Multivariate Machine Learning Methods for Fusing Multimodal Functional Neuroimaging Data

Sven Dähne, Felix Bießmann, Wojciech Samek, *Member, IEEE*, Stefan Haufe, Dominique Goltz, Christopher Gundlach, Arno Villringer, Siamac Fazli, Klaus-Robert Müller, *Member, IEEE,*

***Abstract*—Multimodal data is ubiquitous in engineering, communication, robotics, vision or more generally speaking in industry and the sciences. All disciplines have developed their respective sets of analytic tools to fuse the information that is available in all measured modalities. In this paper we provide a review of classical as well as recent machine learning methods (specifically factor models) for fusing information from functional neuroimaging techniques such as LFP, EEG, MEG, fNIRS and fMRI. Early and late fusion scenarios are distinguished and appropriate factor models for the respective scenarios are presented along with example applications from selected multimodal neuroimaging studies. Further emphasis is given to the interpretability of the resulting model parameters, in particular by highlighting how factor models relate to physical models needed for source localization. The methods we discuss allow to extract information from neural data, which ultimately contributes to (a) better neuroscientific understanding, (b) enhance diagnostic performance and (c) discover neural signals of interest that correlate maximally with a given cognitive paradigm. While we clearly study the multimodal functional neuroimaging challenge, the discussed machine learning techniques have a wide applicability beyond, i.e. in general data fusion and may thus be informative to the general interested reader.**

S. Dähne is with the Machine Learning Group, Department of Computer Science, Berlin Institute of Technology, Berlin, Germany, correspondence to (e-mail: sven.daehne@tu-berlin.de)

F Bießmann is with Amazon, Berlin, Germany.

W. Samek is with the Machine Learning Group, Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Berlin, Germany.

S. Haufe is with the Laboratory for Intelligent Imaging and Neural Computing, Columbia University, New York, NY, USA as well as with the Machine Learning Group, Department of Computer Science, Berlin Institute of Technology, Berlin, Germany.

D. Goltz and C. Gundlach are with the Institute of Psychology, University of Leipzig, Leipzig, Germany as well as the Department for Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany.

A. Villringer is with the Department for Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, the Mind-Brain Institute and Berlin School of Mind and Brain, Charite Universitätsmedizin Berlin and Humboldt-University, Berlin, Germany and the Clinic for Cognitive Neurology, University of Leipzig, Germany.

S. Fazli is with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea, correspondence to e-mail: fazli@korea.ac.kr

K.-R. Müller is with the Machine Learning Group, Department of Computer Science, Berlin Institute of Technology, Berlin, Germany, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Republic of Korea, correspondence to (e-mail: klaus-robert.mueller@tu-berlin.de)

This work was supported by the Brain Korea 21 Plus Program, the SGER Grant 2014055911 through the National Research Foundation of Korea funded by the Ministry of Education, as well as by a Marie Curie International Outgoing Fellowship (grant No. 625991) within the 7th European Community Framework Programme. This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein. This work was also supported in part by BMBF (01IS14013A-E and 01GQ1115).

I. Introduction

Modern neuroscience benefits greatly from a multitude of imaging techniques that, individually, have helped to further our understanding of cognitive processing [1], [2] and improved clinical diagnostics [3], [4]. The combination of several imaging modalities originated in the context of epilepsy imaging [5], [6], [7] but has since then become an important asset in cognitive neuroscience. It is only through multimodal setups that otherwise unparalleled spatial and temporal imaging resolution can be obtained, which allows for combination of complementary information and thereby a better diagnosis and a deeper understanding of how different aspects of brain activity are related.

The most popular multimodal imaging setups combine measurements of electrophysiology with measurements of hemodynamics. Example techniques for measuring electrophysiological properties of neural activity are electrocorticography (ECoG), electroencephalography (EEG), and magnetoencephalography (MEG). Examples for techniques that measure changes in hemodynamic parameters include positron emission tomography (PET), functional near-infrared spectroscopy (fNIRS), and functional magnetic resonance imaging (fMRI). See [8], [9] for reviews on electrophysiological aspects of brain activity and [10], [11] for reviews on hemodynamic aspects.

The task of optimal combination of information from several (imaging) modalities is referred to as *multimodal analysis* or *multimodal fusion*. Multimodal fusion represents an ongoing research endeavor, as there is still no gold standard solution [12], [13].

The following list enumerates some of the key challenges that make multimodal fusion a difficult problem:

- **Different spatial and temporal sampling rates:** The number of recording channels typically range from approximately one hundred for electrophysiology to near one million voxels for hemodynamics. The picture is reversed, however, for temporal sampling rates where electrophysiology is typically sampled in the kHz range while hemodynamics are sampled with rates below 10 Hz.
- **Non-instantaneous and non-linear coupling:** The vascular reaction to a given stimulus is in the range of seconds, while the response in electrophysiological measures (e.g. event-related potentials (ERPs)) occur in the range of milliseconds. Furthermore, non-linear features such as the instantaneous amplitude of neural oscillations may be related to linear features of hemodynamics.

- **The presence of outliers and low signal-to-noise ratio (SNR):** Signals of interest may not be easily detectable at the level of individual measurement channels due to a low SNR. The existence of outliers in the data, (either caused by technological or physiological artifacts) may further shadow the signals of interest and lead estimates of certain statistics of the data astray and thereby hinder successful fusion.
- **Interpretable results:** The aim of multimodal imaging settings is to increase our understanding of the workings of the brain. Therefore, the results of multimodal fusion techniques must be interpretable with respect to functional or anatomical neurophysiological references.

In order to overcome these challenges it is helpful to regard multimodal fusion as modeling as well as an optimization problem. With respect to both of these two views one class of statistical learning methods has become particularly popular for multimodal data analysis: *factor models*. These models assume that the measurements are the result of the activity of a limited set of *components* (see Section III for the formal definition) of which a mixture is observed at the level of the sensors of the measurement device. Un-mixing these components, requires a set of assumptions about the nature of the components. Different assumptions lead to different statistical learning methods and therefore it is important to know these assumptions when choosing an analysis method.

A decomposition into components can be done separately for each modality or jointly for all measurement modalities. We refer to the former approach as *late fusion* scenarios and to the latter approach as *early fusion*. In order to provide a comprehensible review of suitable fusion methods we refrain from covering the entire spectrum of multimodal analysis. Instead focus on methods for these two approaches to the analysis of multimodal functional neuroimaging data. See Fig. 1 for an illustration.

While in principle all of the presented models can be extended to more than two modalities, we here focus – for the sake of readability – on the special case of two different measurement modalities. Furthermore, since the scope of this review is limited to *functional* neuroimaging data, we assume all measurements to be temporally aligned.

The remainder of this manuscript is organized as follows. In section II we briefly summarize what is known about the physiological origins of electrophysiological and hemodynamic signals. In section III we revisit a generative model of multimodal data that expresses the recordings from each modality in terms of a set of hidden variables, which are called components. We then review classical as well as recent methods for the extraction of components from either each modality separately (section IV) or jointly from both modalities (section V). Extensions to these methods are reviewed in section VI. We conclude with a discussion in section VII.

II. Physiological origins

Before discussing analysis methods we very briefly review the physiological origins of electrophysiological and hemodynamic signals. Readers who are familiar with the basics of these types of signals and their coupling may skip ahead to the next section.

A. Physiological origins of electrophysiological measures

Neural activity results in changes of electrical fields [9], which can be measured at various spatial, temporal and functional extents [14]. Intracellular recordings allow for measuring action potentials of single neurons [15]. The activity of single and multiple neurons up to large neuronal assemblies can be extracted with extracellular recording techniques, either invasively with microelectrodes inserted in the brain or EcOG [16] or non-invasively with EEG [17] or MEG [18].

Extracellularly measured local field potentials (LFP) represent a superposition of all currents in the brain, with a distance-dependent contribution of different sources such as synaptic currents, calcium-spikes, action potentials and spike afterpotentials of different neurons [9]. While signals measured from microelectrodes and EcOGs can represent rather focal and localized signatures of neuronal activity, signals measured with EEG rely on synchronous activity of large assemblies of neurons. Such synchronous activity is often resembled in neuronal oscillations [19] and the spatial synchronization strength is reflected in the power of these oscillations [20], [21]. These oscillations have been linked to practically every aspect of cognitive function [22], [23], [24], [25], [19] and are thus also the subject of multimodal analysis settings. Besides these neuronal oscillations, there is synchronized activity of neurons measurable with electrophysiological methods that follows certain events or the presentation of stimuli. Such activity is often termed as event related potentials (ERPs), with different components attributed to various cognitive processes [26], [27].

B. Physiological origins of hemodynamic measures

Hemodynamic activity can be measured invasively by intrinsic optical imaging [28]. Non-invasive alternatives exist in the form of fMRI [29], [30] or fNIRS [31].

fMRI measures the combination of metabolic and vascular response to neural activation, the so-called blood oxygen-level dependent (BOLD) signal [30]. The BOLD signal is inversely related to the local concentration of deoxygenated hemoglobin (HbR), which in turn is influenced by changes in cerebral blood volume (CBV) and cerebral blood flow (CBF) [32]. Since HbR is paramagnetic, while oxygenated hemoglobin (HbO) is not, only changes in the concentration of HbR alter the local magnetic susceptibility and hence give rise to the fMRI signal obtained in a magnetic-resonance (MR) scanner with so-called T2*-weighted pulse sequences.

Functional near infrared spectroscopy (fNIRS) relies on the fact that near-infrared light can traverse biological tissue and thus allows the transmission of photons through the intact brain [31], [33]. The absorption properties of HbR and HbO differ substantially in the infrared range [34]. This enables to measure changes in concentrations of HbR and HbO *in vivo*. When compared to fMRI, fNIRS measurements can be performed with a lightweight and comparatively low-cost setup. Similar to EEG, light emitting and detection devices (so-called *optodes*) can be mounted on a fNIRS cap.

C. Neurovascular coupling

The relationship between neural activity and the vascular response is known as *neurovascular coupling* [10], [32], [35], [36] and the exact nature of this coupling is far from understood [37], [38], [39]. However, recently a number of studies have shown that neural and hemodynamic signals are highly correlated [40], [41], [42], [43], [44], [36], [45], [46]. Simultaneous intracranial electrophysiological recordings and high-resolution fMRI in macaque monkeys, for example, revealed a correlation between the BOLD signal and neuronal activity in the gamma range as a neurovascular coupling mechanism [43], [44]. Similar results have been obtained in cats [47]. However, neurovascular coupling can also be assessed using noninvasive methods such as combined EEG-fNIRS [48], [49] or EEG-fMRI [50], [51], [52], [53]. These and other studies have demonstrated an inverse relationship between the amplitude of neural oscillations in the alpha and beta range as well as a peak in correlation at a time delay of 6 to 8 seconds.

III. The multimodal linear model

A. Nomenclature

In this paper we represent the modalities to be fused by the symbols $\mathbf{x}$ and $\mathbf{y}$. A single observation is denoted by column-vectors $\mathbf{x}(t) \in \mathbb{R}^{M_\mathbf{x}}$ and $\mathbf{y}(t) \in \mathbb{R}^{M_\mathbf{y}}$, where $M_\mathbf{x}$ and $M_\mathbf{y}$ denote the number of recording channels in each modality. The matrices that contain all data points are denoted by $\mathbf{X} = [\mathbf{x}(1), ..., \mathbf{x}(T_\mathbf{x})] \in \mathbb{R}^{M_\mathbf{x} \times T_\mathbf{x}}$ and $\mathbf{Y} = [\mathbf{y}(1), ..., \mathbf{y}(T_\mathbf{y})] \in \mathbb{R}^{M_\mathbf{y} \times T_\mathbf{y}}$. Further symbols used in this article and their meaning is summarized in Table I.

B. The multimodal forward model

The central assumption we make is that the datasets are decomposable into what is called a set of *components* (or *factors*). The notion of a component underlies all of the models presented in this paper. An individual component is identified by a unique temporal and spatial signature and may thus be regarded as a functional unit. The component notation is congruent for $\mathbf{x}$ and $\mathbf{y}$, so we introduce the notation exemplary for $\mathbf{x}$ only. Let the scalar variable $s^i_\mathbf{x}(t)$ denote the temporal signature of a component with the index i at time point t. We will also refer to $s^i_\mathbf{x}(t)$ as the *temporal activity* of this component. The strength with which $s^i_\mathbf{x}(t)$ is expressed at each recording channel is called the *spatial activation pattern* and denoted by the column-vector[1] $\mathbf{a}^i_\mathbf{x} \in \mathbb{R}^{M_\mathbf{x}}$.

Generally a given dataset is assumed to be composed of a set of $K_\mathbf{x} \geq 1$ components. Let $\mathbf{s}_\mathbf{x}(t) \in \mathbb{R}^{K_\mathbf{x}}$ denote the column-vector which represents the temporal activity of the $K_\mathbf{x}$ components and let $\mathbf{A}_\mathbf{x} = \left[\mathbf{a}^1_\mathbf{x}, ..., \mathbf{a}^{K_\mathbf{x}}_\mathbf{x}\right] \in \mathbb{R}^{M_\mathbf{x} \times K_\mathbf{x}}$ denote the matrix in which each column contains the corresponding spatial activation patterns. We will make use of these variables when we consider the mapping from components to recording channels. This mapping is referred to as the *linear forward model*, *linear generative model*, or *encoding model*, for an in depth discussion of these terms see [54]. In this model, the projection of the components to the recording channels is given by

$$\begin{aligned} \mathbf{x}(t) &= \sum_i^{K_\mathbf{x}} \mathbf{a}^i_\mathbf{x} \cdot s^i_\mathbf{x}(t) + \boldsymbol{\epsilon}_\mathbf{x}(t) \\ &= \mathbf{A}_\mathbf{x}\mathbf{s}_\mathbf{x}(t) + \boldsymbol{\epsilon}_\mathbf{x}(t) \,, \end{aligned} \tag{1}$$

where $\boldsymbol{\epsilon}_\mathbf{x}(t) \in \mathbb{R}^{M_\mathbf{x}}$ captures activity that is not explained by the $K_\mathbf{x}$ components and thus considered *noise*. The task of factor model based analyses is to extract estimates of the underlying components from the data. We use $\hat{\mathbf{s}}_\mathbf{x}$ and $\hat{\mathbf{s}}_\mathbf{y}$ to denote these estimates.

The datasets $\mathbf{x}$ and $\mathbf{y}$ are assumed to be related by $K \leq \min(K_\mathbf{x}, K_\mathbf{y})$ pairs of shared component processes among the rows of $\hat{\mathbf{s}}_\mathbf{x}$ and $\hat{\mathbf{s}}_\mathbf{y}$. The exact nature of the relation between the shared components of course depends on the measurement modalities that are being used. However, a very generic connection between the modalities can be constructed based on the assumption that the datasets to represent a common timeline and thereby provide different views upon the

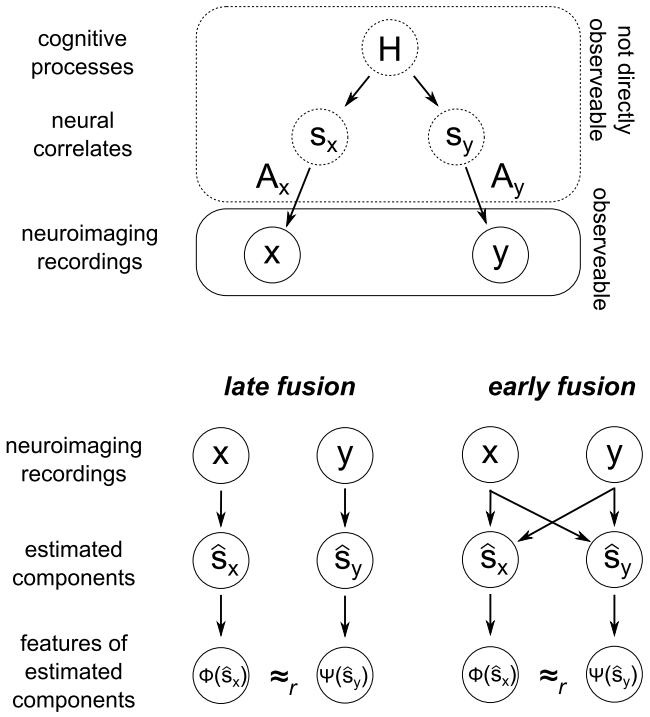

Fig. 1. A multimodal generative model (top) and two generic fusion approaches to multimodal data (bottom). A cognitive phenomenon ($\mathbf{H}$, e.g. attention, stimulus processing) influences certain aspects of modality specific neurophysiological processes, such as electrophysiological or metabolic properties. In the context of this generative model, these processes are modeled by latent variables (also called sources) and denoted by $\mathbf{s}_{\mathbf{x}/\mathbf{y}}$. These latent variables are mapped by a modality specific transformation ($\mathbf{A}_{\mathbf{x}/\mathbf{y}}$) to their respective sensor space variables ($\mathbf{X}/\mathbf{Y}$). Starting from the recorded datasets $\mathbf{X}$ and $\mathbf{Y}$, it is the task of factor model based methods to extract estimates of the latent sources ($\hat{\mathbf{s}}_{\mathbf{x}/\mathbf{y}}$) such that features of the estimated source activity ($\Phi(\hat{\mathbf{s}}_\mathbf{x})$) and $\Psi(\hat{\mathbf{s}}_\mathbf{y})$) are informative about $\mathbf{H}$ itself, or tell us something about how exactly $\mathbf{H}$ exerts influence on $\mathbf{s}_{\mathbf{x}/\mathbf{y}}$. In *early fusion* approaches, information from both modalities is already taken into account when extracting source activity from the data. In *late fusion* approaches, modality-specific sources are extracted without using information from the respective other modality first, and features of the estimated sources are combined thereafter.

[1]Please note that i is not the exponent of the variable but denotes the ith component.

$T_{\mathbf{x/y}}$	Number of data points, modality specific
$M_{\mathbf{x/y}}$	Number of measurement channels, modality specific
$K_{\mathbf{x/y}}$	Number of components (i.e. latent factors), modality specific
K	Number of joint components, i.e. components present in both modalities
$\mathbf{x}(t), \mathbf{y}(t)$	$M_{\mathbf{x}}/M_{\mathbf{y}}$-dimensional column-vector of observed data in modality $\mathbf{x/y}$
$\mathbf{X}, \mathbf{Y}$	$M_{\mathbf{x/y}} \times T_{\mathbf{x/y}}$ matrix containing the observed data, modality specific
$\mathbf{s}_{\mathbf{x/y}}(t), \hat{\mathbf{s}}_{\mathbf{x/y}}(t)$	$K_{\mathbf{x/y}}$-dimensional column-vector of (estimated) components
$\boldsymbol{\epsilon}_{\mathbf{x/y}}(t)$	$M_{\mathbf{x/y}}$-dimensional noise column-vector in forward models
$\mathbf{A}_{\mathbf{x/y}}$	$M_{\mathbf{x/y}} \times K_{\mathbf{x/y}}$ matrix of sensor-space activation patterns in forward models
$\mathbf{W}_{\mathbf{x/y}}$	$M_{\mathbf{x/y}} \times K_{\mathbf{x/y}}$ matrix of filters in backward models
$\mathbf{C}_{\mathbf{xx/yy}}$	Data covariance, modality specific
$\mathbf{C}_{\mathbf{xy/xy}}$	Cross-modal data covariance matrix
$R_{\mathbf{x/y}}$	Number of local brain sources included in a physical model
$\mathbf{U}_{\mathbf{x/y}}$	$R_{\mathbf{x/y}} \times 3$ spatial coordinates of locations of the modeled brain sources
$\mathbf{L}_{\mathbf{x/y}}$	$M_{\mathbf{x/y}} \times R_{\mathbf{x/y}}$ transfer matrix in physical models
$\mathbf{j}_{\mathbf{x/y}}(t), \hat{\mathbf{j}}_{\mathbf{x/y}}(t)$	$R_{\mathbf{x/y}}$-dimensional vector of (estimated) brain source activity
$\boldsymbol{\varepsilon}_{\mathbf{x/y}}(t)$	$M_{\mathbf{x/y}}$-dimensional noise vector in physical models
$\mathbf{F}_{\mathbf{x/y}}$	$R_{\mathbf{x/y}} \times K_{\mathbf{x/y}}$ matrix of source-space activation patterns in joint forward models
$\mathbf{j}_{\not\Leftrightarrow \mathbf{x/y}}(t)$	$R_{\mathbf{x/y}}$ vector of brain activity of no interest in joint forward models
$\mathbf{A}_{\not\Leftrightarrow \mathbf{x/y}}$	$M_{\mathbf{x/y}} \times K_{\mathbf{x/y}}$ part of the activation pattern matrix $\mathbf{A}_{\mathbf{x/y}}$ not explained by brain sources in joint forward models
$\boldsymbol{e}_{\mathbf{x/y}}(t)$	$M_{\mathbf{x/y}}$-dimensional noise vector in joint forward models

TABLE I
NOTATION.

same underlying processes. Therefore, it is to be expected that the time courses of shared components (or certain features of these time courses) exhibit "similar" dynamics. We formalize this notion of similarity by the following expression:

$$\Phi(\hat{s}^i_{\mathbf{x}}) \approx_{\mathrm{r}} \Psi(\hat{s}^i_{\mathbf{y}}), \tag{2}$$

for $i \in \{1, ..., K\}$. The functions $\Phi(\cdot)$ and $\Psi(\cdot)$ extract some feature from the time course of the component pair $\hat{s}^i_{\mathbf{x}}$ and $\hat{s}^i_{\mathbf{y}}$ that is similar in terms of a similarity metric $\approx_{\mathrm{r}}$. Examples for the feature extracting functions $\Phi(\cdot)$ and $\Psi(\cdot)$ could be simply the identity function, a function extracting spectral features, (de-)convolution operators, or functions extracting statistical properties of the distributions of $\hat{s}^i_{\mathbf{x}}$ or $\hat{s}^i_{\mathbf{y}}$. Examples for similarity measuring functions are functions that measure co-modulation in time, such as *covariance* or *correlation*. Another popular choice for the similarity metric is an information theoretic measures that is called *mutual information* (see section IV-A2 for a formal definition). Note that in contrast to covariance and correlation, mutual information captures nonlinear dependencies between variables.

The models we will discuss in section V will be characterized in terms of their specific choices of features to relate from the measurement modalities and the similarity measuring function. Figure 1 summarizes the notions presented in this subsection by outlining the generative model of multimodal neuroimaging data that is adopted here.

C. Estimating components using backward models: Filters

After having expressed the recorded data as a sum of underlying components, where each component is the product of a specific spatial and temporal signature, the question arises how to recover the components from the data. In the most general setting, the factors $\mathbf{A}_{\mathbf{x}}$ and $\mathbf{s}_{\mathbf{x}}$ in Eq. (1) are estimated jointly, a setting that is referred to as blind source separation (BSS). However, the factorization into $\mathbf{A}_{\mathbf{x}}$ and $\mathbf{s}_{\mathbf{x}}$ is not unique and therefore further assumptions have to be made about the nature of the spatial and temporal dynamics. As we will see in later sections, different assumptions lead to different factorization methods.

Estimating both the spatial activation patterns and the time-courses jointly leads to potentially difficult optimization problems. The computational complexity, however, can be reduced by resorting to a so-called linear discriminative (also called *backward* or *decoding*) modeling approach, for a detailed discussion on these types of models and their relationship to *forward models* see [54]. In such an approach, the time-courses of K neural sources are estimated by projecting the data linearly onto a set of spatial *extraction filters* $\mathbf{W}_{\mathbf{x}} = \left[\mathbf{w}^1_{\mathbf{x}}, ..., \mathbf{w}^K_{\mathbf{x}}\right] \in \mathbb{R}^{M_x \times K}$:

$$\hat{\mathbf{s}}_{\mathbf{x}}(t) = \mathbf{W}^{\top}_{\mathbf{x}} \mathbf{x}(t). \tag{3}$$

Note that sometimes (for example in the ICA community) a different convention is adopted in which the extraction filters are in the rows of $\mathbf{W}_{\mathbf{x}}$ instead of the columns as we introduce it here. The coefficients of $\mathbf{W}_{\mathbf{x}}$ determine how to integrate the information from all recording channels in order to optimally extract the time-courses of the components. Several approaches to find, or rather to optimize, these coefficients will be presented in the following sections. However, at this point it is important to discuss some common misconceptions about the interpretability of the coefficients of filters, once they have been obtained.

A prerequisite for determining the anatomical origin and neurophysiological relevance (that is, for enabling "neurophysiological interpretation") of extracted time-courses is to identify the strength with which the time-courses are expressed at each recording channel. Importantly, the coefficients of extraction filters do *not* encode this information and should therefore not be interpreted with respect to the origin of the extracted signal. This is only possible for the activation patterns of forward models [54], [55]. Moreover, it is only the activation patterns that can be subjected to source localization

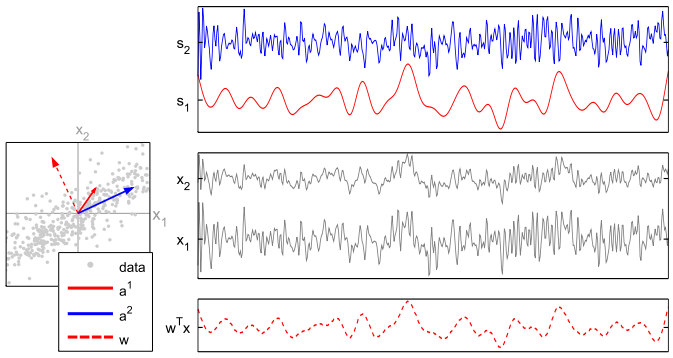

Fig. 2. Illustration of the difference between extraction filters (i.e. the coefficients of backward models) and spatial activation patterns (i.e. the coefficients of forward models) by using simulated data from only one hypothetical modality. The top panel on the right side shows the time-courses of two hidden source components, s_1 and s_2, of which s_1 shall be the signal of interest in this example and s_2 corresponds to a noise component. These time-courses are mapped to two recording channels (x_1 and x_2) by means of Eq. (1), using the matrix $\mathbf{A}_\mathbf{x} = \left[\mathbf{a}^1, \mathbf{a}^2\right]$. The time-courses of the data in channel space is shown in the middle panel on the right side, as well as in a scatter plot on the left side (x_1 on the abscissa and x_2 on the ordinate). The scatter plot also shows the activation patterns (columns of the matrix $\mathbf{A}_\mathbf{x}$) as solid line vectors. Note that $\mathbf{a}^1$ is only half as long as $\mathbf{a}^2$, which means that the noise component s_2 is expressed much stronger in the channel data. Using s_1 as a target variable, the vector $\mathbf{w}$ is the extraction filter of a backward model optimized by means of Eq. (20). $\mathbf{w}$ is shown in the scatter plot as a dashed line vector. Applying $\mathbf{w}$ to the data, i.e. computing $\mathbf{w}^\top\mathbf{x}(t)$, yields a reconstruction of s_1, see the lower panel on the right side. Importantly, while $\mathbf{w}$ extracts the time-course of component s_1, its coefficients are not to be interpreted as to how strong and with what sign s_1 was expressed in the data. Instead, only the coefficients of $\mathbf{a}^1$ contain that information. However, an estimate of $\mathbf{a}^1$ can be derived from $\mathbf{w}$ by means of Eq. (4).

techniques (see below) in order to link cognitive functions to specific brain areas. See Fig. 2 for an illustration of the duality between filters and patterns.

D. Recovering the forward model from a backward model: Patterns

Earlier we have identified interpretability to be one of the key properties that are desired in (multimodal) neuroimaging. In [54] has been established that extraction filters of backward models cannot be interpreted in terms of the studied brain processes (that is, be used to localize these processes to individual sensors). This is due to the fact that extraction filters are generally functions of the signal and the noise and thus heavily influenced by factors not of interest for the neurophysiological interpretation. As a remedy, a corresponding forward model of the form of Eq. (3) may be derived from every linear backward model, the activation patterns of which can be interpreted in the aforementioned way. The transformation of backward model extraction filters into forward model activation patterns is given by

$$\begin{aligned}\mathbf{A}_\mathbf{x} &= \mathbf{C}_{\mathbf{xx}}\mathbf{W}_\mathbf{x}\mathbf{C}^{-1}_{\hat{\mathbf{s}}_\mathbf{x}\hat{\mathbf{s}}_\mathbf{x}} \\ &= \mathbf{C}_{\mathbf{xx}}\mathbf{W}_\mathbf{x}\left(\mathbf{W}_\mathbf{x}^\top\mathbf{C}_{\mathbf{xx}}\mathbf{W}_\mathbf{x}\right)^{-1}\ , \end{aligned} \tag{4}$$

where $\mathbf{C}_{\mathbf{xx}}$ denotes the data covariance matrix and $\mathbf{C}_{\hat{\mathbf{s}}_\mathbf{x}\hat{\mathbf{s}}_\mathbf{x}}$ denotes the covariance matrix of component time-courses.

By virtue of the transformation (4), we can pursue a backward modeling approach, allowing us to conveniently parametrize cost functions solely in terms of the extraction filters (see next section), while being able to achieve neurophysiological interpretability and source localization through analysis of the activation patterns of the corresponding forward model.

E. Source localization for factor models

Unlike backwards models, forward models of the form Eq. (1) allow to identify those sensors that are related to the brain activity under study, and thereby to localize the components of a factor model *in sensor space*. For imaging modalities such as fMRI, for which a one-to-one relationship between sensors and brain locations exist, the analysis of forward model activation patterns is thus sufficient to enable conclusions about the brain areas involved in the studied brain process.

For modalities such as EEG, MEG or fNIRS, which measure effects of brain activity only outside the head, a *source space* representation of the sensor readings has to be inferred in order to draw similar conclusions. To this end, a *physical model* is required, which describes how neural (source) activity in the brain maps to the sensors. Such a model comprises information about the geometries of the different tissues (gray matter, white matter, cerebrospinal fluid, skull, skin) in the studied head, as well as modality-specific properties of these tissues. In case of EEG and MEG, it needs to describe the flow of the extra-cellular ionic return currents occuring in response to the intra-cellular neuronal activity through the volume. The physical properties of interest here are the tissues' electrical conductivities as well as inhomogeneities and anisotropies. Since the quasi-static approximation of Maxwell's equations holds for the frequencies typically studied in EEG/MEG, the relationship between source neuronal currents and the EEG/MEG scalp electrical potentials/magnetic fields generated by the corresponding return currents is linear [56].

In the case of fNIRS, the physical model describes the photon transport through the tissue, and involves optical properties such as absorption and scattering coefficients of different tissue types. The relationship between the internal coefficients in the brain reflecting neuronal activity indirectly through blood de-/oxygenation and the respective coefficients measured at the scalp surface is generally non-linear; however, for small changes in absorption coefficients a linear approximation is reasonable [57]. Therefore, we can assume the following physical model in all of the discussed cases:

$$\mathbf{x}(t) = \mathbf{L}_\mathbf{x}\mathbf{j}_\mathbf{x}(t) + \boldsymbol{\varepsilon}_\mathbf{x}(t)\ . \tag{5}$$

Here, the time-dependent $R_\mathbf{x}$-dimensional vector $\mathbf{j}_\mathbf{x}(t)$ describes the brain source activity at $R_\mathbf{x}$ distinct locations $\mathbf{U}_\mathbf{x} \in \mathbb{R}^{R_\mathbf{x}\times 3}$ in the brain (e. g., points on the cortical surface in MNI coordinates). Notice that for electrophysiological imaging modalities, the source activity is in fact a vector field, since each source location emits a directed electrical current. However, for simplicity, we here assume that the orientations of these currents are fixed (e. g. to be perpendicular to the local cortical surface, which is the predominant direction of the pyramidal cells thought to be the main generators of the

EEG signal). The $M_{\mathbf{x}} \times R_{\mathbf{x}}$ *transfer matrix* $\mathbf{L}_{\mathbf{x}}$ describes the relationship between the source brain activity and the sensor readings, and comprises all geometrical and physical properties of the head discussed above. It it is called the *lead field* in case of EEG/MEG. Finally, $\boldsymbol{\varepsilon}_{\mathbf{x}}(t)$ is a $M_{\mathbf{x}}$-dimensional noise vector.

Given a physical model (that is, transfer matrix $\mathbf{L}_{\mathbf{x}}$), *source localization* of EEG/MEG or fNIRS activity can be carried out by estimating the brain source activity $\mathbf{j}_{\mathbf{x}}(t)$ at locations $\mathbf{U}_{\mathbf{x}}$ giving rise to the measured signals $\mathbf{x}(t)$. This amounts to solving an ill-posed inverse problem, and can be done by introducing prior assumptions on the spatial and/or temporal characteristics of the source activity (e. g., [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [57]).

How exactly factor models relate to physical models needed for source localization has rarely been made explicit in the literature. In the following, we will work out the respective relationships for the case of linear models. Obviously, the physical model Eq. (5) has the same structure as Eq. (1) introduced in the context of factor models, and is in fact also a forward model of the data. However, the meaning of the variables in the two models differs in the following terms.

- The factor model Eq. (1) assumes a *small number* $K_{\mathbf{x}} \leq M_{\mathbf{x}}$ of *components* $\mathbf{s}_{\mathbf{x}}(t)$, each of which captures the activity of a potentially spatially distributed "network" of brain areas, and is thought to have a distinct functional role in terms of the brain processes under study (e. g., those whose activity is consistent across imaging modalities). Components are typically assumed to be mutually uncorrelated if not even statistically independent.
- The physical model Eq. (5), on the other hand, models activity of a *large number* of $R_{\mathbf{x}} \geq M_{\mathbf{x}}$ *brain sources* $\mathbf{j}_{\mathbf{x}}(t)$, each of which corresponds to a single location in the brain. The known relationship between brain sources $\mathbf{j}_{\mathbf{x}}(t)$ and their locations $\mathbf{U}_{\mathbf{x}}$ established by the transfer matrix $\mathbf{L}_{\mathbf{x}}$ enables time-resolved source localization of the measured activity $\mathbf{x}(t)$. Unlike factor model components, different brain sources may very well be correlated and relate to the same cognitive component.
- Technically, the term $\mathbf{A}_{\mathbf{x}}\mathbf{s}_{\mathbf{x}}(t)$ in Eq. (1) captures the part of the data that is correlated with the brain processes of interest, and may consist of genuine brain activity but also artifacts originating outside the brain, while the term $\mathbf{L}_{\mathbf{x}}\mathbf{j}_{\mathbf{x}}(t)$ in Eq. (5) is the part of the data that is explained by physical sources in the brain, and may contain activity related to the brain processes of interest but also unrelated activity.
- Conversely, the noise term $\boldsymbol{\epsilon}_{\mathbf{x}}(t)$ in Eq. (1) captures all measured activity that is not explained by any of the $K_{\mathbf{x}}$ factors and therefore unrelated to the brain processes under study regardless of whether it can be explained by sources in the brain or not, while $\boldsymbol{\varepsilon}_{\mathbf{x}}(t)$ in Eq. (5) captures all activity that is not explained by sources physically located in the brain regardless of whether it is related to the brain processes under study or not.

We have seen that each component $s^i_{\mathbf{x}}(t)$ of a factor model can be localized to sensors through its static activation pattern $\mathbf{a}^i_{\mathbf{x}}$. To achieve a localization to actual brain anatomy, we need to derive analogous patterns in source space. This can be achieved by merging physical and factor models into a theoretical *joint* forward model

$$\mathbf{x}(t) = \underbrace{\mathbf{L}_{\mathbf{x}}\mathbf{j}_{\not\mathbf{s}_{\mathbf{x}}}(t)}_{\text{I. brain only}} + \underbrace{\mathbf{A}_{\not\mathbf{L}_{\mathbf{x}}}\mathbf{s}_{\mathbf{x}}(t)}_{\text{II. factors only}} + \underbrace{\mathbf{L}_{\mathbf{x}}\mathbf{F}_{\mathbf{x}}\mathbf{s}_{\mathbf{x}}(t)}_{\textbf{III. both}} + \underbrace{\boldsymbol{e}_{\mathbf{x}}(t)}_{\text{IV. none}} \tag{6}$$

decomposing the data into parts explained by I. brain processes of no interest, II. artifacts of non-cerebral origin correlated with the brain processes of interest, III. the brain activity of interest and IV. artifacts not correlated with the brain activity of interest. Here, $\mathbf{j}_{\not\mathbf{s}_{\mathbf{x}}}(t) \in \mathbb{R}^{R_{\mathbf{x}}}$ is the brain activity unrelated to any of the $K_{\mathbf{x}}$ factors, $\mathbf{A}_{\not\mathbf{L}_{\mathbf{x}}} \in \mathbb{R}^{M_{\mathbf{x}} \times K_{\mathbf{x}}}$ is the part of the activation pattern matrix that is not explained by brain sources, $\boldsymbol{e}_{\mathbf{x}}(t)$ is non-cerebral noise unrelated to any factor, and finally $\mathbf{F}_{\mathbf{x}} \in \mathbb{R}^{R_{\mathbf{x}} \times K_x}$ are the desired source-space activation patterns localizing each of the K_x components $\mathbf{s}_{\mathbf{x}}(t)$ to brain anatomy.

It is easy to see that the joint forward model Eq. (6) can be obtained from the factor model Eq. (1) by splitting activation pattern and noise terms into parts that can or cannot be explained by physical brain sources through

$$\mathbf{A}_{\mathbf{x}} = \mathbf{L}_{\mathbf{x}}\mathbf{F}_{\mathbf{x}} + \mathbf{A}_{\not\mathbf{L}_{\mathbf{x}}} \tag{7}$$

$$\boldsymbol{\epsilon}_{\mathbf{x}}(t) = \mathbf{L}_{\mathbf{x}}\mathbf{j}_{\not\mathbf{s}_{\mathbf{x}}}(t) + \boldsymbol{e}_{\mathbf{x}}(t)\ . \tag{8}$$

Analogously, the same model can be recovered from the physical model Eq. (5) by splitting brain activity and noise terms into parts that are or are not related to the studied processes via

$$\mathbf{j}_{\mathbf{x}}(t) = \mathbf{F}_{\mathbf{x}}\mathbf{s}_{\mathbf{x}}(t) + \mathbf{j}_{\not\mathbf{s}_{\mathbf{x}}}(t) \tag{9}$$

$$\boldsymbol{\varepsilon}_{\mathbf{x}}(t) = \mathbf{A}_{\not\mathbf{L}_{\mathbf{x}}}\mathbf{s}_{\mathbf{x}}(t) + \boldsymbol{e}_{\mathbf{x}}(t)\ . \tag{10}$$

The decompositions Eq. (7)–(10) suggest that source-space activation patterns $\mathbf{F}_{\mathbf{x}}$ can be obtained without actually carrying out simultaneous estimation of all parameters of the joint forward model in three equivalent ways. First, by exploiting that Eq. (7) is a static version of the physical model Eq. (5), sensor-space activation patterns $\mathbf{a}^i_{\mathbf{x}}$ that have been pre-estimated by factor modeling may be mapped to their source-space equivalents $\mathbf{f}^i_{\mathbf{x}}$ using any inverse source localization method that makes assumptions only on spatial but not temporal properties of the brain sources [59], [60], [62], [64], [66], [65]. Second, by noting that for brain sources $\hat{\mathbf{j}}_{\mathbf{x}}(t)$ that have been pre-estimated from the entire data by source localization methods, Eq. (9) is a source-space version of the factor model Eq. (1), source-space activation patterns $\mathbf{F}_{\mathbf{x}}$ may be obtained by factor modeling using any of the component analysis techniques introduced below. Third, in case that estimates $\hat{\mathbf{s}}_{\mathbf{x}}(t)$ and $\hat{\mathbf{j}}_{\mathbf{x}}(t)$ of both component and brain activations have previously been obtained through appropriate factor and physical modeling, Eq. (9) takes the particularly simple form of a general linear model (GLM), allowing one to estimate $\mathbf{F}_{\mathbf{x}}$ using linear regression. In all cases, the entries of the estimated source-space activation pattern $\mathbf{f}^i_{\mathbf{x}}$ indicate the strength and effect direction with which the i-th factor is expressed at each brain location and thereby link that component's activity to its generating brain structures.

IV. Methods for late fusion

In this section we review multivariate backward models for the extraction of components from a single measurement modality. In the context of multimodal fusion these methods are applied in *late fusion* scenarios, because information from the respective other modality is not considered during the extraction of components. For exemplary purposes we use the variable $\mathbf{x}$ here to represent the observed data, regardless of the measurement modality. Additionally we drop the subscript that indicates the modality from all other variables because in this section there is no need for it.

A further subdivision can be made into *supervised* and *unsupervised* methods. Supervised methods make use of an external target signal during the optimizing of the parameters, while unsupervised methods rely on the statistics of the data alone. Supervised methods are also used in so-called *asymmetric fusion settings* of multimodal analyses. In this setting, one modality, or features extracted from one modality, are used as labels or regressors in order to extract factors from another modality. Examples include studies on correlations of the occipital EEG alpha band power with fMRI data [69] or fNIRS data [50].

A reoccurring theme in all the methods is to guide the search for the weight vector $\mathbf{w}$ (or an entire set of vectors represented by the matrix $\mathbf{W}$) by means of optimizing an *objective function*. We would like to emphasize that for the majority[2] of methods we will discuss below, the objective function takes the form

$$\max_{\mathbf{w}}/\min_{\mathbf{w}} \ \mathbf{w}^\top \mathbf{B}_1 \mathbf{w}, \quad \text{s.t. } \mathbf{w}^\top \mathbf{B}_2 \mathbf{w} = c, \tag{11}$$

where c is a constant. The methods we discuss differ with respect to the choice of the matrices $\mathbf{B}_1$ and $\mathbf{B}_2$. However, if an objective function can be expressed in the above form, the solution is obtained as the solution to the corresponding generalized eigenvalue problem

$$\mathbf{B}_1 \mathbf{w} = \lambda \mathbf{B}_2 \mathbf{w}, \tag{12}$$

where λ denotes what is called the generalized *eigenvalue* that is associated with the *eigenvector* $\mathbf{w}$. Generalized eigenvalue problems have been studied for decades in the field of numerical linear algebra, which has lead to efficient algorithms for solving them [70], [71]. Being able to cast an objective function into the form of a generalized eigenvalue problem is desirable, because it can then be solved using standard numerical linear algebra tools such as MATLAB or R, for example.

A. Unsupervised approaches

1) Principal Component Analysis: Perhaps the most popular and most widely used unsupervised factorization method is the principal component analysis (PCA) [72], [73]. The underlying idea in PCA is to find components in the data that account for as much variance as possible under the constraint that the components are mutually de-correlated.

[2]A notable exception are algorithms for independent component analysis (ICA) discussed in Sections IV-A2 and V-A.

Let us formalize an objective for PCA for a single component. The coefficients of the weight vector $\mathbf{w}$ are to be optimized such that the extracted signal $\mathbf{w}^\top \mathbf{x}(t)$ has maximum variance:

$$\max_{\mathbf{w}} \mathrm{Var}\left(\mathbf{w}^\top \mathbf{x}(t)\right), \quad \text{s.t. } \|\mathbf{w}\|^2 = 1 \tag{13}$$

Expressing the variance of $\mathbf{w}^\top \mathbf{x}(t)$ as

$$\mathrm{Var}(\mathbf{w}^\top \mathbf{x}(t)) = \mathbf{w}^\top \mathbf{C} \mathbf{w}, \tag{14}$$

where the matrix $\mathbf{C}$ is the covariance matrix of the data, we arrive at

$$\max_{\mathbf{w}} \ \mathbf{w}^\top \mathbf{C} \mathbf{w}, \quad \text{s.t. } \mathbf{w}^\top \mathbf{w} = 1. \tag{15}$$

which corresponds to Eq. (11) with $\mathbf{B}_1 = \mathbf{C}$ and $\mathbf{B}_2 = \mathbf{I}$. Thus the corresponding eigenvalue equation is given by

$$\mathbf{C} \mathbf{w} = \lambda \mathbf{w}, \tag{16}$$

and the solution is obtained as the eigen-decomposition of the covariance matrix $\mathbf{C}$.

It can be shown that the eigenvalue of a PCA component corresponds to its variance, i.e. $\lambda = \mathbf{w}^\top \mathbf{C} \mathbf{w}$. Thus the fraction of total variance explained by a subset of $K \leq M$ components is given by the ratio $\sum_i^K \lambda_i / \sum_j^M \lambda_j$. This ratio is often used to determine the size of a suitable PCA component subset that together explains a given percentage of the total variance contained in the data. Here the idea is that the set of components that explains most of the variance in the data are the most "interesting" ones.

***Applications**:* For example, in [74] PCA was used in a multimodal setting involving concurrent EEG and MEG recordings to determine that $K \approx 2$ EEG components explain about 50% variance of sleep spindles while $K \geq 15$ MEG components necessary to account for the same amount of total variance. The findings lead the authors to conclude that the two measurement modalities reflect the activity of different system of neural sources during spindles.

While in the previous example the first K components were deemed the interesting ones, PCA is also often used to *remove* the components with maximal variance from the data, because these are likely to correspond to strong noise that contaminates the actual signal of interest. For example, in simultaneous recordings of fMRI and electrophysiological measures, the changing magnetic field of the MRI scanner induces artifacts in the electrophysiological recordings that are larger by many orders of magnitude. PCA hase been used to clean the data by removing the highest variance components in the context of scanning artifacts [75], [76], pulse artifacts [77], [76], or line noise [78].

2) Independent Component Analysis: A potentially limiting aspect of PCA is the fact that the spatial activations patterns of PCA components are constrained to be orthogonal[3]. This

[3]For PCA, the weight vectors $\mathbf{W}$ are the eigenvectors of the covariance matrix $\mathbf{C}$, i.e. it holds that $\mathbf{C} = \mathbf{W} \Lambda \mathbf{W}^\top$, where Λ is a diagonal matrix and $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$. Using the last two equations and substituting Eq. (16) into Eq. (4) reveals that for PCA it holds that $\mathbf{A} = \mathbf{W}$.

assumption may be too strong in the context of neurophysiological activation patterns. An alternative unsupervised factorization method that does not impose such constraints on the patterns is the independent component analysis (ICA).

ICA is based on the idea that the hidden components are statistically independent. Let $\hat{\mathcal{S}}$ denote the random variable that contains the temporal signature of the extracted components. Then $\hat{\mathcal{S}}$ is parametrized by the weight matrix $\mathbf{W}$, by virtue of the backward modeling approach in Eq. (3). The notion of maximal independence between the individual components, denoted by the random variables $\hat{\mathcal{S}}_i$ for $i \in \{1, ..., K\}$ is equivalent to the notion of minimizing the mutual information (MI) between them. Mathematically, the MI between two variables X and Y is defined as the Kullback-Leibler divergence D_{KL} between the joint probability distribution of X and Y (denoted as p) and the product of the marginal probability distributions

$$\begin{aligned} \mathcal{I}(X,Y) &= D_{KL}(p(X,Y) \,||\, p(X)p(Y)) \\ &= \int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \,. \end{aligned} \quad (17)$$

Furthermore, the mutual information of $\hat{\mathcal{S}}_i$ can be expressed as

$$\mathcal{I}(\hat{\mathcal{S}}) = \sum_i^K \mathcal{H}(\hat{\mathcal{S}}_i) - \mathcal{H}(\hat{\mathcal{S}}), \quad (18)$$

where $\mathcal{H}(\hat{\mathcal{S}}_i)$ denotes the *entropy* of $\hat{\mathcal{S}}_i$. It can be shown that minimizing $\mathcal{I}(\hat{\mathcal{S}})$ can be achieved by minimizing the entropy for all individual components. Since the Gaussian distribution has the maximal entropy among distributions with fixed mean and variance, the mutual information between components can be minimized by extracting components with maximally non-Gaussian distributions. A number of algorithms exist that are based on the this idea (e.g. [79], [80], [81]).

A different approach to ICA is taken by methods that exploit temporal information. These methods are based on the joint (approximate) diagonalisation of time-lagged covariance matrices. Examples are described in [82] and [83].

Note that the independence assumption used in ICA can be applied to either the estimated time courses of the components (as was outlined above) or to their estimated activation patterns. The former approach is referred to as *temporal* ICA, while the latter is called *spatial* ICA. In the context of fMRI, spatial ICA is the more popular version, while in the context of EEG and MEG, temporal ICA is used. See [84] for more discussion on the choice between spatial and temporal ICA.

***Applications**:* ICA algorithms are widely used in preprocessing data to separate artifactual components from components of neural origin, see [85], [86], for example. In the context of multimodal measurements, ICA has proven useful to identify pulse and scanner artifacts [87], [88] and thereby greatly improve the signal quality compared to the non-corrected signal. However, a study that compared several versions of ICA as well as temporal PCA-based approaches [76] in the context of simultaneously acquired EEG and fMRI found that ICA- and PCA-based approaches perform equally well, with ICA requiring more parameter tuning. Applying both approaches in sequence can improve over either approaches individually [89].

Neural oscillations were investigated using EEG/fMRI in study presented in [48]. ICA was used to extract components from the EEG that reflect the sensorimotor rhythm during a movement task. The bandpower time-course of these ICA components correlated inversely with activation in the pre- and postcentral cortex as revealed by fMRI. Differential effects were found for alpha (8 Hz to 12 Hz) and beta (12 Hz to 30 Hz) power, with beta power yielding stronger correlations between the EEG components and the fMRI.

The study reported in [90] used ICA to investigate the origin of auditory ERPs during simultaneous recordings of EEG and fMRI. ICA was applied to both imaging modalities separately. For fMRI, spatial ICA was applied. From the resulting decompositions, one pair showed significant correlations between time-courses. The application of ICA separately to each modality prior to fusion was also adopted in [91].

See [92] for further application examples of ICA.

B. Supervised approaches

In this section we assume that in addition to the data from the imaging modality, we are also given an external target signal, denoted by the scalar variable z. This variable may encode additional information about the stimulus (e.g. type, intensity, latency, etc.), behavioral measurements (reaction times, ratings, etc.), external physiological parameters (skin conductance, heart rate, etc.), or artifactual information (e.g. eye movements, motion parameters, etc.). In general, supervised methods have an advantage over unsupervised approaches because they have more information at their disposal.

1) Regression and classification: Two well known examples of supervised factor models are linear regression and classification by means of linear discriminant analysis (LDA). We will first examine linear regression and then treat LDA as a special case of the former.

The goal of regression is to extract a component with a time-course that co-modulates with the target variable z. Without loss of generality, we assume z to have zero mean and unit variance. One way of quantifying co-modularity between two time series is by way of the *mean squared error* (MSE), given by

$$\mathrm{MSE}\left(\mathbf{w}^\top \mathbf{x}(t), z(t)\right) = \frac{1}{T} \sum_t^T \frac{1}{2} \left(\mathbf{w}^\top \mathbf{x}(t) - z(t)\right)^2 . \quad (19)$$

The spatial filter that minimizes the MSE is given by

$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \mathbf{X}\mathbf{z}^\top, \quad (20)$$

where the row-vector $\mathbf{z} = (z(1), ..., z(T)) \in \mathbb{R}^{1\times T}$ contain the time-course of the target variable and the matrix $\mathbf{X}$ contains the measured data and was defined earlier. This is known as the ordinary least squares (OLS) solution.

Interestingly, the same solution is obtained for the following objective function, which expresses co-modularity in terms of covariance between $\mathbf{w}^\top \mathbf{x}(t)$ and $z(t)$:

$$\max_{\mathbf{w}} \mathrm{Cov}\left(\mathbf{w}^\top \mathbf{x}(t), z(t)\right), \quad \text{s.t. Var}\left(\mathbf{w}^\top \mathbf{x}(t)\right) = 1. \quad (21)$$

Or equivalently in matrix notation expressed as

$$\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{X}\mathbf{z}^\top, \quad \text{s.t. } \mathbf{w}^\top \mathbf{C}\mathbf{w} = 1. \tag{22}$$

Classification of two conditions (or classes) can be treated within the regression framework outlined above. In such a scenario, the target variable z takes on only two values that indicate class membership. Eq. (20) yields the filter that achieves optimal class separation by choosing $z(t) = p(c_1)$ for all time points t that belong to class 1 and $z(t) = -p(c_2)$ for all time points t that belong to class 2, where $p(c_1)$ denotes the prior probability of class 1 and $p(c_2)$ the prior probability of class 2 (see, for example, chapter 4 in [93]. The resulting algorithm is called linear discriminant analysis (LDA).

Applications: Linear regression idea is a special case of the general linear model (GLM) framework that has been successfully applied in the context of unimodal and asymmetric multimodal analysis of fMRI data for almost two decades [94]. In the context of fMRI, the target variable is usually the time-course of an fMRI voxel, while $\mathbf{x}$ is called the *design matrix*. Each column of the design matrix contains a *regressor*, which are explanatory variables such as stimulus level or task-condition for example. As this approach is usually applied to all voxels separately, it is referred to as *mass-univariate* analysis. In the context of EEG, regression-based approaches can be very useful for the extraction and removal of eye movement artifacts [95], for example.

Linear classification methods such as LDA have been found to yield very good performance for fMRI [96] as well as for EEG [55] in unimodal settings. In multimodal settings, LDA has been applied in [49], which we will discuss in more detail in the next subsection. Other examples include, but are not limited to, the studies described in [97], [98]. In [98], LDA was used to extract a component from EEG recordings that best discriminates between two conditions (target vs. standard stimuli). Then the single trial variability of the LDA projection was used as a regressor GLM analysis of the simultaneously recorded fMRI. This procedure revealed that both task dependent as well as task independent networks of fMRI voxels contributed to fluctuations in attention.

2) Regression and classification using band-power features: Given the interest in generators of neural oscillations, we cover supervised backward models that extract oscillatory sources next. Here the instantaneous amplitude (also called *envelope*) is often the subject of investigation. A useful approximation of the (squared) envelope is given by computing the variance of the narrow-band signal in short consecutive time windows, which we refer to as *epochs*. Using the variance approximation of band-power, it is possible to derive algorithms for regression and classification analogously to the previous section. The important difference is that here not the projected signal itself is assumed to co-modulate with the target variable z. Instead the epoch-wise variance (i.e. the power time-course) of $\mathbf{w}^\top \mathbf{x}$ is assumed to co-modulate with z.

Let $\mathbf{X}_e \in \mathbb{R}^{M_\mathbf{x} \times T_\mathbf{x}(e)}$ denote the matrix that contains all samples within an epoch, where the epoch is indexed by e and $T_\mathbf{x}(e)$ denotes all time indices withing the e-th epoch. Because we are using the variance approximation of spectral power in a given frequency band, we assume $\mathbf{x}$ to be bandpass filtered for the band of interest. We further denote the bandpower of $\mathbf{w}^\top \mathbf{x}$ within epoch e by $\phi_\mathbf{w}(e)$, which we define as

$$\phi_\mathbf{w}(e) \stackrel{\text{def}}{=} \mathbf{w}^\top \mathbf{C}(e) \mathbf{w}, \tag{23}$$

where $\mathbf{C}(e)$ denotes the covariance matrix of $\mathbf{x}$ computed for the epoch e, similar to Eq. (14).

Next we formulate the co-modulation objective for $\phi_\mathbf{w}$ and z as a function of $\mathbf{w}$. In analogy to Eq. (21) we use the covariance as a measure for co-modulation, as this allows for a (near) analytical solution. The resulting algorithm is called *source power co-modulation* (SPoC) and was presented in [99]. The objective for source power co-modulation is thus given by

$$\max_{\mathbf{w}} \ \mathrm{Cov}\left(\phi_\mathbf{w}(e), z(e)\right), \quad \text{s.t.} \, \mathrm{Var}\left(\mathbf{w}^\top \mathbf{x}(t)\right) = 1. \tag{24}$$

Again, assuming zero-mean for z, the covariance reduces to the product between the two variables, averaged over epochs:

$$\begin{aligned} \mathrm{Cov}\left(\phi_\mathbf{w}(e), z(e)\right) &\propto \sum_e (\mathbf{w}^\top \mathbf{C}(e) \mathbf{w} \cdot z(e)) \\ &= \mathbf{w}^\top \underbrace{\left(\sum_e \mathbf{C}(e) \cdot z(e) \right)}_{\stackrel{\text{def}}{=} \mathbf{C}_z} \mathbf{w} \ . \end{aligned} \tag{25}$$

Using the last definition we can transform the SPoC objective in Eq.(24) into

$$\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{C}_z \mathbf{w}, \quad \text{s.t. } \mathbf{w}^\top \mathbf{C}\mathbf{w} = 1, \tag{26}$$

which can be further transformed into the following generalized eigenvalue problem:

$$\mathbf{C}_z \mathbf{w} = \lambda \mathbf{C}\mathbf{w}, \tag{27}$$

which is another example of the generalized eigenvalue equation seen above.

In analogy to the previous section on regression and LDA, it can be shown that scenarios in which the bandpower of a component is used to discriminate between two conditions can be subsumed by the SPoC framework with the appropriate choice for the values of z. By again choosing the values of z to reflect the prior probabilities of the two classes (i.e. $z(e) = p(c_1)$ if epoch e belongs to class 1 and $z(e) = -p(c_2)$ otherwise) the SPoC generalized eigenvalue problem turns into

$$(\mathbf{C}_1 - \mathbf{C}_2)\mathbf{w} = \lambda \mathbf{C}\mathbf{w}, \tag{28}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ denote the covariance matrices of the two classes and $\mathbf{C} = \mathbf{C}_1 + \mathbf{C}_2$. The last equation is the solution to the objective function of the common spatial patterns (CSP) algorithm, which is very popular in the field of Brain-Computer Interfaces [100].

Applications: Both CSP and LDA were used in a recent multimodal study to fuse EEG and fNIRS recordings in the context of sensorimotor rhythm (SMR)-based Brain-Computer Interface (BCI) [49]. SMR-based BCIs rely on the voluntary modulation of motor-related μ- and β-bands, which can be induced by actual as well as imagery movements [21], [101], [102]. In the study presented in [49] an EEG-based

classifier for the detection of bandpower changes in a specific frequency range was derived based on a temporal (i.e. band-pass) filter as well as a spatial filter (CSP) in addition to a linear classifier (here LDA). Raw fNIRS was converted into concentration changes of oxygenated [HbO] and deoxygenated [HbR] hemoglobin and the resulting features for each of the 24 fNIRS channels were used to estimate two LDA classifiers: one for [HbO] and one for [HbR]. Finally, the outputs of these three *unimodal* LDA classifiers (one derived from EEG and two from fNIRS) were combined with an *LDA meta-classifier*, which weighs the three signals according to the calibration data.

A different study, reported in [103], used CSP and LDA to deliver real-time feedback of SMR modulations induced by motor imagery (MI), here imagined hand movements. fMRI was recorded at the same time and processed offline. In order to gain insights into the relationship between MI EEG feedback and cortical fMRI signals, extracted EEG bandpower dynamics of the SMR rhythm were related to fMRI activations using GLM analysis. This analysis approach revealed that both EEG and fMRI showed significantly more MI-related activity during feedback blocks compared to no feedback.

For a more detailed review on various hybrid concepts for neurofeedback and BCI, we would like to refer the interested reader to [104].

V. Methods for early fusion

In this section we discuss factor models that are designed to decompose two (or more) datasets at the same time. These approaches integrate information from both measurement modalities for the extraction of components, which makes them applicable in early fusion scenarios. For simplicity we here assume just two modalities, denoted by $\mathbf{x}$ and $\mathbf{y}$, but the concepts presented below can be extended to more than two modalities. In the context of simultaneous measurements of electrophysiology and hemodynamics, $\mathbf{x}$ represents the former and $\mathbf{y}$ the later.

A. Multimodal versions of ICA

Joint ICA (jICA), presented in [105], is a method that enables fusion of multimodal features from several of subjects. Let N_s denote the number of subjects and $\mathbf{D_x} \in \mathbb{R}^{N_s \times N_\mathbf{x}}$ and $\mathbf{D_y} \in \mathbb{R}^{N_s \times N_\mathbf{y}}$ denote the matrices that contain features from the $\mathbf{x}$ and $\mathbf{y}$ modality, respectively.

In the next step the features from the modalities are simply concatenated along the horizontal to yield a multimodal feature matrix $\mathbf{D} = [\mathbf{D_x}, \mathbf{D_y}] \in \mathbb{R}^{N_s \times (N_\mathbf{x} + N_\mathbf{y})}$. Each row in the matrix $\mathbf{D}$ corresponds to the multimodal feature concatenation of a single subject. Joint ICA now assumes the following generative model:

$$\mathbf{D} = \mathbf{G} \cdot \mathbf{V}^\top = \sum_i^K \mathbf{g}^i \cdot {\mathbf{v}^i}^\top, \tag{29}$$

which states that the multimodal feature matrix $\mathbf{D}$ can be decomposed into the sum of $K = \min(N_s, (N_\mathbf{x} + N_\mathbf{y}))$ components. Each of the components is characterized by a multimodal feature profile $\mathbf{v}^i \in \mathbb{R}^{N_\mathbf{x}+N_\mathbf{y}}$ and vector $\mathbf{g}^i \in \mathbb{R}^{N_s}$, for $i \in \{1, ..., N_s\}$, that encodes how strong and with which sign the feature profile is present in each of the subjects. Assuming statistical independence between the feature profiles $\mathbf{v}^i$, a backward modeling approach can be applied to extract an estimate of these profiles by ICA algorithms discussed earlier.

The natural scaling of data from different modalities, i.e. Voltage in EEG vs percent signal change or concentration changes in fMRI or fNIRS, yield quite different histograms and may thus lead methods astray that rely on information-theoretic measures. This is the case for jICA. Additionally, an unequal number of samples between the two modalities leads to jICA giving more priority the modality for which more samples are provided. In order to ensure a balanced representation, up-/downsampling has to be applied.

While jICA assumes a common modulation profile within modalities for all subjects, this assumption is relaxed in an approach called *parallel ICA* (paraICA) [106], [107]. In this approach, a user specified similarity relation between components from the different modalities is optimized simultaneously with modality-specific un-mixing matrices. Thereby paraICA gives more emphasis to subject-specific multimodal components, compared to jICA.

Recently, a fully Bayesian approach to multimodal ICA is proposed in [108], in which the authors presented the so-called *linked ICA*. In contrast to jICA and paraICA, a difference in scaling or noise levels between modalities is not a problem for linked ICA.

Applications: Examples of multimodal fusion using jICA include fusion of EEG and functional MRI [105]. In this application, spatial independence was assumed for fMRI and temporal independence for the EEG ERP data. For fusing ERP components and fMRI activation maps, the multimodal feature maps were constructed as follows. The features in $\mathbf{D_x}$ were the time-course of an averaged event-related potential (ERP) from a single EEG channel, while the features in $\mathbf{D_y}$ were statistical parametric maps obtained from a GLM analysis. This study revealed a cascade of activations, along with their spatio-temporal dynamics, involved in processing of rare events among frequent distractors. While in [105] the data set comprised measurements from 23 healthy subjects, in [109] it was reported that the same analysis pipeline had been subsequently applied to 18 chronic schizophrenia patients. Findings of this analysis included a multimodal component that reliably distinguished between patients and healthy subjects.

A thorough analysis of how and why jICA works has been presented in [110]. In this study, a visual detection task was used to assess jICA performance in the fusion of EEG ERPs and stimulus induced activation maps derived from fMRI. One of the main results was the validation of, what the authors called, the central linking hypothesis. This hypothesis states that large parts of brain activity are visible in both imaging modalities and that a link between them can be established. Another recent study from the same group, presented in [111], used jICA to uncover novel insights into the dynamics of visual contour integration. In this study, jICA revealed spatiotemporal dynamics of the integration process

that would have been missed in a unimodal analysis.

B. CCA and PLS

For finding related components, a useful assumption is temporal co-modulation, which can be captured by finding those transformations for each modality that maximize the correlation between the time-courses of the extracted components. This is the idea of Canonical Correlation Analysis (CCA) [112]. In the simplest case CCA finds a one-dimensional subspace $\mathbf{w_x} \in \mathbb{R}^{M_x}$ and $\mathbf{w_y} \in \mathbb{R}^{M_y}$ for data from two modalities such that the *canonical correlation* of the modalities in that subspace is maximized:

$$\max_{\mathbf{w_x},\mathbf{w_y}} \text{Corr}\left(\mathbf{w_x}^\top \mathbf{x}(t), \mathbf{w_y}^\top \mathbf{y}(t)\right) . \tag{30}$$

The advantage of maximizing the correlation after the linear transformation $\mathbf{w_x}, \mathbf{w_y}$ is that the resulting correlation coefficient is invariant with respect to linear transformations of the data, hence *canonical*. See Fig. 3 for an illustration of the CCA idea. The generalization of the univariate canonical correlation coefficient finds K dimensional subspaces $\mathbf{W_x} \in \mathbb{R}^{M_\mathbf{x} \times K}$ and $\mathbf{W_y} \in \mathbb{R}^{M_\mathbf{y} \times K}$ such that the sum of the correlations is maximized [113]. In matrix notation this objective can be written as

$$\begin{aligned} &\max_{\mathbf{W_x},\mathbf{W_y}} \text{Trace}\left(\mathbf{W_x}^\top \mathbf{X}\mathbf{Y}^\top \mathbf{W_y}\right) \qquad (31)\\ \text{s.t.} \quad &\mathbf{W_x}^\top \mathbf{X}\mathbf{X}^\top \mathbf{W_x} = \mathbf{W_y}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{W_y} = \mathbf{I} \end{aligned}$$

The concept of canonical correlation is very similar to that of the *principal angles* [114] between the spaces spanned by the data matrices $\mathbf{X}$ and $\mathbf{Y}$. The objective of CCA in Eq. (31) can be transformed into the following generalized eigenvalue problem:

$$\begin{bmatrix} 0 & \mathbf{C_{xy}} \\ \mathbf{C_{yx}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{W_x} \\ \mathbf{W_y} \end{bmatrix} = \Lambda \begin{bmatrix} \mathbf{C_x} & 0 \\ 0 & \mathbf{C_y} \end{bmatrix} \begin{bmatrix} \mathbf{W_x} \\ \mathbf{W_y} \end{bmatrix}, \tag{32}$$

where $\mathbf{C_{xy}}, \mathbf{C_{yx}}, \mathbf{C_{xx}}, \mathbf{C_{yy}}$ are defined in Table I. If $\mathbf{C_x}$ and $\mathbf{C_y}$ are assumed to be the identity matrix, that is assuming that the features of $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated, respectively, Eq. 32 solves an optimization problem that is known as *partial least squares* (PLS) [115], [116], which has also found applications in multimodal data fusion [117]. The main difference between PLS and CCA is that CCA aims at finding maximally correlated components, while PLS aims at finding maximally covarying components. While this can be the same in some cases, in practice this is not necessarily so. The correct choice of method depends on what aspects of the data the analyst or experimenter wants to investigate.

Another way of solving the CCA objective is to learn a probabilistic model, as proposed in [118]. Extensions of these probabilistic models, as put forward in [119] also include a factorization of the part of the signal that CCA considers noise – this generalization of CCA is termed *inter-battery factor analysis* in the statistics literature [120]. See also [7] for different Baysian approach multimodal data fusion. CCA has been extended to handle more than two modalities at the same time. These, so-called, *N-way* or *multi-way* extensions of CCA have found application in multimodal neuroimaging as well [121], [122], [123], [124]. See also the review on CCA by [125]. The authors of [125] also discuss the differences between multiway CCA and jICA: jICA seeks to find independent components, which can be too strong an assumption in some cases. More importantly, unlike jICA, multiway CCA does not constrain the activation patterns of components to be the same for both modalities.

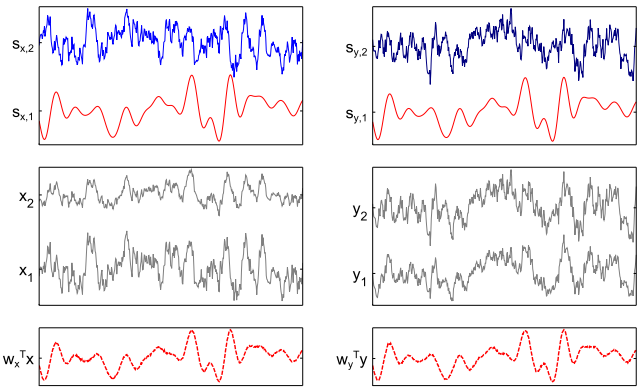

Fig. 3. Illustration of canonical correlation analysis (CCA) for multimodal fusion, exemplary for two datasets $\mathbf{x}$ and $\mathbf{y}$ (shown on the left and right panels, respectively). Two modality specific source spaces are assumed, each containing at least one source that is highly correlated with a corresponding source in the other modality. In this example, the time-courses of $s_{\mathbf{x},1}$ and $s_{\mathbf{y},1}$ are correlated, while $s_{\mathbf{x},2}$ and $s_{\mathbf{y},2}$ are uncorrelated to all other sources. The source signals are projected to the recording channels according to Eq. (1) with modality specific activation patterns, i.e. matrices $\mathbf{A_x}$ and $\mathbf{A_y}$, respectively. CCA optimizes spatial filters $\mathbf{w_x}$ and $\mathbf{w_y}$ such that the correlation between the projections $\mathbf{w_x}^\top \mathbf{x}(t)$ and $\mathbf{w_y}^\top \mathbf{y}(t)$ are maximized.

Note that CCA assumes that the samples of each modality are correlated instantaneously. For neuroimaging data this assumption does often not hold true. One solution is to embed one modality in its temporal context and optimize a time-lag-dependent projection $\mathbf{w_x}(\tau)$ for one modality, such that the canonical correlation is maximized:

$$\max_{\mathbf{w_x}(\tau),\mathbf{w_y}} \text{Corr}\left(\sum_i^{N_\tau} (\mathbf{w_x}(\tau_i)^\top \mathbf{x}(t-\tau_i), \mathbf{w_y}^\top \mathbf{y}(t)\right), \tag{33}$$

for a given set of N_τ time lags $\{\tau_1, ..., \tau_{N_\tau}\}$. The solution to Eq. (33) can be conveniently obtained as the solution to the standard CCA problem in Eq. (30) by applying the trick of *temporal embedding*. Temporal embedding is achieved by first creating N_τ copies of the dataset which is to be embedded (here $\mathbf{X}$), then shifting each copy by one of the specified time lags, and finally stacking the time-shifted copies along the spatial axis of the data matrix. Let the result of this embedding be denoted $\tilde{\mathbf{X}}$, then we have

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_{\tau_1} \\ \vdots \\ \mathbf{X}_{\tau_{N_\tau}} \end{bmatrix} \in \mathbb{R}^{M_\mathbf{x} \cdot N_\tau \times T_\mathbf{x}}, \tag{34}$$

where $\mathbf{X}_{\tau_i}$ denotes the copy of $\mathbf{X}$ that is shifted by time lag τ_i. The optimal N_τ can be found using standard model selection procedures such the *Akaike Information Criterion* which was introduced for this purpose in the context of CCA [126]. In practice however, it is sufficient to use the

well established knowledge about the neurovascular coupling dynamics to restrict the length of the temporal window to less than 20 seconds. Using the definition above we can substitute $\tilde{\mathbf{X}}$ into the original CCA objective in Eq. (30) from which we obtain a temporally embedded spatial filter $\mathbf{w}_\mathbf{x}(\tau)$.

The *spatio-temporal* filter $\mathbf{w}_\mathbf{x}(\tau)$ is recovered from $\tilde{\mathbf{w}}_\mathbf{x}$ as

$$\tilde{\mathbf{w}}_\mathbf{x} = \begin{bmatrix} \mathbf{w}_\mathbf{x}(\tau_1) \\ \vdots \\ \mathbf{w}_\mathbf{x}(\tau_{N_\tau}) \end{bmatrix} \in \mathbb{R}^{M_\mathbf{x} \cdot N_\tau \times 1}. \tag{35}$$

Unfortunately if there are only a limited number of samples available and at the same time the dimensionality of the data is large, then this temporal embedding will lead to ill-conditioned covariance matrices. However one can apply the kernel trick and solve the dual formulation of the problem instead (see Section VI). This approach is proposed as temporal kernel CCA in [127].

Application: Figure 4 shows an example of the patterns obtained from recordings of spontaneous neural activity in the anesthetized macaque monkey. Electrophysiological signals were obtained by intracranial electrodes and high-resolution fMRI data in a spherical region-of-interest around the recording electrode was measured simultaneously. Experimental details are described in [128]. Filters were estimated using temporal kernel CCA and patterns were obtained using Eq. (4). The structure of the fMRI pattern $\mathbf{a}_y$ reflects a smooth hemodynamic spatial response that is in line with the anatomical structure around the electrode: the coefficients along the cortical laminae are large and decay quickly perpendicular to the cortical laminae. Similarly, the coefficients of the neurovascular time-frequency response pattern $\mathbf{a}_\mathbf{x}(\tau)$ reflect clearly the well known physiology of the neurovascular response. The temporal profile shows a clear peak at 5 s and a later undershoot at about 15 s. The frequency profile indicates that the strongest hemodynamic response is in the high gamma frequency range.

Also in the context of EEG-fMRI recordings, CCA has become widely used. One of the major advantages of CCA is that it is straightforward to extend to analyses of more than one subject. This approach was taken e.g. in [129]. Here the authors investigate amplitude modulations of event-related potentials (ERPs) recorded with EEG simultaneously with fMRI during an auditory oddball paradigm. After preprocessing, the EEG data and the fMRI data were whitened and subjected to a multi-way CCA analysis that finds those components that maximally correlate between all pairs of subjects and between the two modalities. Similar approaches in the field of unimodal data analyses also make use of multi-way CCA in order to integrate data from multiple subjects, see e.g. [123], [124].

C. mSPoC

In section IV-B2 we have seen that the co-modulation between component power and a scalar target variable z can be modeled using the SPoC objective function. Here we extend this notion to the multimodal case by assuming that the target function z is the time-course of a component that is to be extracted from the other modality. Thus we set

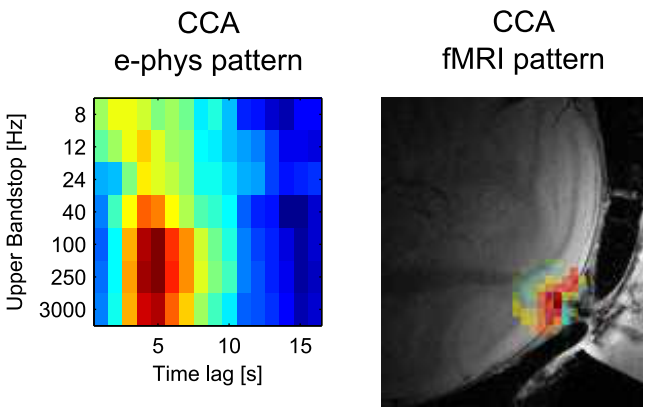

Fig. 4. Application example of CCA. Here an intracranially measured electrophysiology signal was fused with high resolution fMRI. A time-frequency representation was derived from the univariate intracranial electrode signal. This, now multivariate time-frequency signal, was temporally embedded and, together with the fMRI signal subjected to CCA analysis. Shown are the resulting activation patterns for the electrode on the left and the fMRI signal on the right. The fMRI activation pattern was superimposed on an anatomical scan. See main text for interpretation of the results.

$z(e) = \mathbf{w}_\mathbf{y}^\top \mathbf{y}(e)$ and formulate the objective function for the *multimodal source power co-modulation* analysis (mSPoC) [130] as

$$\max_{\mathbf{w}_\mathbf{x}, \mathbf{w}_\mathbf{y}} \operatorname{Cov}\left(\phi_{\mathbf{w}_\mathbf{x}}(e), \mathbf{w}_\mathbf{y}^\top \mathbf{y}(e)\right) \tag{36}$$
$$\text{s.t.} \quad \operatorname{Var}\left(\mathbf{w}_\mathbf{x}^\top \mathbf{x}(t)\right) = \operatorname{Var}\left(\mathbf{w}_\mathbf{y}^\top \mathbf{y}(e)\right) = 1,$$

where $\phi_{\mathbf{w}_\mathbf{x}}$ was defined in Eq. (23). Note that here we require $\mathbf{y}$ to be indexed by the epoch index e and thus have $\mathbf{Y} \in \mathbb{R}^{M_\mathbf{y} \times N_e}$, where N_e denotes the number of samples of the $\mathbf{y}$ modality that can be aligned to short epochs in the $\mathbf{x}$ modality. Figure 5 illustrates the ideas underlying mSPoC.

Unlike the SPoC or the CCA objective, the mSPoC objective does not lead directly to generalized eigenvalue problem. However, it turns out that the mSPoC objective can be broken down into sub-problems that each have straight forward solutions which have been discussed above.

To see this, let us assume that $\mathbf{w}_\mathbf{x}$ is already known. Then $\phi_{\mathbf{w}_\mathbf{x}}$ evaluates to a row vector and the mSPoC objective becomes the regression objective given in Eq. (21) and the weight vector $\mathbf{w}_\mathbf{y}$ is obtained through Eq. (20). Now let us assume that $\mathbf{w}_\mathbf{y}$ is known. In this case $\mathbf{w}_\mathbf{y}^\top \mathbf{y}(e)$ evaluates to a scalar function of epoch-index e and the mSPoC objective reduces to the previously discussed SPoC objective (see Eq. (24)) and can be solved by means of the corresponding generalized eigenvalue problem shown in Eq. (27). Thus a simple way to optimize the mSPoC objective is to randomly initialize $\mathbf{w}_\mathbf{x}$ and then iterate Eq. (20) and Eq. (27) until convergence [130].

In order to model non-instantaneous interaction between bandpower dynamics of a component in $\mathbf{x}$ and the time course of a component in $\mathbf{y}$ the mSPoC objective can be extended to include a convolution of the bandpower dynamics. This is done by introducing a finite impulse response (FIR) filter $\mathbf{w}_\tau \in \mathbb{R}^{N_\tau}$, the coefficients of which can either be set using prior knowledge or estimated from the data.

Using the trick of temporal embedding that we have seen

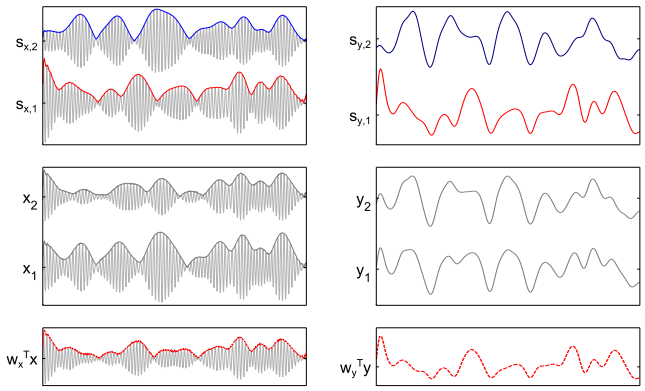

Fig. 5. Illustration of multimodal source power correlation (mSPoC) for multimodal fusion, exemplary for two datasets $\mathbf{x}$ and $\mathbf{y}$ (shown on the left and right panels, respectively). Two modality specific source spaces are assumed. One of the source spaces (here the $\mathbf{x}$-source space) contains at least one oscillatory source with variable amplitude dynamics that are correlated to the time-course of a corresponding source in the other source space. In this example, the amplitude modulations of $s_{\mathbf{x},1}$ are correlated to the time-course of $s_{\mathbf{y},1}$, while $s_{\mathbf{x},2}$ (and its amplitude dynamics) as well as $s_{\mathbf{y},2}$ are uncorrelated to all other sources. The source signals are projected to the recording channels according to Eq. (1) with modality specific activation patterns, i.e. matrices $\mathbf{A_x}$ and $\mathbf{A_y}$, respectively. mSPoC optimizes spatial filters $\mathbf{w_x}$ and $\mathbf{w_y}$ such that the correlation between the amplitude dynamics of $\mathbf{w}_\mathbf{x}^\top \mathbf{x}(t)$ and the time-course of $\mathbf{w}_\mathbf{y}^\top \mathbf{y}(t)$ are maximized.

in the discussion of CCA, we express the convolution as a the product of the vector $\mathbf{w}_\tau$ and the matrix that contains the temporally embedded $\phi_{\mathbf{w_x}}$, which we denote by $\tilde{\Phi}_{\mathbf{w_x}} \in \mathbb{R}^{N_\tau \times N_e}$. Then the objective function of a temporal mSPoC, expressed in matrix notation for $K = 1$, reads

$$\max_{\mathbf{w_x},\mathbf{w}_\tau,\mathbf{w_y}} \mathbf{w}_\tau^\top \tilde{\Phi}_{\mathbf{w_x}} \mathbf{Y}^\top \mathbf{w}_\mathbf{y}^\top \tag{37}$$

$$\text{s.t.} \quad \mathbf{w}_\mathbf{x}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w_x} = \mathbf{w}_\tau^\top \tilde{\Phi}_{\mathbf{w_x}} \tilde{\Phi}_{\mathbf{w_x}}^\top \mathbf{w}_\tau = \mathbf{w}_\mathbf{y}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{w_y} = 1.$$

Note that this modeling approach does *not* assume the dynamics of the hemodynamic activation to be identical for all positions in the brain, which would imply the existence of location-independent *canonical hemodynamic response function* (HRF). Instead, by optimizing $\mathbf{w}_\tau$ anew together with each component pair $\mathbf{w_x}$ and $\mathbf{w_y}$, the temporal mSPoC approach explicitly models a potentially space-varying and non-instantaneous coupling between EEG bandpower dynamics and fMRI activations. This is in line with the known variability of the HRF across space and subjects [131], [132], [128].

Similar to before, the weight vectors in this objective can be optimized by reducing the optimization problem above to sub-problems to which we have already seen the solution. If $\mathbf{w}_\tau$ is to be estimated from the data, the temporal mSPoC objective can be solved by starting with a randomly initialized $\mathbf{w_x}$ and then iterating the CCA objective and the SPoC objective until a suitable convergence criterion is met. In applications in which $\mathbf{w}_\tau$ is known, optimizing the mSPoC objective reduces to alternating SPoC and regression.

Applications: The utility of mSPoC has been demonstrated in [130] in the context of fusing EEG and fNIRS measurements. mSPoC was shown to outperform CCA in terms of obtained correlations between the modalities by extracting physiologically plausible components.

Here we further illustrate the application of mSPoC for the fusion of simultaneously recorded EEG and fMRI during transient hand movements of the right hand. For this purpose one subject was placed in an 3 T MRI scanner and instructed to squeeze a soft ball five consecutive times with a frequency of 1 to 2 Hz each time an auditory brief tone was presented. 31-channel EEG was simultaneously recorded. mSPoC was applied to investigate the co-modulation between induced power dynamics of the sensorimotor rhythm (here in the β band, i.e. 16 to 25 Hz) and BOLD signal changes. After mSPoC analysis, spatial activation patterns were computed for EEG and fMRI according to Eq. (4).

The brain region generating the EEG mSPoC component was localized using Eq. (7), that is by estimating a source-space equivalent $\mathbf{f_x}$ of the EEG sensor-space activation pattern $\mathbf{a_x}$ provided by the mSPoC algorithm. Thus, similar to the multiple signal classification (MUSIC) approach [58], we scanned through $R_\mathbf{x} = 74,661$ dipole locations on the tesselated cortical surface and measured, using the corresponding part of the lead field, to what extent a single dipole at each location can explain the activation pattern $\mathbf{a_x}$.

Figure 6 shows the activation pattern of the coupled EEG and fMRI component as estimated by mSPoC. The largest activation cluster covered the left sensorimotor cortex, with maximal activations in the premotor cortex, primary motor cortex, and primary somatosensory cortex. This is in line with previous studies showing that the strength of the sensorimotor rhythm is inversely correlated with activity in motor and somatosensory cortex [48]. The EEG activation pattern was best approximated by a dipole in the left primary motor cortex (right-most panel of Fig. 6). The EEG source was thus estimated to be less than 2 cm away from the corresponding fMRI activation in primary motor cortex and the best-fitting dipole almost perfectly explained the mSPoC activation pattern ($v = 98\,\%$ explained variance).

VI. Extensions

In this section we discuss concepts to extend the methods presented in the previous sections in order to incorporate non-linear interaction and to robustify them. Note that, due to space limitations, we only exemplified these concepts here, rather than working out the details of these extensions for all analysis approaches.

A. Nonlinear interactions

In many real-world settings, the interaction between observables (and non-observables) is of nonlinear nature. These relations are thus not adequately modeled by methods that assume linear relations, such as regression or CCA, for example. Here we discuss some approaches to deal with non-linear interactions.

1) Nonlinear mapping and explicit modeling of non-linear interaction: A simple but efficient trick is to map the data into a nonlinear feature space and apply the linear approach therein, assuming a linear relationship between the nonlinearly

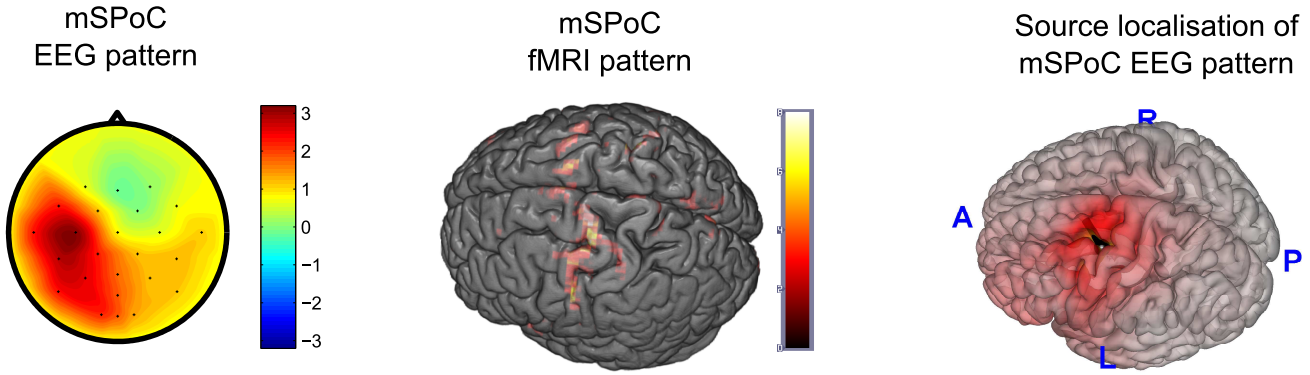

Fig. 6. Application example of mSPoC. Bandpower modulations of EEG were fused with simultaneously measured fMRI in a motor task (transient right hand movement). EEG was bandpass filtered to amplify oscillations in the β band (16 Hz to 25 Hz). The left panel shows the sensor space activation pattern of the EEG mSPoC component and the middle panel shows the fMRI activation pattern of the corresponding mSPoC component. These components were identified based on task-induced co-modulation of amplitude dynamics in the EEG and the time-course of BOLD dynamics in the fMRI. The right panel shows an estimate of the source-space pattern of the EEG component based on the the sensor space pattern, computed using the the MUSIC algorithm [58]. See main text for further discussion.

transformed variables. While this approach is often successful, it does come with certain caveats – in particular in light of the reviewed duality between filters and patterns and the source-space localization of sensor-space patterns.

As an example, let us re-visit the SPoC setting. In this setting we assume co-modulation between a target variable z and the bandpower of a component in the data. Note that computing bandpower is a nonlinear feature of the oscillatory signal. However, here it is wrong to first compute bandpower at each recording channel and then try to find a projection using linear regression. Let $\phi(\mathbf{X})$ denote the data matrix in which bandpower time-courses have been computed for each recording channel. If regression is applied to find a filter $\mathbf{w}$ such that $\mathbf{w}^\top\phi(\mathbf{X})$ maximally co-modulates with z, then the spatial pattern that corresponds to $\mathbf{w}$ *cannot* be source localized. Therefore, it is more appropriate here to explicitly model the nonlinearity in the objective function and compute power on the temporal signature of the to-be-extracted component, as is the case in SPoC and mSPoC.

We quantify this notion using a realistic simulation of EEG, in which we compare the two approaches, i.e. (i) regression on channelwise bandpower features (i.e. $z \approx \mathbf{w}^\top\phi(\mathbf{X})$) and (ii) source power co-modulation (i.e. $z \approx \phi(\mathbf{w}^\top\mathbf{X})$). The simulated EEG is generated according to the generative model in Eq. (1). The target function z is chosen to be the bandpower modulation of one of the simulated sources. Further details of the simulation can be found in [99] and [133]. The results are depicted in Fig. 7 as a function of sensor-space signal-to-noise ratio. The SPoC approach yields better approximation of the target function. More importantly however, the sensor-space pattern of the SPoC component is more similar to the true target component than is the case for the "power pattern" obtained for regression. This is reflected by a better dipole fit and less source reconstruction error for SPoC patterns.

Thus, while nonlinear mappings can be used to "linearize" nonlinear relations, care has to be taken with respect to interpretation of the model parameters. Explicit modeling of nonlinearity, if feasible, may preserve interpretability.

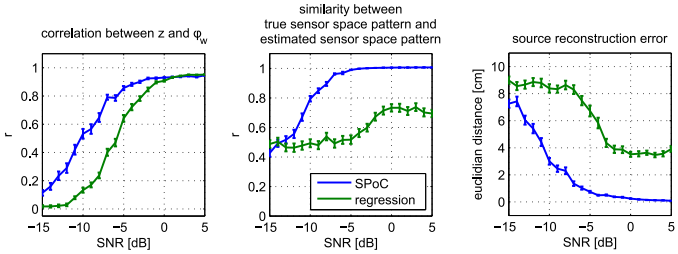

Fig. 7. A simulation that illustrates potential problems with parameter interpretation if the assumptions of the generative model are not respected. In this EEG simulation the task was to extract a bandpower signal that co-modulates with a given target function z. The target signal corresponds to the true amplitude modulation of one of the simulated source components. Regression was applied to channelwise computed bandpower $\phi(\mathbf{X})$ and estimated z as $\phi_{\mathbf{w}} = \mathbf{w}^\top\phi(\mathbf{X})$. SPoC was applied to the sensor signal and estimated z as $\phi_{\mathbf{w}} = \phi(\mathbf{w}^\top\mathbf{X})$. Sensor-space patterns were obtained for both methods using Eq. (4). Note that $\phi(\mathbf{w}^\top\mathbf{X}) \neq \mathbf{w}^\top\phi(\mathbf{X})$ because the computation of bandpower is a nonlinear operation. The left panel shows that SPoC yields better estimation of the target variable, compared to regression. More importantly, the resulting sensor space patterns show high similarity with the pattern of the true source only in the case of SPoC (middle panel). Source localization of the sensor-space patterns using dipole fitting reveals that SPoC patterns can be well explained by dipoles that are close to the location of the true simulated dipole (right panel). The simulation were repeated with new data 100 times for each signal-to-noise ratio (SNR). The results shown were obtained on test data that was not used to train (i.e. to optimize the parameters of) the algorithms. Lines correspond to means over repetitions, errorbars to $10 \cdot$ SE.

2) The kernel trick: Another approach to extend the presented methods to nonlinear domains is based on the so-called *kernel trick* [134], [135]. The essence of this trick is to implicitly map the variables into a higher (possibly infinite) dimensional feature space $\mathcal{F}$ and to apply the linear machinery there. Practically, this can be achieved by substituting the linear inner product in the original formulation of the algorithm by kernel functions $k(\cdot,\cdot)$ which represent inner products in feature space

$$k(\mathbf{x},\mathbf{y}) = \langle\xi(\mathbf{x}),\xi(\mathbf{y})\rangle_{\mathcal{F}} \tag{38}$$

Thus the resulting algorithm can be interpreted as running the original algorithm on the (nonlinearly) mapped objects

$\xi(\mathbf{x})$ and $\xi(\mathbf{y})$. The choice of the kernel largely influences the algorithm's ability to model particular types of nonlinearity. A popular kernel which works very well in practice is the Gaussian kernel $k(\mathbf{x},\mathbf{y}) = e^{-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}}$.

Many algorithms have been "kernelized" including PCA [135], CCA [127], FDA [136], [137] and ICA [138]. For kernel CCA one can show that the objective is of the same type as the one in Eq. (32), but the covariance matrices are substituted by kernel matrices which implicitly model the correlation between variables in feature space. Mathematically, the kernel CCA filters are obtained by solving the following generalized eigenvalue problem

$$\begin{bmatrix} 0 & \mathbf{K_{xy}} \\ \mathbf{K_{yx}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{W_x} \\ \mathbf{W_y} \end{bmatrix} = \Lambda \begin{bmatrix} \mathbf{K_x} & 0 \\ 0 & \mathbf{K_y} \end{bmatrix} \begin{bmatrix} \mathbf{W_x} \\ \mathbf{W_y} \end{bmatrix}, \quad (39)$$

where $k(\mathbf{x}_i,\mathbf{x}_j)$ is the ij-th element of $\mathbf{K_x}$ and the other kernel matrices are defined analogously.

3) Higher moments: A third avenue towards nonlinearity is based on the analysis of higher moments. Since many factor models we discussed use second-order moments to measure relationships between variables, they implicitly assume Gaussianity and linear mappings. By considering higher moments, however, this restriction is relaxed. One popular higher moment based measure of dependency is the Mutual Information (MI), introduced in Eq. (17).

With this measure the correlation in objective function of (nonlinear) CCA can be replaced by MI and formulated as

$$\max_{\mathbf{w_x},\mathbf{w_y}} D_{KL}(p(\mathbf{w_x}^\top\mathbf{X},\mathbf{w_y}^\top\mathbf{Y}) \,||\, p(\mathbf{w_x}^\top\mathbf{X})p(\mathbf{w_y}^\top\mathbf{Y})) \quad (40)$$

The authors of [139] showed that when p is the Gaussian distribution, then the divergence formulation in Eq. (40) reduces to the CCA problem introduced in Section V-B. In the general case the Mutual Information based version of the algorithm considers nonlinear dependencies and can not be solved as generalized eigenvalue problem. The authors of [139] proposed an algorithm based on kernel density estimation for solving this type of optimization problems.

B. Robustifying

Finally we address the issue of robustifying the presented approaches against the tendency to overfit and the adverse impact of outliers. Since factor models such as the ones discussed in this paper maximize an objective on a dataset with finite (sometimes very small) sample size, they do not necessary find a solution which works well in *general* but a solution which is optimal (with respect to some possibly non-robust error measure) on the particular dataset. This poses a severe problem especially when analyzing neuroimaging data because generalization is a key property of neurophysiologically meaningful solutions. The lack of generalization is termed *overfitting* and may have two reasons.

First, overfitting occurs when the complexity of the solution is too high relative to the sample size. In other words, there is not enough data to reliably fit the complex model. One way to avoid the overfitting problem in this case is to restrict the complexity of the solution [140], e.g., by adding a *regularization term* to the objective function of the algorithm [141]. One popular choice, the Tikhonov regularization term [142], penalizes the complexity of the solution. Tikhonov regularization is often called L_2-norm regularization, which refers to the mechanism used to stabilize the algorithm. The key idea is to impose a penalty on the euclidean norm (i.e. the L_2-norm) of the subspace to be found. This penalty can be easily incorporated into the standard formulations of most algorithms discussed above by adding a ridge to the diagonal of the covariance matrices, hence the regression case of Tikhonov regularization is often called *ridge regression*. The solution obtained by Eq. 20 becomes in the case of Tikhonov regularization

$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \mathbf{I}\lambda_\mathbf{x}\right)^{-1}\mathbf{X}\mathbf{z}^\top, \quad (41)$$

where λ controls the amount of regularization, the higher, the smaller the norm of $\mathbf{w}$ will be. Effectively this smaller norm constraint will lead to more similar and smaller coefficients of $\mathbf{w}$. Taking a probabilistic perspective on the regularized least squares regression setting, it is easy to derive, that the amount of regularization translates directly into the noise assumed to be present in the data, see e.g. [93]. Analogously in the case of CCA, the L_2 regularized version of CCA results in a generalized eigenvalue equation just like in Eq. 32, with the slight modification that a ridge of height $\lambda_\mathbf{x},\ \lambda_\mathbf{y}$ is added to the covariance matrices on the right hand side of the equation, such that

$$\begin{bmatrix} 0 & \mathbf{C_{xy}} \\ \mathbf{C_{yx}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix} = \Lambda \begin{bmatrix} \mathbf{C_x} + \mathbf{I}\lambda_\mathbf{x} & 0 \\ 0 & \mathbf{C_y} + \mathbf{I}\lambda_\mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix}, \quad (42)$$

where $\lambda_\mathbf{x},\ \lambda_\mathbf{y}$ are the regularizers for each modality, respectively. For an introduction to the relationship between standard CCA and regularized CCA see [143]. Similarly to the case of regression, also in regularized CCA the regularization constants are proportional to the amount of noise assumed in the data, see for instance [118].

Next to these simple cases of euclidean norm constraint regularizations, there is a spectrum of other approaches to regularize the solutions of factor models. Many approaches impose a mixture of L_2 and L_1 norms, this is often referred to as *elastic net* regularization. Other methods penalize the L_1 norm of the factor subspace or on the sources themselves. This approach is popular in the dictionary learning community, see e.g. [144], [145]. More sophisticated regularization schemes impose structured sparsity constraints on the solution [146].

Other regularization strategies minimize nonstationarity [147], apply shrinkage [55], early stopping [148], weight decay [149] or asymptotic model selection criteria [150] have also been successfully used in the past.

Another reason for overfitting are outliers. If an analysis model is parametric, i.e., assumes a particular distribution of the variables and requires parameter estimation, the presence of outliers deviating from the assumptions may heavily bias the solution. Robust parameter estimators such as M-estimators [151] can minimize the impact of outliers and largely improve

the quality of the solution. For algorithms that can be formulated as divergence maximization problems there exist also an alternative way of reducing the impact of outliers, namely the usage of robust divergence. For instance, the objective function presented in Eq. (40) can be significantly robustified when using Beta divergence [152] instead of KL-divergence. This robustification strategy has been applied to CCA [139], [153] and to other algorithms such as Common Spatial Patterns [154], [155], and Independent Component Analysis [156].

VII. Discussion and Conclusion

late fusion scenarios

features of interest	amplitude modulations of oscillatory sources	ERPs, hemodynamics
suggested methods	SPoC, CSP, ICA, PCA	Regression, GLM, LDA, ICA, PCA

early fusion scenarios

features of interest	amplitude modulations of oscillatory sources + hemodynamics	ERPs + hemodynamics
suggested methods	mSPoC	CCA/PLS, jICA

TABLE II
SUGGESTIONS FOR WHEN TO CHOOSE WHICH OF THE METHODS WE DISCUSSED.

Multimodal data contains a wealth of information that reflects different aspects of underlying physical processes. While it may appear very promising to capture multiple characteristics of such a process, a number of challenges have to be addressed to make practical use of the multimodal data sources before we can finally fuse them to obtain more accurate results and better insight. Data sources to be fused will inevitably contain different signal-to-noise characteristics, a different percentage of outliers, the spatial and temporal sampling as well as the dimensionality may disagree, moreover, the underlying physics may give rise to a high variability in how the modalities might be coupling (e.g. linear vs. nonlinear) and finally, multimodal analysis tools may in practice only be useful if they can be interpreted and thus allow better understanding. We have contributed here by placing these generic challenges of multimodal data analysis into the context of neuroimaging and reviewed a set of tools that we consider of practical use.

The following enumeration summarizes how these challenges are addressed by the methods presented in this paper.

Spatio-temporal sampling: Different spatial and temporal resolution can be addressed by computing PCA or ICA along either the spatial or the temporal dimension, depending on which dimension has a more favorable ratio of number of input dimensions vs number of samples. This way, dimensionality can be reduced to a set of components while preserving relevant information and inference can be conducted in the component space (sections IV and V).

Non-instantaneous coupling: Temporal embedding, as was shown for tkCCA [127], can model non-instantaneous interactions. Alternatively, time-lagged interactions can be modeled explicitly by including a convolution operator in the model as is the case in mSPoC [130] (sections V-B and V-C).

Nonlinear coupling: The kernel trick (cf. [157], [135], [134]) can be applied in order to model nonlinear interactions, thereby mapping input variables into a feature space in which the interaction is more linear. Alternatively, a similarity relation can be employed that does not assume linear relations. One example is mutual information (section VI).

Signal-to-noise and robustness: Robust estimates of model parameters can be achieved by regularization [158], [159] or, alternatively, by using robust divergences such as beta divergences [154], [155]. All backward models discussed here are inherently multivariate, which means they integrate information from all recording channels by means of filters and thus yield higher SNRs than univariate approaches (section VI).

Interpretation and source localization: For filters obtained from backward models, corresponding activation patterns can be obtained by virtue of Eq. (4) [95], [55], [54]. This makes their parameters interpretable in the context of generative forward models. When applying multimodal analysis it should be emphasized that the underlying generative models should be respected. That is, when studying correlations or nonlinear couplings these should be computed in source space, as reviewed in the context of SPoC/mSpoc (see section IV-B2) [99]. Failure to respect the underlying generative models may lead to systematic estimation errors, loss of robustness and also to inaccuracies in localization of the results of multimodal analysis (sections III-D, III-E, and VI as well as Fig. 7) .

Note that throughout this paper, we have used the term multimodal in the context of multiple measurement modalities. However, the term could also be interpreted in a wider sense, namely that the data to be fused may come from different sources, irrespective of the physical measurement modality. For example, multimodal models may be used to combine information from different subjects or experiments. The investigation of inter-subject-correlations (ISC) [160], [161], [162], [163] or hyperscanning [164], [165] in the context of social neuroscience are actively researched fields. Here, the application of multivariate fusion techniques such as the ones reviewed in this paper, have the potential to increase understanding of the involved cognitive processes, as for instance in [124], where the authors used CCA to extend the concept of ISCs to that of *canonical-ISCs* in order to investigate the neural underpinnings of 2D vs 3D perception. Similarly, an extension of the SPoC technique presented in [133] allows to asses inter-subject-envelope-correlations of neural oscillations.

We would like to point out that the methods reviewed here can (and perhaps should) be combined for the synthesis of new and improved analysis approaches. mSPoC [99] can be considered an example for this, because it combines the ideas of SPoC and CCA. Further examples are the combination of CCA with jICA presented in [166] or the combination of jICA with PCA presented in [167]. This shows that the fusion of analysis approaches can be just as fruitful as the fusion of

multiple data modalities.

Given the multitude of analysis approaches available, the most relevant question for the practitioner is of course which method to choose. Unfortunately there is no unique answer to this, because it depends on (i) the preferred analysis scenario (e.g. late fusion vs early fusion), (ii) the assumptions being made about the data (e.g. what type of coupling between modalities), (iii) the features of interest in the analysis (e.g. spectral features or time-domain features), (iv) what additional information is available (e.g. condition/class labels), and other aspects. However, in order to narrow down the possible choices we present a systematic overview over the methods presented in this paper in table II.

While this work has reviewed a number of generic tools for multimodal data analysis in neuroimaging, the authors firmly believe that these analysis techniques are applicable beyond the realm of neuroscience, where similarly structured challenges are known to occur. For example in social media analysis vast communication statistics are being recorded, here, tkCCA has allowed to fuse geostatistical information and tweet patterns to quantify the information spread of news [168].

A number of open problems remain: If the data is nonstationary (cf. [169], [147]), how can we extract similar types of stationary or non-stationary processes from multimodal data sources? How can symbolic modalities be combined with other continuous measurements, in particular in the context of non-sampling errors? How can we perform causal inference across modalities? And, finally – here transfer learning and multimodal data analysis become very related – how can multiple trained models be of use for enhancing the statistical power of multimodal data.

REFERENCES

[1] T. Eichele, K. Specht, M. Moosmann, M. L. A. Jongsma, R. Q. Quiroga, H. Nordby, and K. Hugdahl, "Assessing the spatiotemporal evolution of neuronal activation with single-trial event-related potentials and functional mri," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 17798–803, Dec 2005.

[2] S. Debener, M. Ullsperger, M. Siegel, and A. K. Engel, "Single-trial EEG-fMRI reveals the dynamics of cognitive function," *Trends in Cognitive Sciences*, vol. 10, pp. 558–63, Dec 2006.

[3] A. M. Dale and E. Halgren, "Spatiotemporal mapping of brain activity by integration of multiple imaging modalities," *Current Opinion in Neurobiology*, vol. 11, pp. 202–8, Apr 2001.

[4] S. Vulliemoz, D. W. Carmichael, K. Rosenkranz, B. Diehl, R. Rodionov, M. C. Walker, A. W. McEvoy, and L. Lemieux, "Simultaneous intracranial EEG and fMRI of interictal epileptic discharges in humans," *NeuroImage*, vol. 54, pp. 182–190, Jan 2011.

[5] J. Ives, S. Warach, F. Schmitt, R. Edelman, and D. Schomer, "Monitoring the patient's EEG during echo planar MRI," *Electroencephalography and Clinical Neurophysiology*, vol. 87, no. 6, pp. 417–420, 1993.

[6] L. Lemieux, P. J. Allen, F. Franconi, M. R. Symms, and D. K. Fish, "Recording of EEG during fMRI experiments: patient safety," *Magnetic Resonance in Medicine*, vol. 38, no. 6, pp. 943–952, 1997.

[7] J. Daunizeau, C. Grova, G. Marrelec, J. Mattout, S. Jbabdi, M. Pélégrini-Issac, J.-M. Lina, and H. Benali, "Symmetrical event-related EEG/fMRI information fusion in a variational Bayesian framework," *NeuroImage*, vol. 36, no. 1, pp. 69–87, 2007.

[8] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[9] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents-EEG, ECoG, LFP and spikes," *Nature Reviews Neuroscience*, vol. 13, no. 6, pp. 407–420, 2012.

[10] D. Attwell and C. Iadecola, "The neural basis of functional brain imaging signals," *Trends in Neurosciences*, vol. 25, no. 12, pp. 621–625, 2002.

[11] O. J. Arthurs and S. Boniface, "How well do we understand the neural origins of the fMRI BOLD signal?," *Trends in Neurosciences*, vol. 25, no. 1, pp. 27–31, 2002.

[12] F. Bießmann, S. M. Plis, F. C. Meinecke, T. Eichele, and K.-R. Müller, "Analysis of multimodal neuroimaging data," *Biomedical Engineering, IEEE Reviews in*, vol. 4, pp. 26–58, 2011.

[13] R. J. Huster, S. Debener, T. Eichele, and C. S. Herrmann, "Methods for simultaneous EEG-fMRI: an introductory review," *The Journal of Neuroscience*, vol. 32, no. 18, pp. 6053–6060, 2012.

[14] M. Scanziani and M. Häusser, "Electrophysiology in the age of light," *Nature*, vol. 461, no. 7266, pp. 930–939, 2009.

[15] A. L. Hodgkin and A. F. Huxley, "Action potentials recorded from inside a nerve fibre," *Nature*, vol. 144, no. 3651, pp. 710–711, 1939.

[16] A. R. Wyler, G. A. Ojemann, E. Lettich, and A. A. Ward Jr, "Subdural strip electrodes for localizing epileptogenic foci," *Journal of Neurosurgery*, vol. 60, no. 6, pp. 1195–1200, 1984.

[17] H. Berger, "Über das Elektroenkephalogramm des Menschen," *Archiv fur Psychiatrie und Nervenkrankheiten*, vol. 87, pp. 527–570, 1929.

[18] D. Cohen, "Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents," *Science*, vol. 161, no. 3843, pp. 784–786, 1968.

[19] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, pp. 1926–1929, June 2004.

[20] M. Denker, S. Roux, H. Lindén, M. Diesmann, A. Riehle, and S. Grün, "The local field potential reflects surplus spike synchrony," *Cerebral Cortex*, vol. 21, no. 12, pp. 2681–2695, 2011.

[21] G. Pfurtscheller and F. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[22] S. Makeig and T. P. Jung, "Tonic, phasic, and transient EEG correlates of auditory awareness in drowsiness.," *Cognitive Brain Research*, vol. 4, no. 1, pp. 15–25, 1996.

[23] G. Thut, A. Nietzel, S. A. Brandt, and A. Pascual-Leone, "Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection.," *The Journal of Neuroscience*, vol. 26, no. 37, pp. 9494–9502, 2006.

[24] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis.," *Brain Research Reviews*, vol. 29, no. 2-3, pp. 169–195, 1999.

[25] O. Jensen, J. Kaiser, and J.-P. Lachaux, "Human gamma-frequency oscillations associated with attention and memory.," *Trends in Neurosciences*, vol. 30, no. 7, pp. 317–324, 2007.

[26] D. Regan, *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*. New York: Elsevier, 1989.

[27] S. J. Luck, *An introduction to the event-related potential technique*. MIT press, 2014.

[28] A. Grinvald, E. Lieke, R. D. Frostig, C. D. Gilbert, and T. N. Wiesel, "Functional architecture of cortex revealed by optical imaging of intrinsic signals," *Nature*, vol. 324, pp. 361–364, 1986.

[29] K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, and R. Turner, "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 12, pp. 5675–5679, 1992.

[30] S. Ogawa, T.-M. Lee, A. S. Nayak, and P. Glynn, "Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields," *Magnetic Resonance in Medicine*, vol. 14, no. 1, pp. 68–78, 1990.

[31] F. F. Jobsis, "Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters," *Science*, vol. 198, no. 4323, pp. 1264–1267, 1977.

[32] R. B. Buxton, K. Uludağ, D. J. Dubowitz, and T. T. Liu, "Modeling the hemodynamic response to brain activation," *NeuroImage*, vol. 23, pp. S220–S233, 2004.

[33] M. Thorniley, L. Livera, Y. Wickramasinghe, S. Spencer, and P. Rolfe, "The non-invasive monitoring of cerebral tissue oxygenation," *Advances in Experimental Medicine and Biology*, vol. 277, p. 323, 1990.

[34] S. Wray, M. Cope, D. T. Delpy, J. S. Wyatt, and E. O. R. Reynolds, "Characterization of the near infrared absorption spectra of cytochrome aa_3 and haemoglobin for the non-invasive monitoring of cerebral oxygenation," *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, vol. 933, no. 1, pp. 184–192, 1988.

[35] E. Hamel, "Perivascular nerves and the regulation of cerebrovascular tone," *Journal of Applied Physiology*, vol. 100, no. 3, pp. 1059–1064, 2006.

[36] N. K. Logothetis and B. A. Wandell, "Interpreting the BOLD signal," *Annual Review of Physiology*, vol. 66, pp. 735–769, 2004.

[37] D. J. Heeger and D. Ress, "What does fMRI tell us about neuronal activity?," *Nature Reviews Neuroscience*, vol. 3, no. 2, pp. 142–151, 2002.

[38] N. K. Logothetis, "What we can do and what we cannot do with fMRI," *Nature*, vol. 453, no. 7197, pp. 869–878, 2008.

[39] C. N. Hall, C. Reynell, B. Gesslein, N. B. Hamilton, A. Mishra, B. A. Sutherland, F. M. OFarrell, A. M. Buchan, M. Lauritzen, and D. Attwell, "Capillary pericytes regulate cerebral blood flow in health and disease," *Nature*, vol. 508, no. 7494, pp. 55–60, 2014.

[40] J. Berwick, D. Johnston, M. Jones, J. Martindale, C. Martin, A. Kennerley, P. Redgrave, and J. Mayhew, "Fine detail of neurovascular coupling revealed by spatiotemporal analysis of the hemodynamic response to single whisker stimulation in rat barrel cortex," *Journal of Neurophysiology*, vol. 99, no. 2, pp. 787–798, 2008.

[41] G. Bonvento, N. Sibson, and L. Pellerin, "Does glutamate image your thoughts?," *Trends in Neurosciences*, vol. 25, no. 7, pp. 359–364, 2002.

[42] A. Devor, I. Ulbert, A. K. Dunn, S. N. Narayanan, S. R. Jones, M. L. Andermann, D. A. Boas, and A. M. Dale, "Coupling of the cortical hemodynamic response to cortical and thalamic neuronal activity," *Proceedings of the National Academy of Sciences*, vol. 102, no. 10, pp. 3822–3827, 2005.

[43] J. Goense and N. K. Logothetis, "Neurophysiology of the BOLD fMRI signal in awake monkeys," *Current Biology*, vol. 18, no. 9, pp. 631–640, 2008.

[44] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, no. 6843, pp. 150–157, 2001.

[45] J. Martindale, J. Mayhew, J. Berwick, M. Jones, C. Martin, D. Johnston, P. Redgrave, and Y. Zheng, "The hemodynamic impulse response to a single neural event," *Journal of Cerebral Blood Flow & Metabolism*, vol. 23, no. 5, pp. 546–555, 2003.

[46] Y. B. Sirotin, E. M. Hillman, C. Bordier, and A. Das, "Spatiotemporal precision and hemodynamic mechanism of optical point spreads in alert primates," *Proceedings of the National Academy of Sciences*, vol. 106, no. 43, pp. 18390–18395, 2009.

[47] J. Niessing, B. Ebisch, K. E. Schmidt, M. Niessing, W. Singer, and R. A. Galuske, "Hemodynamic signals correlate tightly with synchronized gamma oscillations," *Science*, vol. 309, no. 5736, pp. 948–951, 2005.

[48] P. Ritter, M. Moosmann, and A. Villringer, "Rolandic alpha and beta EEG rhythms' strengths are inversely related to fMRI-BOLD signal in primary somatosensory and motor cortex," *Human Brain Mapping*, vol. 30, no. 4, pp. 1168–1187, 2009.

[49] S. Fazli, J. Mehnert, J. Steinbrink, G. Curio, A. Villringer, K.-R. Müller, and B. Blankertz, "Enhanced performance by a Hybrid NIRS-EEG Brain Computer Interface," *NeuroImage*, vol. 59, no. 1, pp. 519–529, 2012. Open Access.

[50] M. Moosmann, P. Ritter, I. Krastel, A. Brink, S. Thees, F. Blankenburg, B. Taskin, H. Obrig, and A. Villringer, "Correlates of alpha rhythm in functional magnetic resonance imaging and near infrared spectroscopy," *NeuroImage*, vol. 20, no. 1, pp. 145–158, 2003.

[51] H. Laufs, A. Kleinschmidt, A. Beyerle, E. Eger, A. Salek-Haddadi, C. Preibisch, and K. Krakow, "EEG-correlated fMRI of human alpha activity," *NeuroImage*, vol. 19, no. 4, pp. 1463–1476, 2003.

[52] D. Mantini, M. G. Perrucci, C. Del Gratta, G. L. Romani, and M. Corbetta, "Electrophysiological signatures of resting state networks in the human brain," *Proceedings of the National Academy of Sciences*, vol. 104, no. 32, pp. 13170–13175, 2007.

[53] H. Laufs, J. Daunizeau, D. Carmichael, and A. Kleinschmidt, "Recent advances in recording electrophysiological data simultaneously with magnetic resonance imaging," *NeuroImage*, vol. 40, no. 2, pp. 515–528, 2008.

[54] 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.

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

[56] S. Baillet, J. Mosher, and R. Leahy, "Electromagnetic brain mapping," *IEEE Signal Processing Magazine*, vol. 18, pp. 14–30, 2001.

[57] C. Habermehl, J. Steinbrink, K.-R. Müller, and S. Haufe, "Optimizing the regularization for image reconstruction of cerebral diffuse optical tomography," *Journal of Biomedical Optics*, vol. 19, no. 9, p. 096006, 2014.

[58] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 43, no. 3, pp. 276–280, 1986.

[59] M. S. Hämäläinen and R. J. Ilmoniemi, "Interpreting magnetic fields of the brain: minimum norm estimates," *Medical & Biological Engineering & Computing*, vol. 32, pp. 35–42, 1994.

[60] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain," *International Journal of Psychophysiology*, vol. 18, pp. 49–65, 1994.

[61] B. D. Van Veen, W. Van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 9, pp. 867–880, 1997.

[62] J. C. Mosher and R. M. Leahy, "Source localization using recursively applied and projected (RAP) MUSIC," *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 332–340, 1999.

[63] J. Gross, J. Kujala, M. Hmlinen, L. Timmermann, A. Schnitzler, and R. Salmelin, "Dynamic imaging of coherent sources: Studying neural interactions in the human brain," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 694–699, 2001.

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

[65] L. Ding and B. He, "Sparse source imaging in EEG with accurate field modeling," *Human Brain Mapping*, vol. 29, no. 9, pp. 1053–1067, 2008.

[66] S. Haufe, V. V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte, "Estimating vector fields using sparse basis field expansions," in *Advances in Neural Information Processing Systems 21* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 617–624, Cambridge, MA: MIT Press, 2008.

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

[68] A. Gramfort, D. Strohmeier, J. Haueisen, M. Hämäläinen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations," *NeuroImage*, vol. 70, no. 0, pp. 410–422, 2013.

[69] R. I. Goldman, J. M. Stern, J. Engel, and M. S. Cohen, "Simultaneous EEG and fMRI of the alpha rhythm," *Neuroreport*, vol. 13, pp. 2487–92, Dec 2002.

[70] J. G. Francis, "The QR transformation a unitary analogue to the lr transformationpart 1," *The Computer Journal*, vol. 4, no. 3, pp. 265–271, 1961.

[71] V. N. Kublanovskaya, "On some algorithms for the solution of the complete eigenvalue problem," *USSR Computational Mathematics and Mathematical Physics*, vol. 1, no. 3, pp. 637–657, 1962.

[72] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[73] I. T. Jolliffe, "A note on the use of principal components in regression," *Applied Statistics*, pp. 300–303, 1982.

[74] N. Dehghani, S. S. Cash, A. O. Rossetti, C. C. Chen, and E. Halgren, "Magnetoencephalography demonstrates multiple asynchronous generators during human sleep spindles," *Journal of Neurophysiology*, vol. 104, no. 1, pp. 179–188, 2010.

[75] M. Negishi, M. Abildgaard, T. Nixon, and R. Todd Constable, "Removal of time-varying gradient artifacts from EEG data acquired during continuous fmri," *Clinical Neurophysiology*, vol. 115, no. 9, pp. 2181–2192, 2004.

[76] R. Niazy, C. Beckmann, G. Iannetti, J. Brady, and S. Smith, "Removal of fMRI environment artifacts from EEG data using optimal basis sets," *NeuroImage*, vol. 28, no. 3, pp. 720–737, 2005.

[77] C.-G. Bénar, Y. Aghakhani, Y. Wang, A. Izenberg, A. Al-Asmi, F. Dubeau, and J. Gotman, "Quality of EEG in simultaneous EEG-fMRI for epilepsy," *Clinical Neurophysiology*, vol. 114, no. 3, pp. 569–580, 2003.

[78] A. de Cheveigné and L. C. Parra, "Joint decorrelation, a versatile tool for multichannel data analysis," *NeuroImage*, 2014.

[79] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[80] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings F (Radar and Signal Processing)*, vol. 140, no. 6, pp. 362–370, 1993.

[81] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[82] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

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

[84] V. D. Calhoun, J. Liu, and T. Adalı, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and erp data," *NeuroImage*, vol. 45, no. 1, pp. S163–S172, 2009.

[85] A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio, "Artifact reduction in magnetoneurography based on time-delayed second-order correlations," *IEEE Transactions on Biomedical Engineering*, vol. 47, pp. 75–87, January 2000.

[86] 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, no. 1, p. 30, 2011.

[87] D. Mantini, M. G. Perrucci, S. Cugini, A. Ferretti, G. L. Romani, and C. Del Gratta, "Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis," *NeuroImage*, vol. 34, no. 2, pp. 598–607, 2007.

[88] S. Debener, K. J. Mullinger, R. K. Niazy, and R. W. Bowtell, "Properties of the ballistocardiogram artefact as revealed by EEG recordings at 1.5, 3 and 7 T static magnetic field strength," *International Journal of Psychophysiology*, vol. 67, no. 3, pp. 189–199, 2008.

[89] Z. Liu, J. A. de Zwart, P. van Gelderen, L.-W. Kuo, and J. H. Duyn, "Statistical feature extraction for artifact removal from concurrent fMRI-EEG recordings," *NeuroImage*, vol. 59, no. 3, pp. 2073–2087, 2012.

[90] T. Eichele, V. D. Calhoun, M. Moosmann, K. Specht, M. L. Jongsma, R. Q. Quiroga, H. Nordby, and K. Hugdahl, "Unmixing concurrent EEG-fMRI with parallel independent component analysis," *International Journal of Psychophysiology*, vol. 67, no. 3, pp. 222–234, 2008.

[91] L. Dong, D. Gong, P. A. Valdes-Sosa, Y. Xia, C. Luo, P. Xu, and D. Yao, "Simultaneous EEG-fMRI: Trial level spatio-temporal fusion for hierarchically reliable information discovery," *NeuroImage*, vol. 99, pp. 28–41, 2014.

[92] T. Eichele, V. D. Calhoun, and S. Debener, "Mining EEG–fMRI using independent component analysis," *International Journal of Psychophysiology*, vol. 73, no. 1, pp. 53–61, 2009.

[93] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[94] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[95] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda, "Recipes for the linear analysis of EEG," *NeuroImage*, vol. 28, no. 2, pp. 326–341, 2005.

[96] M. Misaki, Y. Kim, P. A. Bandettini, and N. Kriegeskorte, "Comparison of multivariate classifiers and response normalizations for pattern-information fMRI," *NeuroImage*, vol. 53, no. 1, pp. 103–118, 2010.

[97] R. I. Goldman, C.-Y. Wei, M. G. Philiastides, A. D. Gerson, D. Friedman, T. R. Brown, and P. Sajda, "Single-trial discrimination for integrating simultaneous EEG and fMRI: identifying cortical areas contributing to trial-to-trial variability in the auditory oddball task," *NeuroImage*, vol. 47, no. 1, pp. 136–147, 2009.

[98] J. M. Walz, R. I. Goldman, M. Carapezza, J. Muraskin, T. R. Brown, and P. Sajda, "Simultaneous EEG–fMRI reveals a temporal cascade of task-related and default-mode activations during a simple target detection task," *NeuroImage*, 2013.

[99] S. Dähne, F. C. Meinecke, S. Haufe, J. Höhne, M. Tangermann, K.-R. Müller, and V. V. Nikulin, "SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters," *NeuroImage*, vol. 86, no. 0, pp. 111–122, 2014.

[100] 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, 2008.

[101] G. Dornhege, M. Krauledat, K.-R. Müller, and B. Blankertz, "General signal processing and machine learning tools for BCI," in *Toward Brain-Computer Interfacing* (G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, eds.), pp. 207–233, Cambridge, MA: MIT Press, 2007.

[102] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[103] C. Zich, S. Debener, C. Kranczioch, M. G. Bleichner, I. Gutberlet, and M. De Vos, "Real-time EEG feedback during simultaneous EEG-fMRI identifies the cortical signature of motor imagery," *NeuroImage*, 2015. in revision.

[104] S. Fazli, S. Dähne, W. Samek, F. Bießmann, and K.-R. Müller, "Learning from more than one data source: data fusion techniques for sensorimotor rhythm-based Brain-Computer Interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 891–906, 2015.

[105] V. D. Calhoun, T. Adali, G. Pearlson, and K. Kiehl, "Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data," *NeuroImage*, vol. 30, no. 2, pp. 544–553, 2006.

[106] J. Liu and V. Calhoun, "Parallel independent component analysis for multimodal analysis: application to fMRI and EEG data," in *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro 2007*, pp. 1028–1031, IEEE, 2007.

[107] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero, and V. Calhoun, "Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Human Brain Mapping*, vol. 30, no. 1, pp. 241–255, 2009.

[108] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, "Linked independent component analysis for multimodal data fusion," *NeuroImage*, vol. 54, no. 3, pp. 2198–2217, 2011.

[109] V. D. Calhoun, T. Adali, K. A. Kiehl, R. Astur, J. J. Pekar, and G. D. Pearlson, "A method for multitask fMRI data fusion applied to schizophrenia," *Human Brain Mapping*, vol. 27, no. 7, pp. 598–610, 2006.

[110] B. Mijovi, K. Vanderperren, N. Novitskiy, B. Vanrumste, P. Stiers, B. Van den Bergh, L. Lagae, S. Sunaert, J. Wagemans, S. Van Huffel, and M. De Vos, "The why and how of JointICA: Results from a visual detection task," *NeuroImage*, vol. 60, no. 2, pp. 1171–1185, 2012.

[111] B. Mijovi, M. De Vos, K. Vanderperren, B. Machilsen, S. Sunaert, S. Van Huffel, and J. Wagemans, "The dynamics of contour integration: A simultaneous EEGfMRI study," *NeuroImage*, vol. 88, pp. 10–21, 2014.

[112] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[113] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[114] C. Jordan, "Essai sur la Géometrie à n dimensions," *Bulletin de la Société Mathématique de France*, vol. 3, pp. 103–174, 1875. Tome III, Gauthiers-Villars, Paris, 1962, 79-149.

[115] L. Sun, S. Ji, S. Yu, and J. Ye, "On the equivalence between canonical correlation analysis and orthonormalized partial least squares," in *International Joint Conference on Artifical Intelligence (IJCAI)*, 2009.

[116] A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi, "Partial least squares (PLS) methods for neuroimaging: A tutorial and review," *NeuroImage*, vol. 56, no. 2, pp. 455–475, 2011.

[117] E. Martínez-Montes, P. A. Valdés-Sosa, F. Miwakeichi, R. I. Goldman, and M. S. Cohen, "Concurrent EEG/fMRI analysis by multiway partial least squares," *NeuroImage*, vol. 22, no. 3, pp. 1023–1034, 2004.

[118] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis.," tech. rep., UC Berkeley, 2006.

[119] C. Archambeau and F. R. Bach, "Sparse probabilistic projections," in *Advances in Neural Information Processing Systems 21*, pp. 73–80, 2009.

[120] L. R. Tucker, "An inter-battery method of factor analysis," *Psychometrika*, vol. 23, no. 2, pp. 111–136, 1958.

[121] N. M. Correa, Y.-O. Li, T. Adali, and V. D. Calhoun, "Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 998–1007, 2008.

[122] N. Correa, Y.-O. Li, T. Adali, and V. Calhoun, "Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, pp. 385–388, 2009.

[123] G. Varoquaux, S. Sadaghiani, P. Pinel, A. Kleinschmidt, J. Poline, and B. Thirion, "A group model for stable multi-subject ICA on fMRI datasets," *NeuroImage*, vol. 51, no. 1, pp. 288–299, 2010.

[124] M. Gaebler, F. Bießmann, J.-P. Lamke, K.-R. Müller, H. Walter, and S. Hetzer, "Stereoscopic depth increases intersubject correlations of brain networks," *NeuroImage*, vol. 100, pp. 427–434, 2014.

[125] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 39–50, 2010.

[126] H. Akaike, "Canonical correlation analysis of time series and the use of an information criterion," in *System Identification Advances and Case Studies* (R. K. Mehra and D. G. Lainiotis, eds.), vol. 126 of *Mathematics in Science and Engineering*, pp. 27 – 96, Elsevier, 1976.

[127] F. Bießmann, F. C. Meinecke, A. Gretton, A. Rauch, G. Rainer, N. Logothetis, and K.-R. Müller, "Temporal kernel canonical correlation analysis and its application in multimodal neuronal data analysis," *Machine Learning*, vol. 79, no. 1-2, pp. 5—27, 2009.

[128] F. Bießmann, Y. Murayama, N. K. Logothetis, K.-R. Müller, and F. C. Meinecke, "Improved decoding of neural activity from fMRI signals using non-separable spatiotemporal deconvolutions," *NeuroImage*, vol. 61, no. 4, pp. 1031–1042, 2012.

[129] N. M. Correa, T. Eichele, T. Adal, Y.-O. Li, and V. D. Calhoun, "Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI," *NeuroImage*, vol. 50, no. 4, pp. 1438–1445, 2010.

[130] S. Dähne, F. Bießman, F. C. Meinecke, J. Mehnert, S. Fazli, and K.-R. Müller, "Integration of multivariate data streams with bandpower signals," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1001–1013, 2013.

[131] D. A. Handwerker, J. M. Ollinger, and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *NeuroImage*, vol. 21, no. 4, pp. 1639–1651, 2004.

[132] F. Pedregosa, M. Eickenberg, P. Ciuciu, B. Thirion, and A. Gramfort, "Data-driven HRF estimation for encoding and decoding models," *NeuroImage*, vol. 104, pp. 209–220, 2015.

[133] S. Dähne, V. V. Nikulin, D. Ramírez, P. J. Schreier, K.-R. Müller, and S. Haufe, "Finding brain oscillations with power dependencies in neuroimaging data," *NeuroImage*, vol. 96, pp. 334–348, 2014.

[134] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Neural Networks*, vol. 12, pp. 181–201, May 2001.

[135] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[136] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, pp. 41–48, IEEE, 1999.

[137] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. J. Smola, and K.-R. Müller, "Learning discriminative and invariant nonlinear features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 623–628, 2003.

[138] S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller, "Kernel-based nonlinear blind source separation," *Neural Computation*, vol. 15, pp. 1089–1124, 2003.

[139] R. Iaci and T. N. Sriram, "Robust multivariate association and dimension reduction using density divergences," *Journal of Multivariate Analysis*, vol. 117, pp. 281–295, May 2013.

[140] V. Vapnik, *The nature of statistical learning theory*. springer, 2000.

[141] K.-R. Müller, C. W. Anderson, and G. E. Birch, "Linear and non-linear methods for brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 165–169, 2003.

[142] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*. Winston Washington, DC, 1977.

[143] T. D. Bie and B. D. Moor, "On the regularization of canonical correlation analysis," *International Symposium on Independent Component Analysis and Blind Signal Separation*, Jan 2003.

[144] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[145] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion, "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity," in *Information Processing in Medical Imaging*, pp. 562–573, Springer, 2011.

[146] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion, "Multiscale mining of fMRI data with hierarchical structured sparsity," *SIAM Journal on Imaging Sciences*, vol. 5, no. 3, pp. 835–856, 2012.

[147] 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.

[148] S. Amari, N. Murata, K.-R. Müller, M. Finke, and H. H. Yang, "Asymptotic statistical theory of overtraining and cross-validation," *Neural Networks, IEEE Transactions on*, vol. 8, no. 5, pp. 985–996, 1997.

[149] C. M. Bishop, *Neural networks for pattern recognition*. Clarendon press Oxford, 1995.

[150] N. Murata, S. Yoshizawa, and S. Amari, "Network information criterion-determining the number of hidden units for an artificial neural network model," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 865–872, 1994.

[151] P. J. Huber, *Robust statistics*. Springer, 2011.

[152] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," *Tokyo Institute of Statistical Mathematics, Tokyo, Japan, Technical Report*, 2001.

[153] A. Mandal and A. Cichocki, "Non-linear canonical correlation analysis using alpha-beta divergence," *Entropy*, vol. 15, no. 7, pp. 2788–2804, 2013.

[154] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems 26* (L. Bottou, C. Burges, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), pp. 1007–1015, MIT Press, 2013.

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

[156] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," *Neural Computation*, vol. 14, no. 8, pp. 1859–1886, 2002.

[157] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, ACM, 1992.

[158] 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, 2011.

[159] R. Tomioka and K. R. Müller, "A regularized discriminative framework for EEG analysis with application to brain-computer interface," *NeuroImage*, vol. 49, pp. 415–432, 2010.

[160] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, 2004.

[161] R. Hari and M. V. Kujala, "Brain basis of human social interaction: from concepts to brain imaging," *Physiological Reviews*, vol. 89, no. 2, pp. 453–479, 2009.

[162] J. P. Dmochowski, P. Sajda, J. Dias, and L. C. Parra, "Correlated components of ongoing EEG point to emotionally laden attention–a possible marker of engagement?," *Frontiers in Human Neuroscience*, vol. 6, 2012.

[163] J. P. Dmochowski, M. A. Bezdek, B. P. Abelson, J. S. Johnson, E. H. Schumacher, and L. C. Parra, "Audience preferences are predicted by temporal reliability of neural processing," *Nature Communications*, vol. 5, 2014.

[164] P. R. Montague, G. S. Berns, J. D. Cohen, S. M. McClure, G. Pagnoni, M. Dhamala, M. C. Wiest, I. Karpov, R. D. King, N. Apple, *et al.*, "Hyperscanning: simultaneous fMRI during linked social interactions," *NeuroImage*, vol. 16, no. 4, pp. 1159–1164, 2002.

[165] F. Babiloni, L. Astolfi, F. Cincotti, D. Mattia, A. Tocci, A. Tarantino, M. Marciani, S. Salinari, S. Gao, A. Colosimo, *et al.*, "Cortical activity and connectivity of human brain during the prisoner's dilemma: an EEG hyperscanning study," in *29th Annual International Confonference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4953–4956, 2007.

[166] J. Sui, H. He, G. D. Pearlson, T. Adali, K. A. Kiehl, Q. Yu, V. P. Clark, E. Castro, T. White, B. A. Mueller, B. C. Ho, N. C. Andreasen, and V. D. Calhoun, "Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia," *NeuroImage*, vol. 66, pp. 119–132, 2013.

[167] J. Sui, T. Adali, G. D. Pearlson, and V. D. Calhoun, "An ICA-based method for the identification of optimal fMRI features and components using combined group-discriminative techniques," *NeuroImage*, vol. 46, no. 1, pp. 73–86, 2009.

[168] F. Bießmann, J.-M. Papaioannou, M. Braun, and A. Harth, "Canonical trends: Detecting trend setters in web data," in *Proceedings of the International Conference on Machine Learning*, 2012.

[169] P. von Bünau, F. C. Meinecke, F. Király, and K.-R. Müller, "Finding stationary subspaces in multivariate time series," *Physical Review Letters*, vol. 103, p. 214101, 2009.

Sven Dähne obtained a B.Sc. in Cognitive Science in 2007 from the University of Osnabrück and a M.Sc. in Computational Neuroscience in 2010 from TU Berlin and the Bernstein Center for Computational Neuroscience (BCCN) in Berlin. He is currently working towards his Ph.D. in the Machine Learning Group at TU Berlin, while being associated to the BCCN as well as the Berlin Big Data Center (BBDC). He has been working on machine learning methods for online-adaptation of EEG based BCIs and investigations in the causes for non-stationarity in BCI performance. His current research focuses on the development of analysis methods for spectral modulations in the context of uni- and multimodal neuroimaging data.

Felix Bießmann obtained a BSc in Cognitive Science in 2005 from the University of Osnabrück, a MSc in Neuroscience at the International Max-Planck Research School, Tübingen, and a PhD in Machine Learning from Berlin Institute of Technology. From 2013 to 2014 he was Assistant Professor at Korea University. He currently is with the Machine Learning Group at Amazon Development Center Berlin. His research interests include statistical learning methods for multimodal data with a focus on neuroscientific and biomedical data.

Wojciech Samek (Member, IEEE) received a Diploma degree in computer science from Humboldt University of Berlin, Berlin, Germany, in 2010 and the Ph.D. degree in machine learning from the Technische Universität Berlin, Berlin, Germany, in 2014. In the same year he founded the Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, which he currently directs. He is associated with the Berlin Big Data Center (BBDC) and was a Scholar of the German National Academic Foundation and a Ph.D. Fellow at the Bernstein Center for Computational Neuroscience (BCCN) Berlin. He was visiting Heriot-Watt University, Edinburgh, U.K., and the University of Edinburgh, Edinburgh, from 2007 to 2008 and in 2009 he was with the Intelligent Robotics Group at NASA Ames Research Center in Mountain View, CA, USA. His research interests include machine learning, biomedical engineering, neuroscience, and computer vision.

Stefan Haufe is a Marie-Curie postdoctoral fellow at Columbia University, New York. He was previously a postdoctoral researcher at the City College of New York and Technische Universität Berlin. He received a Ph.D. degree in natural sciences from TU Berlin in 2011 and a Diploma in computer science from Martin-Luther-Universität Halle-Wittenberg in 2005. His research interests include brain connectivity analysis, electrical and optical forward and inverse modeling, statistical source separation, decoding, interpretation in neuroscience, brain-computer interfacing, mental state monitoring, and the study of attention and emotion using hyperscanning paradigms.

Dominique Goltz obtained a B.Sc. in Cognitive Science in 2007 from the University of Osnabrück and a M.Sc. in Neuro-Cognitive Psychology in 2009 from the Ludwig-Maximilians-Universität München. She is currently working towards her Ph.D. at the Max Planck Institute for Human Cognitive and Brain Sciences and University of Leipzig. Her current research focuses on different cognitive processes during somatosensory information processing and multimodal imaging techniques.

Christopher Gundlach received his Diploma in Psychology in 2010 from the University of Leipzig. He is currently working towards his Ph.D. at the Max Planck Institute for Human Cognitive and Brain Sciences and the University of Leipzig and is now a research associate at the Department of Experimental Psychology and Methods at the University of Leipzig. His research focuses on neuronal oscillations, their relation to perception and cognitive processes like attention, their modulation by non-invasive brain stimulation techniques and multimodal imaging techniques.

Arno Villringer received is Doctorate M.D. in Medicine from the University of Freiburg, Germany. He was a resident at the University Munich from 1986 until 1992. He received his Habilitation in Neurology in 1994 and was a Consultant Physician at the Department of Neurology at the Charite Hospital in Berlin, Germany from 1993 until 1996. He was the vice chairman at the Department of Neurology at the Charite from 1996 until 2004. and a C3 professor of Neurology at the Charite from 1997 until 2008. In 2007 he became the director of the Department of Neurology at the Max Planck Institute for Human Cognitive and Brain Sciences as well as the director of the Clinic of Cognitive Neurology at the University Clinic Leipzig, Germany. Since 1999 he is the coordinator of the Competence Network Stroke. In 2006 he co-founded the Berlin School of Mind and Brain in the as part of an Excellence Initiative and since 2006 he is the speaker of the Berlin School of Mind and Brain. In 2008 he became an honorary professor at the Charite University Medicine Berlin. Since 2009 he is a Professor for Cognitive Neurology at the University of Leipzig.

Siamac Fazli received his B.Sc. Physics degree from the University of Exeter in 2002, his M.Sc. in Medical Neurosciences from the Humboldt University Berlin in 2004 and his Ph.D. from the Berlin Institute of Technology in 2011. From 2011-2013 he worked as a Postdoc researcher at the Berlin Institute of Technology for the Bernstein Focus Neurotechnology. Since 2013 he works as an Assistant Professor at Korea University. His current research interests include neuroscience, machine learning, multi-modal neuroimaging and brain-computer interfacing.

Klaus-Robert Müller (Member, IEEE) has been a professor of computer science at Technische Universität Berlin since 2006; at the same time he has been the director of the Bernstein Focus on Neurotechnology Berlin, since 2014 he co-directs the Berlin Big Data Center. He studied physics in Karlsruhe from 1984 to 1989 and obtained his Ph.D. degree in computer science at Technische Universität Karlsruhe in 1992. After completing a postdoctoral position at GMD FIRST in Berlin, he was a research fellow at the University of Tokyo from 1994 to 1995. In 1995, he founded the Intelligent Data Analysis group at GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. From 1999 to 2006, he was a professor at the University of Potsdam. He was awarded the 1999 Olympus Prize by the German Pattern Recognition Society, DAGM, and, in 2006, he received the SEL Alcatel Communication Award. In 2012, he was elected to be a member of the German National Academy of Sciences-Leopoldina and in 2014 he was awarded the Berlin Science Award. His research interests are intelligent data analysis, machine learning, signal processing, and brain-computer interfaces.