hammer, Andrea Kübler, Klaus-Robert Müller, Gabriel Curio, and Thorsten Dickhaus. Neurophysiological predictor of smr-based bci performance. Neuroimage, 51 (4):1303–1309, 2010.

Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. The Journal of Machine Learning Research, 15(1):1455–1459, 2014.

Stephanie Brandl, Laura Frølich, Johannes Höhne, Klaus-Robert Müller, and Wojciech Samek. Brain-computer interfacing under distraction: An evaluation study. Submitted, 2016.

Rasmus Bro, Claus A Andersson, and Henk AL Kiers. Parafac2-part ii. modeling chromatographic data with retention time shifts. Journal of Chemometrics, 13(3-4):295–309, 1999.

György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents–eeg, ecog, lfp and spikes. Nature reviews neuroscience, 13(6):407–420, 2012.

J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. Psychometrika, 35(3):283–319, 1970.

Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. High-speed spelling with a noninvasive brain–computer interface. Proceedings of the National Academy of Sciences, 112 (44):E6058–E6067, 2015.

Wei Cheng, Seungchul Lee, Zhousuo Zhang, and Zhengjia He. Independent component analysis based source number estimation and its comparison for mechanical systems. Journal of Sound and Vibration, 331(23):5153 – 5167, 2012. ISSN 0022-460X. doi: http://dx.doi.org/10.1016/j.jsv.2012.06.021.

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons, 2009.

Pierre Comon. Independent component analysis, a new concept? Signal Process., 36(3):287–314, April 1994. ISSN 0165-1684. doi: 10.1016/0165-1684(94)90029-9.

Rodney J Croft and Robert J Barry. Issues relating to the subtraction phase in eog artefact correction of the eeg. International Journal of Psychophysiology, 44(3):187–195, 2002.

Fernando Lopes da Silva. Eeg: origin and measurement. In EEG-fMRI, pages 19–38. Springer, 2009.

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

Arnaud Delorme, Jason Palmer, Julie Onton, Robert Oostenveld, and Scott Makeig. Independent eeg sources are dipolar. PloS one, 7(2):e30135, 2012.

Jeng-Ren Duann, Tzyy-Ping Jung, Scott Makeig, and Terrence J Sejnowski. Repeated decompositions reveal the stability of infomax decomposition of fmri data. In Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, pages 5324–5327. IEEE, 2006.

Mads Dyrholm, Christoforos Christoforou, and Lucas C Parra. Bilinear discriminant component analysis. The Journal of Machine Learning Research, 8:1097–1111, 2007.

Kasper Eskelund, Ewen N MacDonald, and Tobias S Andersen. Face configuration affects speech perception: Evidence from a mcgurk mismatch negativity study. Neuropsychologia, 66:48–54, 2015.

Paul L Fackler. Notes on matrix calculus. North Carolina State University, 2005.

Oliver Faugeras, Joël Janin, Frédéric Cazals, and Pierre Kornprobst. Modeling in Computational Biology and Biomedicine: A Multidisciplinary Endeavor. Springer Science & Business Media, 2012.

Bruce J Fisch and Rainer Spehlmann. Fisch and Spehlmann's EEG primer: basic principles of digital and analog EEG. Elsevier Health Sciences, 1999.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2):179–188, 1936.

Laura Frølich and Irene Winkler. Removal of muscular artifacts in eeg signals: A comparison of ica and other linear decomposition methods. Submitted, 2016.

Laura Frølich, T. S. Andersen, and Morten Mørup. Classification of independent components of eeg into multiple artifact classes. Psychophysiology, 52(1):32–45, Jan 2015a.

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

Laura Frølich, Tobias S. Andersen, and Morten Mørup. Multi-way strategies for single-trial classification of electroencephalography data. Submitted, 2016.

G Geetha and SN Geethalakshmi. Noise cancellation of ocular and muscular artifacts from eeg signals based on adaptive filtering. International Journal of Computer and Electrical Engineering, 4(5):785, 2012.

Germán Gómez-Herrero, Mercedes Atienza, Karen Egiazarian, and Jose L Cantero. Measuring directional coupling between eeg sources. Neuroimage, 43(3): 497–508, 2008.

Klaus Gramann, Thomas Töllner, and Hermann J Müller. Dimension-based attention modulates early visual processing. Psychophysiology, 47(5):968–978, 2010.

Cuntai Guan, Manoj Thulasidas, and Jiankang Wu. High performance p300 speller for brain-computer interface. In Biomedical Circuits and Systems, 2004 IEEE International Workshop on, pages S3–5. IEEE, 2004.

Harold W Gutch and Fabian J Theis. Independent subspace analysis is unique, given irreducibility. In Independent Component Analysis and Signal Separation, pages 49–56. Springer, 2007.

Richard Harshman. Parafac2: Extensions of a procedure for explanatory factor analysis and multidimensional scaling. The Journal of the Acoustical Society of America, 51(1A):111–111, 1972. doi: http://dx.doi.org/10.1121/1.1981298.

Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis. UCLA Working Papers in Phonetics, 1970.

Trevor J.. Hastie, Robert John Tibshirani, and Jerome H Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage, 87:96 – 110, 2014. ISSN 1053-8119.

Sven Hoffmann and Michael Falkenstein. The correction of eye blink artefacts in the eeg: a comparison of two prominent methods. PLoS One, 3(8):e3004, 2008.

Johannes Höhne, Martijn Schreuder, Benjamin Blankertz, and Michael Tangermann. A novel 9-class auditory erp paradigm driving a predictive text entry system. Frontiers in neuroscience, 5, 2011.

Kenneth Hugdahl and René Westerhausen. The two halves of the brain: Information processing in the cerebral hemispheres. MIT Press, 2010.

Aapo Hyvärinen. Independent component analysis in the presence of gaussian noise by maximizing joint likelihood. Neurocomputing, 22(1):49–67, 1998.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. Neural Networks, IEEE Transactions on, 10(3):626–634, 1999.

Aapo Hyvärinen. Independent component analysis: recent advances. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 371(1984), 2012. ISSN 1364-503X. doi: 10.1098/rsta.2011.0534.

Aapo Hyvärinen and Pavan Ramkumar. Testing independent component patterns by inter-subject or inter-session consistency. Frontiers in Human Neuroscience, 7(94), 2013. ISSN 1662-5161. doi: 10.3389/fnhum.2013.00094.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, editors. Independent Component Analysis. Wiley, 2001. missing.

Aapo Hyvärinen, Pavan Ramkumar, Lauri Parkkonen, and Riitta Hari. Independent component analysis of short-time fourier transforms for spontaneous eeg/meg analysis. NeuroImage, 49(1):257–271, 2010.

BW Jervis, M Coelho, and GW Morgan. Effect on eeg responses of removing ocular artefacts by proportional eog subtraction. Medical and Biological Engineering and Computing, 27(5):484–490, 1989.

Joe-Air Jiang, Chih-Feng Chao, Ming-Jang Chiu, Ren-Guey Lee, Chwan-Lu Tseng, and Robert Lin. An automatic analysis method for detecting and eliminating ecg artifacts in eeg. Computers in biology and medicine, 37(11): 1660–1671, 2007.

Murray Johns, Kate Crowley, Robert Chapman, Andrew Tucker, and Christopher Hocking. The effect of blinks and saccadic eye movements on visual reaction times. Attention, Perception, & Psychophysics, 71(4):783–788, 2009.

Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J Mckeown, Vicente Iragui, and Terrence J Sejnowski. Removing electroencephalographic artifacts by blind source separation. Psychophysiology, 37(02):163–178, 2000.

Motoaki Kawanabe and Carmen Vidaurre. Improving bci performance by modified common spatial patterns with robustly averaged covariance matrices. In World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany, pages 279–282. Springer, 2009.

Motoaki Kawanabe, Wojciech Samek, Klaus-Robert Müller, and Carmen Vidaurre. Robust common spatial filters with a maxmin approach. Neural computation, 26(2):349–376, 2014.

Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro. Parafac2-part i. a direct fitting algorithm for the parafac2 model. Journal of Chemometrics, 13(3-4): 275–294, 1999.

Albert Kim and Lee Osterhout. The independence of combinatory semantic processing: Evidence from event-related potentials. Journal of Memory and Language, 52(2):205–225, 2005.

Timo Kirschstein and Rüdiger Köhling. What is the source of the eeg? Clinical EEG and neuroscience, 40(3):146–149, 2009.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. SIAM review, 51(3):455–500, 2009.

Tamara Gibson Kolda. Multilinear operators for higher-order decompositions. United States. Department of Energy, 2006.

Gundars Korats, Steven Le Cam, Radu Ranta, and Mohamed Hamid. Applying ica in eeg: Choice of the window length and of the decorrelation method. In Biomedical Engineering Systems and Technologies, pages 269–286. Springer, 2012.

Te-Won Lee. Independent Component Analysis: Theory and Applications. Springer US, 1999. ISBN 978-1-4419-5056-7. doi: 10.1007/978-1-4757-2851-4.

Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural computation, 11(2):417–441, 1999.

Ming Li and Baozong Yuan. 2d-lda: A statistical linear discriminant analysis for image matrix. Pattern Recognition Letters, 26(5):527–532, 2005.

Qun Li and D. Schonfeld. Multilinear discriminant analysis for higher-order tensor data classification. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 36(12):2524–2537, Dec 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2014.2342214.

Xinhai Liu, Wolfgang Glänzel, and Bart De Moor. Hybrid clustering of multi-view data via tucker-2 model and its application. Scientometrics, 88(3):819–839, 2011.

Fabien Lotte and Cuntai Guan. Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms. Biomedical Engineering, IEEE Transactions on, 58(2):355–362, 2011.

S. J. Luck. An Introduction to Event-Related Potentials and Their Neural Origins (Chapter 1). MIT Press, Cambridge, 2005.

S Makeig, M Westerfield, T-P Jung, S Enghoff, J Townsend, E Courchesne, and TJ Sejnowski. Dynamic brain sources of visual evoked responses. Science, 295(5555):690–694, 2002.

Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, Terrence J Sejnowski, et al. Independent component analysis of electroencephalographic data. Advances in neural information processing systems, pages 145–151, 1996.

Andrea Mognon, Jorge Jovicich, Lorenzo Bruzzone, and Marco Buiatti. Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. Psychophysiology, 48(2):229–240, 2011.

Morten Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1):24–40, 2011.

Suresh Daniel Muthukumaraswamy. High-frequency brain activity and muscle artifacts in meg/eeg: a review and recommendations. Frontiers in human neuroscience, 7, 2013.

Christa Neuper and Wolfgang Klimesch. Event-related dynamics of brain oscillations, volume 159. Elsevier, 2006.

Christa Neuper and Gert Pfurtscheller. Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. International journal of psychophysiology, 43(1):41–58, 2001.

Vadim V Nikulin, Guido Nolte, and Gabriel Curio. A novel method for reliable and fast extraction of neuronal eeg/meg oscillations on the basis of spatio-spectral decomposition. NeuroImage, 55(4):1528–1535, 2011.

Julie Onton and Scott Makeig. High-frequency broadband modulations of electroencephalographic spectra. Frontiers in human neuroscience, 3, 2009.

Jason A Palmer, Scott Makeig, Kenneth Kreutz-Delgado, and Bhaskar D Rao. Newton method for the ica mixture model. In ICASSP, pages 1805–1808, 2008.

Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda. Recipes for the linear analysis of eeg. Neuroimage, 28(2):326–341, 2005.

Christopher A Paynter, Kenneth Kotovsky, and Lynne M Reder. Problem-solving without awareness: an erp investigation. Neuropsychologia, 48(10): 3137–3144, 2010.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. Version 20121115.

Matthias Petschow, Elmar Peise, and Paolo Bientinesi. High-performance solvers for dense hermitian eigenproblems. SIAM Journal on Scientific Computing, 35(1):C1–C22, 2013.

Michael Plöchl, José P Ossandón, and Peter König. Combining eeg and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. Front. Hum. Neurosci, 6(278):10–3389, 2012.

P. M. Quilter, B. B. MacGillivray, and D. G. Wadbrook. The removal of eye movement artefact from EEG signals using correlation techniques. In IEEE Conference Random Signal Analysis, volume 159, pages 93–100. IEEE Conference Publication, 1977.

T Radüntz, J Scouten, O Hochmuth, and B Meffert. Eeg artifact elimination by extraction of ica-component features using image processing algorithms. Journal of neuroscience methods, 243:84–93, 2015.

I Rejer and P Gorski. Benefits of ica in the case of a few channel eeg. In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pages 7434–7437. IEEE, 2015.

WH Ridder and A Tomlinson. Suppression of contrast sensitivity during eyelid blinks. Vision research, 33(13):1795–1802, 1993.

O Ryynänen, J Hyttinen, and J Malmivuo. Effect of skull resistivity and measurement noise on the spatial resolution of eeg. International Journal of Bioelectromagnetism, 7(1), 2005.

Wojciech Samek and Klaus-Robert Muller. Information geometry meets bci spatial filtering using divergences. In Brain-Computer Interface (BCI), 2014 International Winter Workshop on, pages 1–4. IEEE, 2014.

Claudia Sannelli, Thorsten Dickhaus, Sebastian Halder, Eva-Maria Hammer, Klaus-Robert Müller, and Benjamin Blankertz. On optimal channel configurations for smr-based brain–computer interfaces. Brain topography, 23(2): 186–193, 2010.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. Biomedical Engineering, IEEE Transactions on, 51(6):1034–1043, 2004.

Donald L Schomer and Fernando Lopes Da Silva. Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins, 2012.

Alwin Stegeman. Degeneracy in candecomp/parafac and indscal explained for several three-sliced arrays with a two-valued typical rank. Psychometrika, 72 (4):601–619, 2007.

Jeroen J Stekelenburg and Jean Vroomen. Neural correlates of multisensory integration of ecologically valid audiovisual events. Journal of Cognitive Neuroscience, 19(12):1964–1973, 2007.

Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(10):1700–1715, 2007.

Shozo Tobimatsu, You Min Zhang, and Motohiro Kato. Steady-state vibration somatosensory evoked potentials: physiological characteristics and tuning function. Clinical neurophysiology, 110(11):1953–1958, 1999.

Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3):279–311, 1966.

Fay S Tyner, John Russell Knott, and W Brem Mayer. Fundamentals of EEG technology: Clinical correlates, volume 2. Lippincott Williams & Wilkins, 1989.

Rolf Verleger, Claudia Paulick, Joachim Möcks, Janette L Smith, and Karsten Keller. Parafac and go/no-go: Disentangling cnv return from the p3 complex by trilinear component analysis. International Journal of Psychophysiology, 87(3):289–300, 2013.

William J Vetter. Matrix calculus operations and taylor expansions. SIAM review, 15(2):352–369, 1973.

François-Benoit Vialatte, Jordi Solé-Casals, Monique Maurice, Charles Latchoumane, Nigel Hudson, Sunil Wimalaratna, Jaeseung Jeong, and Andrzej Cichocki. Improving the quality of eeg data in patients with alzheimer's disease using ica. In Advances in Neuro-Information Processing, pages 979–986. Springer, 2008.

Ricardo Vigário, Veikko Jousmäki, M Hämäläninen, R Haft, and Erkki Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. Advances in neural information processing systems, pages 229–235, 1998.

Filipa Campos Viola, Jeremy Thorne, Barrie Edmonds, Till Schneider, Tom Eichele, and Stefan Debener. Semi-automatic identification of independent components representing eeg artifact. Clinical Neurophysiology, 120(5):868–877, 2009.

Muriel Visani, Christophe Garcia, and J-M Jolion. Normalized radial basis function networks and bilinear discriminant analysis for face recognition. In Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on, pages 342–347. IEEE, 2005.

James R Voss, Luis Rademacher, and Mikhail Belkin. Fast algorithms for gaussian noise invariant independent component analysis. In Advances in Neural Information Processing Systems, pages 2544–2552, 2013.

Irene Winkler, Stefan Haufe, and Michael Tangermann. Automatic classification of artifactual ica-components for artifact removal in eeg signals. Behavioral and Brain Functions, 7(1):1, 2011.

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. In IEEE Engineering in Medicine and Biology Society (EMBC), pages 4101–4105, 2015.

S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang. Discriminant analysis with tensor representation. In IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2005.

Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 26(1):131–137, 2004.

Zheng Ye, Yue-jia Luo, Angela D Friederici, and Xiaolin Zhou. Semantic and syntactic processing in chinese sentence comprehension: Evidence from event-related potentials. Brain research, 1071(1):186–196, 2006.

Florian Yger, Fabien Lotte, and Masashi Sugiyama. Averaging covariance matrices for eeg signal classification based on the csp: An empirical study. In Signal Processing Conference (EUSIPCO), 2015 23rd European, pages 2721–2725. IEEE, 2015.

Danhua Zhu, Jordi Bieger, Gary Garcia Molina, and Ronald M Aarts. A survey of stimulation methods used in ssvep-based bcis. Computational intelligence and neuroscience, 2010:1, 2010.

Andreas Ziehe and Klaus-Robert Müller. Tdsep, an an efficient algorithm for blind separation using time structure. In Lars Niklasson, Mikael Boden, and Tom Ziemke, editors, ICANN 98, Perspectives in Neural Computing, pages 675–680. Springer London, 1998. ISBN 978-3-540-76263-8. doi: 10.1007/978-1-4471-1599-1_103.