NeuroImage 87 (2014) 96–110

Contents lists available at ScienceDirect

NeuroImage

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

On the interpretation of weight vectors of linear models in multivariate neuroimaging☆

Stefan Haufe [a,b,*], Frank Meinecke [c,a], Kai Görgen [d,e,f], Sven Dähne [a], John-Dylan Haynes [d,e,b], Benjamin Blankertz [f,b], Felix Bießmann [g,a,*]

[a] Fachgebiet Maschinelles Lernen, Technische Universität Berlin, Germany
[b] Bernstein Focus: Neurotechnology, Berlin, Germany
[c] Zalando GmbH, Berlin, Germany
[d] Bernstein Center for Computational Neuroscience, Charité — Universitätsmedizin, Berlin, Germany
[e] Berlin Center for Advanced Neuroimaging, Charité — Universitätsmedizin, Berlin, Germany
[f] Fachgebiet Neurotechnologie, Technische Universität Berlin, Germany
[g] Korea University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Accepted 31 October 2013
Available online 15 November 2013

Keywords:
Neuroimaging
Multivariate
Univariate
fMRI
EEG
Forward/backward models
Generative/discriminative models
Encoding
Decoding
Activation patterns
Extraction filters
Interpretability
Regularization
Sparsity

ABSTRACT

The increase in spatiotemporal resolution of neuroimaging devices is accompanied by a trend towards more powerful multivariate analysis methods. Often it is desired to interpret the outcome of these methods with respect to the cognitive processes under study. Here we discuss which methods allow for such interpretations, and provide guidelines for choosing an appropriate analysis for a given experimental goal: For a surgeon who needs to decide where to remove brain tissue it is most important to determine the origin of cognitive functions and associated neural processes. In contrast, when communicating with paralyzed or comatose patients via brain–computer interfaces, it is most important to accurately extract the neural processes specific to a certain mental state. These equally important but complementary objectives require different analysis methods. Determining the origin of neural processes in time or space from the parameters of a data-driven model requires what we call a *forward model* of the data; such a model explains how the measured data was generated from the neural sources. Examples are general linear models (GLMs). Methods for the extraction of neural information from data can be considered as *backward models*, as they attempt to reverse the data generating process. Examples are multivariate classifiers. Here we demonstrate that the parameters of forward models are neurophysiologically interpretable in the sense that significant nonzero weights are only observed at channels the activity of which is related to the brain process under study. In contrast, the interpretation of backward model parameters can lead to wrong conclusions regarding the spatial or temporal origin of the neural signals of interest, since significant nonzero weights may also be observed at channels the activity of which is statistically independent of the brain process under study. As a remedy for the linear case, we propose a procedure for transforming backward models into forward models. This procedure enables the neurophysiological interpretation of the parameters of linear backward models. We hope that this work raises awareness for an often encountered problem and provides a theoretical basis for conducting better interpretable multivariate neuroimaging analyses.

Introduction

For many years, *mass-univariate* methods (e.g., Friston et al., 1994; Luck, 2005; Pereda et al., 2005) have been the most widely used for analyzing multivariate neuroimaging data. In such methods, every single measurement channel (e.g., functional magnetic resonance imaging (fMRI) voxel or electroencephalography (EEG) electrode) is individually related to a target variable, which represents, for example, behavioral or stimulus parameters, which are considered as a model for neural activation. In contrast, *multivariate* methods combine information from different channels. This approach makes it possible to cancel out noise and thereby to extract the brain signals of interest with higher sensitivity and specificity (Bießmann et al., 2009; Blankertz et al., 2002, 2008, 2011; Comon, 1994; Dähne et al., 2014; Dolce & Waldeier, 1974; Donchin & Heffley, 1978; Haufe et al., 2010; Hyvärinen et al., 2001; Koles et al., 1995; Kragel et al., 2012; Kriegeskorte et al., 2006; Lemm et al., 2011; Nikulin et al., 2011; Nolte et al., 2006; Parra et al., 2003, 2008; von Bünau et al., 2009).

* Corresponding author at: Fachgebiet Maschinelles Lernen, Technische Universität Berlin, Germany.
E-mail addresses: stefan.haufe@tu-berlin.de (S. Haufe), felix.biessmann@tu-berlin.de (F. Bießmann).

1053-8119/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.10.067

The goals of neuroimaging analyses can be broadly categorized in two classes as illustrated by the following typical application scenarios.

Interpretability for neuroscience and clinical use. Basic neuroscience research is often concerned with determining the brain regions (or measurement channels), frequencies, or time intervals reflecting a certain cognitive process. Here we call analyses, for which this is possible, *interpretable* with respect to these processes. In extreme cases, interpretable methods could even be used to answer questions like "Where can a surgeon cut, without damaging a certain brain function?"

Accurate brain state estimation for BCIs. In other applications such as brain–computer interfacing (BCI, Dornhege et al., 2007; Wolpaw & Wolpaw, 2012), researchers are mainly interested in estimating (or decoding) brain states from neuroimaging data, or vice versa. For analysis methods in this scenario, the accuracy of decoding is more important than the interpretability of the model parameters.

There is generally no reason to believe that the decoding models used for BCIs should at the same time be interpretable. But this is exactly what is sometimes implicitly assumed. For example, one may contrast the brain activity in two experimental conditions using a multivariate classifier. Although classifiers are designed for a different purpose (estimation of brain states, that is), it is common to interpret their parameters with respect to properties of the brain. *A widespread misconception about multivariate classifier weight vectors is that (the brain regions corresponding to) measurement channels with large weights are strongly related to the experimental condition.* In fact, such conclusions can be unjustified. Classifier weights can exhibit small amplitudes for measurement channels containing the signal-of-interest, but also large amplitudes at channels *not* containing this signal. In an extreme scenario, in which a surgeon bases a decision about which brain areas to cut on, e.g., classifier weights, both Type I and Type II errors may thus occur, with potentially severe consequences: the surgeon may cut wrong brain areas and actually miss correct ones. The goal of this paper is to raise awareness of this problem in the neuroimaging community and to provide practitioners with easy recipes for making their models interpretable with respect to the neural processes under study. Doing so, we build on prior work contained in Parra et al. (2005), Hyvärinen et al. (2009), Blankertz et al. (2011), Naselaris et al. (2011) and Bießmann et al. (2012b).

While we here focus on *linear* models, nonlinear ones suffer from the same interpretational difficulties. Besides their simplicity, linear models are often preferred to nonlinear approaches in decoding studies, because they combine information from different channels in a weighted sum, which resembles the working principle of neurons (Kriegeskorte, 2011). Moreover, they typically yield comparable estimation accuracy in many applications (Misaki et al., 2010).

The article is structured as follows. We start in the Methods section with three simple examples illustrating how coefficients of linear classifiers may severely deviate from what would reflect the simulated "physiological" truth. Next, we establish a distinction of the models used in multivariate data analysis into forward and backward models. Roughly speaking, forward models express the observed data as functions of some underlying variables, which are of interest for the particular type of analysis conducted (e.g., are maximally mutually independent, or allow the best estimation with respect to certain brain states, etc.). In contrast, backward models express those variables of interest as functions of the data. We point out that the interpretability of a model depends on the direction of the functional relationship between observations and underlying variables: the parameters of forward models are interpretable, while those of backward models typically are not. However, we provide a procedure for transforming backward models into corresponding forward models, which works for the linear case. By this means, interpretability can be achieved for methods employing linear backward models such as linear classifiers.

In the Experiments and Experimental results sections we demonstrate the benefit of the proposed transformation for a number of established multivariate methods using synthetic data as well as real EEG and fMRI recordings. In the Discussion section, we discuss theoretical and practical issues related to our findings, as well as non-linear generalizations and relations to the popular searchlight approach in neuroimaging (Chen et al., 2011; Kriegeskorte et al., 2006). Conclusions are drawn in the Conclusions section.

Methods

Our considerations apply in the same way to EEG, fMRI and any other measurements. Moreover, it is not required that each dimension of the data exactly corresponds to one physical sensor (fMRI voxel, EEG electrode). For example, one may as well consider "spatial features", where every data channel corresponds to a different time point or interval of the same physical measurement sensor (see Example 3 in the Three classification examples section). Generally, the data may be composed of any features derived from the original measurements through linear or nonlinear processing, and may even comprise higher-order interaction measures between physical sensors, as in Shirer et al. (2012). We refer to all such features simply as data *channels*.

In the following, the number of channels will be denoted by M and the data of channel m (with $m \in \{1, \ldots, M\}$) will be called x_m. Furthermore, to obtain a concise notation, we combine all channels' data into the vector $\mathrm{x} = [x_1, \ldots, x_M]^{\mathrm{T}} \in \mathbb{R}^M$. Finally, we will assume that N data samples $\mathbf{x}(n) = 1, \ldots, N$ are available, where in the neuroimaging context the index n may often refer to time. In analogy, we will assume the presence of K so-called latent factors in the data (see Forward models and activation patterns and Backward models and extraction filters sections), where the n-th sample of these factors is summarized as $\mathbf{s}(n) = [s_1(n), \ldots, s_K(n)]^{\mathrm{T}} \in \mathbb{R}^K$. Finally, in *supervised* settings, each latent factor $s_k(n)$ is linked to an externally given target variable $y_k(n)$. These targets can either take continuous (e.g., stimulus intensities or reaction times) or discrete (e.g., class labels indicating the experimental condition) values. The n-th sample of target variables is denoted by $\mathbf{y}(n) = [\mathrm{y}_1(n), \ldots, y_k(n)]^{\mathrm{T}} \in \mathbb{R}^K$. Generally, we set scalar values in italic face, while vector-valued quantities and matrices are set in bold face. An overview of the notation is given in Table 1. Denoting $\mathbf{x}(n)$ the measured variable and target variables as $\mathbf{y}(n)$ we follow the standard convention in the machine learning community. Although we are aware of the convention in the fMRI literature to denote the design matrix as $\mathbf{X}$, we deliberately chose the machine learning nomenclature: the problem of interpretability arises when using multivariate classifiers, which are more associated with machine learning than with standard fMRI methods.

Three classification examples

Example 1. Consider a binary classification setting in which we want to contrast the brain activity in two experimental conditions based on the

Table 1
Notation.

N	Number of data points
M	Number of measurement channels
K	Number of latent factors or target variables
$\mathbf{x}(n)$	M-dimensional vector of observed data
$\mathbf{s}(n), \hat{\mathbf{s}}(n)$	K-dimensional vector of latent factors
$\mathbf{y}(n)$	K-dimensional vector of target variables
$\epsilon(n)$	M-dimensional noise vector in forward models
$\mathbf{A}$	$M \times K$ matrix of patterns in forward models
$\mathbf{W}$	$M \times K$ matrix of filters in backward models
Σ_{x}	Data covariance
$\Sigma_{\hat{\mathbf{s}}}$	Covariance of the latent factors
$\Sigma_{\epsilon\epsilon}$	Noise covariance in forward models

observations in two channels, $x_1(n)$ and $x_2(n)$. Imagine that $x_1(n)$ contains the signal of interest $s(n)$ (e.g., the presence or absence of an experimental condition), but also a strong distractor signal $d(n)$ (e.g., heart beat) that overshadows the signal of interest. For example, $x_1(n) = s(n) + d(n)$. Channel $x_2(n)$ measures the same distractor signal, but *not* the signal of interest, i.e. $x_2(n) = d(n)$. Combining the information from both channels, the signal of interest can easily be recovered by taking the difference $x_1(n) - x_2(n) = \mathbf{w}^T\mathbf{x}(n)$, where $\mathbf{w} = [1, -1]^T$ is the weight vector of the optimal linear classifier. Importantly, this classifier gives equally "strong" weights to both channels. Thus, interpreting those weights as evidence that the signal-of-interest is present in a channel would lead to the erroneous conclusion that the signal is (also) present in channel $x_2(n)$ — which is not the case. In fact, in the course of this paper we will demonstrate that the *only* inference about the signal-of-interest one can draw from the fact that there are nonzero weights on both channels is that that signal is present in at least one of the channels.

Example 2. Fig. 1 illustrates a slightly more complex two-dimensional scenario, which exemplifies that classifier weights on channels not containing the signal of interest might also be positive, and might generally also have a larger magnitude than those of signal-related channels. Consider that the data measured in two conditions (classes) are multivariate Gaussian distributed with different means and equal covariance matrices. Here, we choose the class means to be $\mu_+ = [1.5, 0]^T$ and $\mu_- = [1.5, 0]^T$, and the common covariance matrix to be $\Sigma = \begin{bmatrix} 1.02 & -0.30 \\ -0.30 & 0.15 \end{bmatrix}$ (these values were determined in order to obtain a particular weight vector $\mathbf{w}_{\mathrm{LDA}}$, see below).

Projecting the data onto x_1 by means of the linear transformation $\mathbf{w}_{x_1}^T\mathbf{x}(n) = x_1(n)$ with $\mathbf{w}_{x1} = [1, 0]^T$ yields a reasonable separation of the two classes (see the bottom left panel of Fig. 1), with the correlation of the class label $y(n) \in \{-1, +1\}$ and channel data $x_1(n)$ being $r = \mathrm{Corr}(y, x_1) = 0.83$, where $\mathrm{Corr}(x_1, x_2) = \mathrm{Cov}(x_1, x_2) / (\mathrm{Std}(x_1) \cdot \mathrm{Std}(x_2))$. In contrast, projecting the data onto channel x_2 using $\mathbf{w}_{x_2} = [0, 1]^T$ provides no separation at all (see bottom center panel). Here, $r = 0.04$. Multivariate classification according to linear discriminant analysis (LDA) – which is Bayes-optimal in this specific scenario – achieves the best possible separation of $r = 0.92$ using the projection vector $\mathbf{w}_{\mathrm{LDA}} \propto [1,2]^T$ (see Activation patterns obtained from multivariate OLS decoding are equivalent to a mass-univariate analysis section and Appendix C). Thus, in order to maximize class separation, the weight on $x_2(n)$ must be twice as large as the weight on $x_1(n)$, although $x_2(n)$ does not contain class-specific information at all.

Example 3. Finally, the interpretability issues outlined above do not only hold for data with *spatial* structure as derived from different EEG electrodes or fMRI channels, but for arbitrary features derived from data. To illustrate that, imagine the classification of stimulus-evoked vs. baseline neural activity based on pre- and poststimulus (that is, *temporal*) features. Here, the difference of pre- and poststimulus activity should be highly class-specific. Thus, large (temporal) classifier weights might be assigned to both features. Analyzing these weights one might incorrectly conclude that the pre-stimulus interval contains class-specific information.

These three examples demonstrate that classifier weights may convey misleading information about the class-related directions in the data, where the LDA classifier is representative of an entire class of methods estimating so-called backward models, for which such seemingly counter-intuitive behavior is necessary and systematic. We will discuss backward models after introducing their counterparts, forward models, in the following.

Forward models and activation patterns

Forward models express the observed data as functions of some latent (that is, hidden) variables called components or *factors*. Since they provide a model for the generation of the observed data, they are also referred to as generative models in the machine learning literature.

In the linear case, the data $\mathbf{x}(n)$ are expressed as the sum of K factors $s_k(n)$ $(k \in \{1, \ldots, K\})$, which are weighted by their

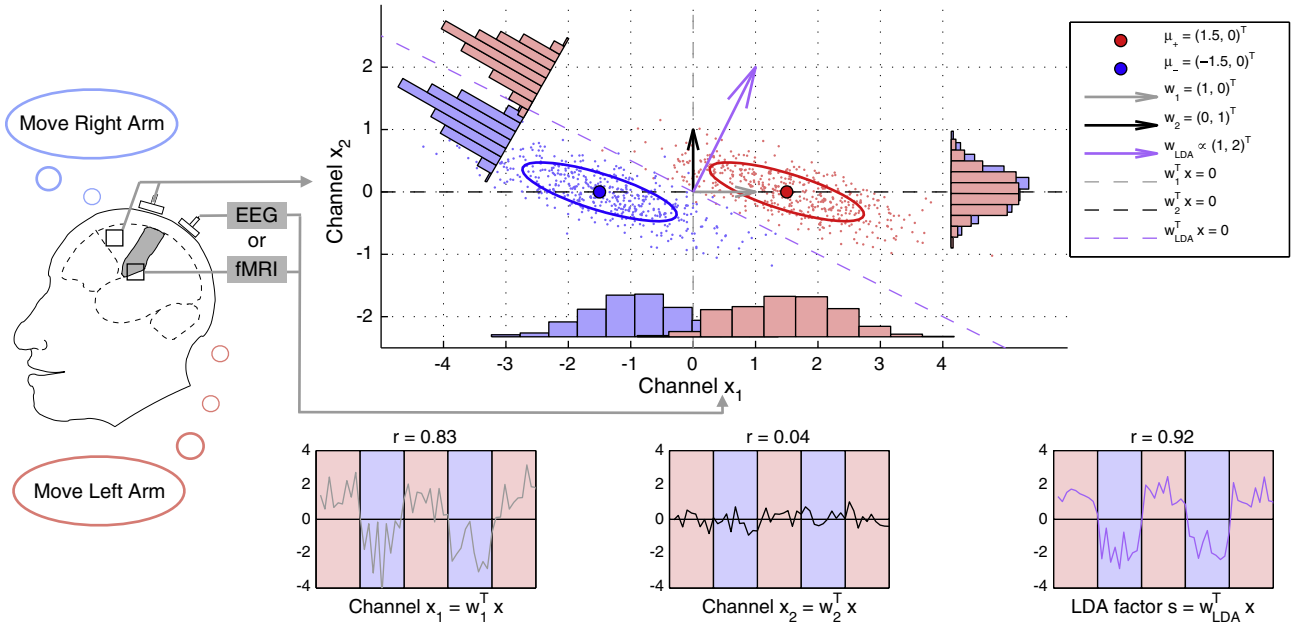

Fig. 1. Two-dimensional example of a binary classification setting. The class-conditional distributions are multivariate Gaussians with equal covariance matrix. The class means differ in channel $x_1(n)$, but not in channel $x_2(n)$. Thus, channel $x_2(n)$ does not contain any class-related information. Nevertheless, Bayes-optimal classification according to linear discriminant analysis (LDA) projects the data onto the weight vector (extraction filter) $\mathbf{w}_{\mathrm{LDA}} \propto [1, 2]^T$, i.e., assigns twice the weight of channel $x_1(n)$ to channel $x_2(n)$. This large weight on $x_2(n)$ is needed for compensating the skewed correlation structure of the data, and must not be interpreted in the sense that the activity at $x_2(n)$ is class-specific. By transforming the LDA projection vector into a corresponding activation pattern $\mathbf{a}_{\mathrm{LDA}}$ using Eq. (7), we obtain $\mathbf{a}_{\mathrm{LDA}} \propto [1, 0]^{T}$, which correctly indicates that $x_1(n)$ is class-specific, while $x_2(n)$ is not.

Table 2
Comparison of linear *forward* and *backward* modeling perspectives associated with *activation patterns* and *extraction filters*, respectively, as well as the special supervised cases of *encoding* and *decoding*.

	Forward model	Backward model
Alternative name	Generative model	Discriminative model
Model (linear case)	$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \epsilon(n)$	$\mathbf{W}^\top \mathbf{x}(n) = \hat{\mathbf{s}}(n)$
Purpose	Factorize the data into *latent factors* $\mathbf{s}(n)$ and their corresponding *activation patterns* (columns of $\mathbf{A}$), plus noise $\epsilon\epsilon(n)$.	Extract *latent factors* $\hat{\mathbf{s}}(n)$ from the data by multiplying with *extraction filters* (columns of $\mathbf{W}$).
Interpretable	$\mathbf{A}$, $\mathbf{s}(n)$	$\hat{\mathbf{s}}(n)$
Supervised case	Encoding: Replace latent factors $\mathbf{s}(n)$ by known external target variables $\mathbf{y}(n)$ or pre-estimated factors $\hat{\mathbf{s}}(n)$. Thus, estimate how $\mathbf{y}(n)$ or $\hat{\mathbf{s}}(n)$ are *encoded* in the measurement.	Decoding: Seek latent factors $\hat{\mathbf{s}}(n)$ to approximate known external target variables $\mathbf{y}(n)$. Thus, estimate how $\mathbf{y}(n)$ can be *decoded* from the measurement.

corresponding *activation patterns* $\mathbf{a}_k \in R^M$, plus additive noise. That is, $\mathbf{x}(n) = \Sigma_k \mathbf{a}_k \mathbf{s}_k(n) + \epsilon(n)$, or, in matrix notation

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \epsilon(n), \tag{1}$$

where $\mathbf{s}(n) = [s_1(n), \ldots, s_K(n)]^\top \in \mathbb{R}^K$ and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] \in \mathbb{R}^{M \times K}$, cf. also the linear model presented in Parra et al. (2005). In the most general case, $\mathbf{A}$ and $\mathbf{s}(n)$ are estimated jointly, which is referred to as the blind source separation (BSS) setting. Since the factorization into $\mathbf{A}$ and $\mathbf{s}(n)$ is not unique, assumptions have to be imposed, where different assumptions generally also lead to different factorizations.

Each estimated factor s_k can be thought of as a specific signal-of-interest "hidden" in the data, e.g., a brain response to a stimulus, which is isolated from other signals of cerebral or extracerebral origin. The corresponding activation pattern $\mathbf{a}_k$ encodes the strength and polarity with which the factor's activity is present in each channel. As such, activation patterns have a clear physiological interpretation, which can be summarized as follows:

> *The entries of the activation pattern* $\mathbf{a}_k$ *show, in which channels (e.g., fMRI voxels or EEG electrodes) the signal* s_k *is reflected. Value (and sign) are directly related to its strength (and effect direction) at different channels.*

Intuitively, one might think of latent factors as being produced by a specific brain region or network implementing a certain mental function or processing. In EEG, where the physics of volume conduction implies a linear mapping from brain sources to sensors, they are indeed commonly equated with the activity of underlying electrical brain sources. The estimation of a linear forward model here amounts to implicitly solving the EEG inverse problem (see The relevance of linear models for EEG data section).

Imaging modalities such as fMRI measure (the hemodynamic response to) neuronal activity directly inside the brain, such that no inverse problem needs to be solved. Nevertheless, assuming a linear forward model of the data is still adequate, if we assume that each measurement is a summation of the activities of multiple concurrent brain (and noise) processes, each of which might be expressed with different strengths in different subsets of voxels. This assumption of linear superposition builds the foundation of standard fMRI analyses such as using a general linear model (GLM). In the classification example depicted in Fig. 1, for instance, the data actually follow the forward model $\mathbf{x}(n) = \mathbf{a}y(n) + \epsilon(n)$, where the binary class label $y(n) \in \{-1/2, 1/2\}$ plays the role of the latent signal, the mean difference $\mu_+ - \mu_- = (3, 0)^\top = \mathbf{a}$ is the corresponding activation pattern, and where the noise $\epsilon \sim N(0, \Sigma)$ is distributed according to a zero-mean bivariate Gaussian distribution with covariance matrix Σ.

Supervised methods that directly estimate activation patterns of a forward model given a target variable $\mathbf{y}(n)$ are typically referred to as *encoding* approaches (Naselaris et al., 2011). Thus, encoding is the supervised special case of forward modeling (cf., Table 2). As with all forward models, the estimated patterns can be interpreted in the desired way as outlined above. Classifiers such as LDA, however, do not fit into the encoding framework, because they employ backward models as discussed below.

Backward models and extraction filters

Backward models "extract" latent factors $\hat{\mathbf{s}}(n)$ as functions of the observed data, i.e., reverse the direction of the functional dependency between factors and data compared to forward models. They are typically used if there is no need to model the generation of the entire observations, because one is only interested in transforming them into a (potentially low-dimensional) representation in which they exhibit certain desired characteristics. As such, backward models roughly[1] correspond to discriminative models in machine learning.

In the linear case, the mapping from observations to factors can be summarized in the transformation matrix $\mathbf{W} \in \mathbb{R}^{M \times K}$. The backward model then reads

$$\mathbf{W}^\top \mathbf{x}(n) = \hat{\mathbf{s}}(n). \tag{2}$$

As in forward modeling, the model parameters must be fitted under appropriate assumptions, where the assumptions on $\hat{\mathbf{s}}(n)$ may generally be similar to those outlined in the Forward models and activation patterns section. In fact, most of the blind source separation techniques mentioned there can be formulated both in a forward and in a backward modeling context (see Interpreting results of backward modeling: obtaining activation patterns from extraction filters section).

In supervised backward modeling, $\mathbf{W}$ is chosen such that $\hat{\mathbf{s}}(n)$ approximates a target variable, where typically $K < M$. In analogy to the term "encoding" for supervised forward modeling, one here also speaks of *decoding* (Naselaris et al., 2011). An overview of the properties of forward and backward models, as well as their supervised variants of encoding and decoding, is provided in Table 2.

Generally, each column $\mathbf{W}_k \in \mathbb{R}^M$ of $\mathbf{W}$ extracts one factor $\hat{s}_k(n)$, and is referred to as the *extraction filter* for that factor. Hence, just as in forward modeling, every factor is associated with an M-dimensional weight vector. *However, in contrast to forward models these weight vectors appear now in a complementary role.* Instead of multiplying it with a latent factor $\hat{s}_k(n)$ in order to obtain the contribution of that factor to the measured data $\mathbf{x}(n)$, we now multiply it to the measured data $\mathbf{x}(n)$ in order to obtain the latent factor $\hat{s}_k(n)$. Accordingly, there is no general reason why the filter vector $\mathbf{w}_k$ should be similar to the activation pattern $\mathbf{a}_k$ of the same factor $\hat{s}_k(n)$, and its interpretation is different:

> *When projecting observed data onto an extraction filter* $\mathbf{w}_k$*, the result will be a latent component exhibiting certain desired properties (e.g., allow good classification or maximize the similarity to a target variable).*

[1] While backward modeling encompasses both supervised and unsupervised approaches, the term "discriminative" is typically used only for supervised approaches.

The purpose of a filter is two-fold: it should amplify a signal of interest, while it should at the same time suppress all "signals of no interest".[2] The first task alone is achieved best by the activation pattern of the target signal (vector $\mathbf{w}_1$ in Fig. 1), but the second task requires the filter to deviate from that direction in order to be "as perpendicular" as possible to the activation patterns of the strongest disturbing noise sources (Eigenvectors corresponding to the largest Eigenvalues of Σ in Fig. 1). If those latter patterns are not orthogonal to the signal pattern, the trade-off between amplifying the target signal and suppressing the noise requires a rather complex spatial structure in which the "meaning" of filter weights for a certain channel cannot be disentangled between those two tasks. *In particular, the filter weights do not allow one to draw conclusions about the features (e.g., brain voxels) in which the corresponding factor is expressed.*

Moreover, the sign of a particular filter weight does not give an indication about the activity at the respective channels being positively or negatively related to the experimental condition in a classification or regression task. Put short, extraction filters are generally complicated functions of signal *and* noise components in the data.

Interpreting results of backward modeling: obtaining activation patterns from extraction filters

Backward modeling amounts to transforming the data into a supposedly more informative representation, in which the signals of interest are isolated as low-dimensional components or factors. However, we have seen that the filters in $\mathbf{W}$ only tell us how to combine information from different channels to *extract* these factors from data, but not how they are *expressed* in the measured channels. Obviously, if we aim at a neurophysiological interpretation or just a meaningful visualization of the weights, we have to answer the latter question. In other words, we have to construct *activation patterns* from *extraction filters*.

The square case $K = M$

For linear backward modeling approaches extracting exactly $K = M$ linearly independent factors, the extraction filters $\mathbf{W}$ form an invertible square matrix. By multiplying Eq. (2) with $\mathbf{W}^{-\mathrm{T}}$ from the left, where $\mathbf{W}^{-\mathrm{T}}$ denotes the transpose of the inverse of $\mathbf{W}$, we obtain

$$\mathrm{x}(n) = \mathbf{W}^{-\mathrm{T}}\hat{\mathrm{s}}(n),$$

which has the form of a noise-free forward model of the data $\mathbf{x}(n)$ with activation patterns $\mathbf{A} = \mathbf{W}^{-\mathrm{T}}$. As mentioned, this duality of forward and backward linear modeling holds for many BSS methods including most ICA variants.[3] When interpreting the parameters of these methods it is just important that the forward modeling view is adopted, i.e., that the activation patterns $\mathbf{A}$ are interpreted rather than the extraction filters $\mathbf{W}$.

The general case $K \leq M$

For backward modeling approaches estimating a reduced set of $K < M$ factors, obtaining an equivalent forward model is not straightforward, since the filter matrix $\mathbf{W}$ is not invertible anymore. Nonetheless, our goal here is again to find a pattern matrix $\mathbf{A}$ indicating those measurement channels in which the extracted factors are reflected. Therefore, we seek a linear forward model having the form of Eq. (1):

$$\mathbf{x}(n) = \mathbf{A}\hat{\mathbf{s}}(n) + \epsilon(n). \tag{3}$$

In the following, we assume w.l.o.g. that $\mathbb{E}[\mathbf{x}(n)]_n = \mathbb{E}[\hat{\mathbf{s}}(n)]_n = \mathbb{E}[\epsilon(n)]_n = 0$, where $\mathbb{E}[\cdot]_n$ denotes expectation over samples. Then, the associated covariance matrices are given by $\Sigma_{\mathbf{x}} = \mathbb{E}\left[\mathbf{x}(n)\mathbf{x}(n)^{\top}\right]_n$, $\Sigma_{\hat{\mathbf{s}}} = \mathbb{E}\left[\hat{\mathbf{s}}(n)\hat{\mathbf{s}}(n)^{\top}\right]_n$ and $\Sigma_{\epsilon} = \mathbb{E}[\epsilon(n)\epsilon(n)^{\mathrm{T}}]_n$. Moreover, we assume that the latent factors $\hat{\mathbf{s}}(n)$ are linearly independent, which implies that $\mathbf{W}$ must have full rank, i.e., rank $(\mathbf{W}) = K$.

If the noise term $\epsilon(n)$ is uncorrelated with the latent factors $\hat{\mathbf{s}}$, i.e.,

$$\mathbb{E}\left[\epsilon(n)\hat{\mathbf{s}}(n)^{\top}\right]_n = 0, \tag{4}$$

we call Eq. (3) a *corresponding* forward model to the discriminative model Eq. (2). Assuming a corresponding forward model ensures that any variation that can be explained by the latent factors is captured in the term $\mathbf{A}\hat{\mathbf{s}}(n)$, and not in $\epsilon(n)$. This approach leads directly to the following result.

Theorem 1. *For any backward model*

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

the corresponding forward model is unique, and its parameters are obtained by

$$\mathbf{A} = \Sigma_{\mathbf{x}}\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1}. \tag{6}$$

The columns of $\mathbf{A}$ are activation patterns, which, unlike their associated filters contained in $\mathbf{W}$, enable the desired interpretation, i.e., indicate the effect directions and strengths of the extracted latent factors in the measurement channels.

The proof is given in Appendix A. A proof of the existence of a corresponding forward model is given in Appendix B.

Four remarks

Remark 1. While the above result is based on population covariances Σ_x, $\Sigma_{\hat{\mathbf{s}}}$ and Σ_{ϵ}, those can be exchanged by their sample empirical counterparts in order to derive activation patterns in practice.

Remark 2. In the square case, in which the backward modeling method extracts exactly $K = M$ linearly independent factors, $\mathbf{A} = \Sigma_x\mathbf{W}\Sigma_{\hat{s}}^{-1}$ directly reduces to $\mathbf{W}^{-\mathrm{T}}$, the inverse of $\mathbf{W}^{\mathrm{T}}$. Here, $\epsilon(n) = 0$, since all noise is contained in $\hat{\mathbf{s}}(n)$ In the general case $K < M$, however, $\mathbf{A} \neq (\mathbf{W}^{\mathrm{T}})^{+}$; that is, the patterns do *not* coincide with the columns of the familiar Moore Penrose pseudoinverse of $\mathbf{W}^{\mathrm{T}}$, given by $(\mathbf{W}^{\mathrm{T}})^{+} = \lim_{\alpha \to 0}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W} + \alpha\mathbf{I})^{-1}$.

Remark 3. Although the corresponding forward model to a given backward model is uniquely determined by Eq. (6), that does not imply that the decomposition of the data is unique itself. The "correctness" of the activation patterns derived using Eq. (6) thus solely depends on the correctness of the data decomposition provided by the backward modeling step, which in turn depends on the appropriateness of the assumptions used to estimate the backward model (see Backward models can be made interpretable section for a discussion).

Remark 4. Since filters and patterns are dual to each other, we can also construct extraction filters from given activation patterns. In practice this just means solving Eq. (6) for $\mathbf{W}$. Situations, in which this could be useful, are outlined in the Backward models can be made interpretable section.

Simplifying conditions

If the estimated factors $\hat{\mathbf{s}}(n)$ are uncorrelated, which is the default for many (but not all) backward modeling approaches, and trivially true for $K = 1$, we can obtain activation patterns simply as the covariance between data and latent factors:

$$\mathbf{A} \propto \Sigma_x\mathbf{W} = \mathrm{Cov}[\mathbf{x}(n), \hat{\mathbf{s}}(n)], \tag{7}$$

[2] The "signals of no interest" are collectively called noise, although they might comprise not only measurement noise and technical artifacts, but also the activity of all brain processes not currently under study.

[3] In ICA terminology the activation patterns correspond to the *mixing* matrix, the extraction filters to the *demixing* matrix.

where

$$Cov[\mathbf{x}(n), \hat{\mathbf{s}}(n)] = \begin{bmatrix} Cov(x_1(n), \hat{s}_1(n)) & \cdots & Cov[x_1(n), \hat{s}_K(n)] \\ \vdots & \ddots & \vdots \\ Cov[x_M(n), \hat{s}_1(n)] & \cdots & Cov[x_M(n), \hat{s}_K(n)] \end{bmatrix}.$$

This relationship has also been pointed out by Hyvärinen et al. (2009) in the context of independent component analysis for image processing.

Note that in a *decoding* setting, one might simply replace the factor estimate $\hat{\mathbf{s}}(n)$ in Eq. (7) by the external target variable $\mathbf{y}(n)$. In fact, it can be shown that this approximation is exact for ordinary least squares (OLS) decoding (see Activation patterns obt ained from multivariate OLS decoding are equivalent to a mass-univariate analysis section). That is, the activation pattern can be approximated by calculating the covariance Cov[$\mathbf{x}(n)$, $\mathbf{y}(n)$] (*not* the correlation Corr[$\mathbf{x}(n)$, $\mathbf{y}(n)$]) of each single channel's data with the target variable, which amounts to a purely mass-univariate analysis.

Eq. (6) also shows under which conditions extraction filters are proportional to activation patterns, i.e., when filter weights can be interpreted directly. This is possible, *only if* also the individual channels in the observed data (in addition to the factors) are uncorrelated. However, this assumption is hardly ever met for real neuroimaging data (cf., Experiments and Experimental results sections). Therefore, it is indeed crucial to draw the distinction between filters and patterns.

Regression approach

A different way of constructing activation patterns from latent factors is provided by Parra et al. (2005). They propose to find a pattern that, when multiplied by the extracted latent factors, explains the observed data best in the least-squares sense. That is, to fit a forward model using the pre-estimated factors $\hat{\mathbf{s}}(n)$ under the assumption that the noise $\hat{\epsilon}$ is Gaussian distributed. Interestingly, this approach leads to the exact same solution as in Eq. (6):

$$\begin{aligned} \mathbf{A}_{\mathrm{OLS}} &= arg\min_{\widetilde{\mathbf{A}}} \sum_n \left(\mathbf{x}(n) - \widetilde{\mathbf{A}}\hat{\mathbf{s}}(n)\right)^2 \\ &= \Sigma_x \mathbf{W} \Sigma_{\hat{s}}^{-1}. \end{aligned} \tag{8}$$

Treating the calculation of activation patterns as a regression problem provides an interesting perspective, since it suggests a straightforward way to integrate prior knowledge into the activation pattern estimation, which could genuinely improve interpretability in the presence of too few or excessively noisy data. For example, if the underlying factors are believed to contribute only to a few channels, a ℓ_1-norm penalty might be added in order to sparsify **A**. The resulting estimator

$$\mathbf{A}_{\ell_1} = arg\min_{\widetilde{\mathbf{A}}} \sum_n \left(\mathbf{x}(n) - \widetilde{\mathbf{A}}\hat{\mathbf{s}}(n)\right)^2 + \lambda \left\|\widetilde{\mathbf{A}}\right\|_1 \tag{9}$$

is known as the "LASSO" in the statistics literature (Tibshirani, 1996). Note that this is fundamentally different from sparsifying the filters themselves (see Regularization does not make backward models better interpretable section). Other penalties might enforce spatially smooth patterns or sparsity in a suitable function space (e.g., Gramfort et al., 2013; Haufe et al., 2008, 2009, 2011; Vega-Hernández et al., 2008).

Experiments

Simulations

We performed simulations to assess the extent to which mass-univariate measures as well as weight vectors (filters) and corresponding activation patterns of multivariate methods are able to recover the spatial distribution of an underlying simulated factor in a binary classification task as well as in an unsupervised blind source separation (BSS) setting. For these simulations, we extend the simple 2-dimensional example presented in the Three classification examples section to 64 dimensions.

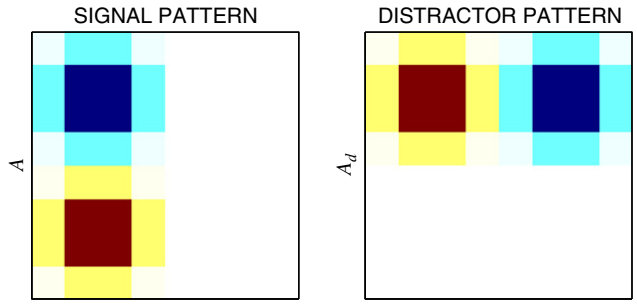

Fig. 2. True activation patterns of the simulated signal and distractor components.

Data generation

The data were generated according to model Eq. (1). There were $K = 2$ components plus additional noise giving rise to a pseudo-measurement at $M = 64$ channels, which are arranged in an 8×8 grid (cf., Fig. 2). The noise was generated according to a 64-dimensional multivariate Gaussian distribution with zero mean and a random covariance matrix, which was uniformly sampled from the set of positive-semi-definite matrices. Of the two designated components, one is a distractor component, while the other one contains the signal-of-interest. The distractor component factor was sampled from a standard normal (Gaussian) distribution. The corresponding activation pattern consists of two blobs with opposite signs in the upper left and upper right corners of the grid. The signal component's activation pattern also consists of two blobs with opposite signs: one in the upper left and one in the lower left corner. Thus, the two components spatially overlap in the upper left corner. The activation patterns of the signal and distractor components are depicted in Fig. 2. In the classification setting, the temporal signature of the signal component factor was set to be the class label $y(n) \in \{-1, +1\}$ plus standard normal distributed noise. In the BSS setting, it was sampled from a univariate Laplace distribution and thus the only non-Gaussian distributed component in the data. The observations were composed as the sum of a 10% signal component portion, a 60% distractor component portion, and a 30% noise portion.

Tested methods

We generated 100 datasets for the classification setting, and 100 corresponding ones for the BSS setting. Each dataset consists of 1000 pseudo-measurements, which were generated using distinct random noise covariance matrices. For each BSS dataset, independent component analysis was carried out using joint approximate diagonalization (JADE, Cardoso & Souloumiac, 1996), where we used the original code provided by the authors. The analysis was restricted to the space of the ten largest principle components. From the resulting ten ICA components, only the one best correlating with the simulated Laplace-distributed factor was investigated.

Note that, while JADE may not be as popular as other ICA algorithms, the specific choice of the ICA algorithm plays only a minor role for demonstrating the different abilities of extraction filters and activation patterns to indicate sensors containing the signal-of-interest, which is the purpose of this simulation rather than demonstrating the performance of a particular method. Therefore, we here designed the non-Gaussian factor-of-interest to be reliably found by any ICA algorithm. In fact, nearly identical results are obtained with FastICA (Hyvärinen, 1999, author's implementation with default parameters, $g(u) = u^3$ nonlinearity). More generally, we could replace the ICA example by a different combination of underlying signals with specific properties

and a corresponding BSS method optimized to extracting such factors, with similar outcome.

For each classification dataset, we computed the mass-univariate correlation Corr[$\mathbf{x}(n), y(n)$] of the class label with each channel reading, as well as the mass-univariate covariance Cov[$\mathbf{x}(n), y(n)$]. Moreover, we conducted multivariate classification by applying linear logistic regression (LLR) on the 64-dimensional data. We considered the unregularized variant of LLR as well as two variants in which either the ℓ_1-norm (LLR-L1) or the ℓ_2-norm (LLR-L2) of the estimated weight vector $\mathbf{w}$ is penalized. The regularization strength in the latter two cases was adjusted using 5-fold cross-validation. Each method gives rise to an extraction filter $\mathbf{w}$. Corresponding activation patterns $\mathbf{a}$ were obtained by transforming the extraction filters according to Eq. (6). Moreover, we used LASSO regression (Eq. (9)) to obtain sparsified patterns $\mathbf{a}_{\ell_1}$. The regularization parameter here was 5-fold cross-validated.

Taken together, we computed the following fourteen 64-dimensional weight vectors in each experiment.

- Indep. component analysis (JADE): $\mathbf{w}$, $\mathbf{a}$, $\mathbf{a}_{\ell_1}$
- Linear logistic regression (LLR): $\mathbf{w}$, $\mathbf{a}$, $\mathbf{a}_{\ell_1}$
- ℓ_1-norm regularized LLR (LLR-L1): $\mathbf{w}$, $\mathbf{a}$, $\mathbf{a}_{\ell_1}$
- ℓ_2-norm regularized LLR (LLR-L2): $\mathbf{w}$, $\mathbf{a}$, $\mathbf{a}_{\ell_1}$
- Mass-univariate correlation: Corr[$\mathbf{x}(n), y(n)$]
- Mass-univariate covariance: Cov[$\mathbf{x}(n), y(n)$].

For further analysis and visualization, all weight vectors were normalized. The weights provided by JADE were moreover adjusted to match the sign of the true pattern.

Performance evaluation

Focusing on the eleven weight vectors computed in the classification context, we evaluated two performance measures. Most importantly, the reconstruction of the true signal component pattern was quantified by means of the correlation between the true pattern and the estimated weight vector. Secondly, the stability of the estimation was assessed by taking the channel-wise variance of the weight vectors across the 100 repetitions, and averaging it across channels.

Spatio-spectral decomposition of EEG data

Oscillatory activity in the alpha-band (8–13 Hz) is the strongest neural signal that can be observed in the EEG. There exist multiple rhythms in the alpha range, each of which reflects synchronization within a specific macroscopic cortical network during idling. We here analyzed a 5-min relaxation measurement recorded prior to an in-car EEG-study on attentional processes (Schmidt et al., 2009). During the recording, the subject sat relaxedly in the driver's seat of a car with his eyes closed. The engine and all electronic devices apart from the EEG instrumentation were switched off. Electroencephalography was recorded from 59 electrodes (located according to the extended 10–20 system, 1000 Hz sampling rate, low cut-off: 0.016 Hz; high cut-off: 250 Hz, nose reference) using BrainAmp recording hardware (Brain Products GmbH, Munich). The EEG signal was digitally low-pass-filtered to 50 Hz and down-sampled to 100 Hz.

We applied spatio-spectral decomposition (SSD) (Nikulin et al., 2011) in order to extract EEG components with strong peaks in the alpha band. By means of SSD, we obtained a full decomposition of the data with invertible pattern and filter matrices, and with factor time series ordered by the ratio of the power in the alpha (8–13 Hz) band and the power in a slightly wider (7–14 Hz) band. Obviously, the power ratio is bounded in [0, 1], and the alpha peak of a component is the more pronounced, the closer the power ratio approaches one. We analyzed the first five SSD components, achieving power ratios between 0.93 and 0.97. A single-dipole scan (Schmidt, 1986) was conducted for each of the spatial activation patterns as well as for each of the extraction filters of the selected components in order to attempt to localize the electrical generators of alpha-band activity in the brain. Note, that performing source localization from filter weights is conceptually wrong (see The relevance of linear models for EEG data section). However, it was performed here for demonstration purposes. The dipole fits were carried out in a realistically shaped three-shell head model based on nonlinearly averaged MRI scans of 152 subjects (Fonov et al., 2011). Forward calculations were performed according to Nolte & Dassios (2005).

fMRI and intracranial neurophysiology

Hemodynamic measurements obtained by fMRI are the most popular neuroimaging modality (Friston, 2009). However they only reflect neural activity indirectly. Thus simultaneous intracranial electrophysiological measurements and fMRI measurements are needed to investigate the complex neurovascular coupling mechanisms that give rise to the fMRI signal (Logothetis et al., 2001). Many multivariate analysis approaches to these multimodal data sets compute filters (Bießmann et al., 2011) which are not interpretable. In this example we illustrate how the relationship between filters and patterns can be applied to this kind of multimodal data. This not only exemplifies an fMRI application of the filter pattern relationship; the example also shows that patterns in multimodal neuroimaging analyses offer a more physiologically plausible and better interpretable perspective on neurovascular coupling mechanisms. The data are simultaneous multimodal recordings of intracranial electrophysiology and high resolution fMRI during spontaneous activity in primary visual cortex of the macaque monkey. For experimental details see Murayama et al. (2010) and Bießmann et al. (2012a,b). Let $\mathbf{x}(n) \in \mathbb{R}^F$ be the electrophysiological band power in F frequency bins and time bins of 1 s duration, recorded at an intracranial electrode in primary visual cortex and $\mathbf{y}(n) \in \mathbb{R}^V$ the functional magnetic resonance data recorded in V voxels and the same time bins as those of the electrophysiological data within a spherical region of interest around the electrode. We applied temporal kernel canonical correlation analysis (tkCCA) (Bießmann et al., 2009) in order to obtain the optimal filters $\mathbf{w}_x(\tau)$, $\mathbf{w}_y$ such that

$$\left\{\mathbf{w}_x(\tau), \mathbf{w}_y\right\} = \underset{\widetilde{\mathbf{w}}_x(\tau), \widetilde{\mathbf{w}}_y}{argmax} \; Corr\left[\sum_{\tau=-N_\tau}^{0} \widetilde{\mathbf{w}}_x(\tau)^\top \mathbf{x}(n+\tau), \widetilde{\mathbf{w}}_y^\top \mathbf{y}(n)\right]. \quad (10)$$

The time–frequency filter $\mathbf{w}_x(\tau)$ transforms the neural spectrogram into decorrelated neural components that approximate the fMRI signal best and the spatial filter $\mathbf{w}_y$ extracts decorrelated components from the fMRI signal which correlate best with the extracted neural components.

Experimental results

Simulations

Fig. 3 depicts an instance of the fourteen weight vectors calculated in one particular simulation. For reference, see the true activation pattern of the signal in Fig. 2 (left). The mass-univariate correlation Corr[$\mathbf{x}(n)$, $\mathbf{y}(n)$] shows high (negative) values only in the lower left corner of the grid. Due to the influence of the distractor component in the upper left corner, the correlation in that area drops dramatically, although the class-specific factor is equally strongly expressed in that corner. The mass-univariate *covariance* Cov[$\mathbf{x}(n)$, $\mathbf{y}(n)$] recovers both active areas equally well.

The spatial filters of the four multivariate methods LLR, LLR-L2, LLR-L1 and JADE show a great variety. The LLR filter has a highly complex structure which does not resemble the true signal pattern at all. The three other filters have less complex structures which show certain similarities to the true pattern. As expected, LLR-L1 delivers a sparse filter vector, while LLR, LLR-L2 and JADE do not. Notably, elastic net

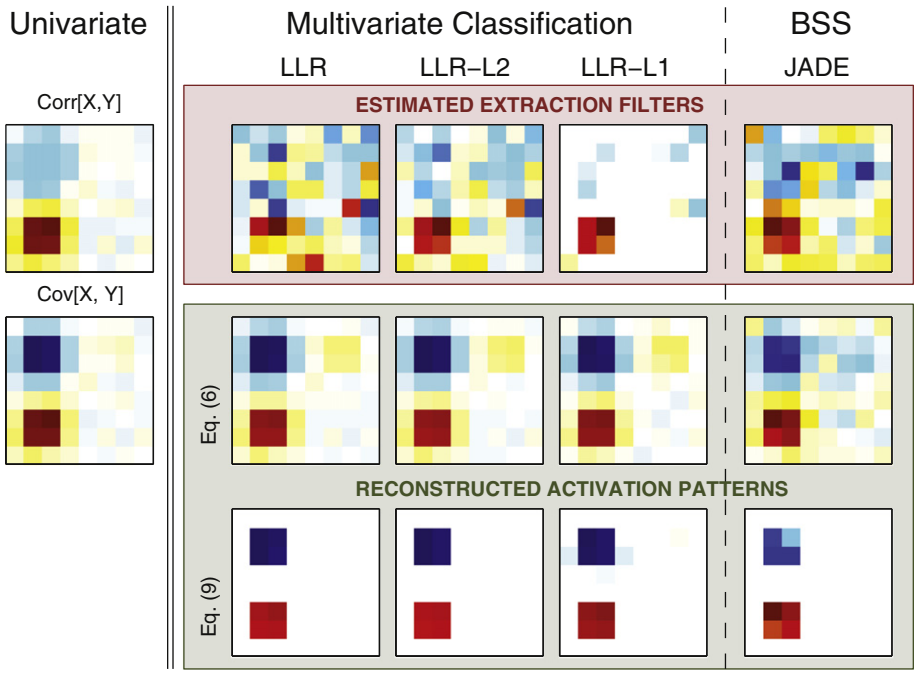

Fig. 3. Mass-univariate measures, as well as extraction filters and corresponding reconstructed activation patterns of multivariate methods in one particular experiment.

regularization (Zou & Hastie, 2005) as used in Carroll et al. (2009) is a hybrid of ℓ_1-norm and ℓ_2-norm regularization, and therefore delivers solutions which are in between LLR-L2 and LLR-L2 in terms of sparsity. Importantly, all filters show large (mostly) negative weights in the upper right corner, where there is no task-related activity at all. These weights are highly stable across repetitions of the experiment; and would be found to significantly differ from zero using statistical testing.

The patterns analytically obtained by Eq. (6) as well as the ℓ_1-norm (i.e., LASSO) regularized patterns estimated by Eq. (9) for all four multivariate approaches are very similar, and generally resemble the true signal activation pattern very well. This is particularly surprising considering the diverse spatial structure of the underlying filters. As expected, the patterns estimated using LASSO are generally slightly sparser. Note, however, that the benefit of performing sparse pattern estimation depends on whether the underlying factors indeed exhibit sparse activations.

Fig. 4 shows the mean pattern reconstruction error and the variance of the entries of the eleven weight vectors calculated in the classification setting. The reconstruction quality (upper panel) is between $r = 0.96$ and $r = 0.99$ (and thus close to the perfect score of $r = 1$) for all of the six pattern estimates, but also for the mass-univariate covariance Cov[$\mathbf{x}(n)$, $\mathbf{y}(n)$]. Although the performance of the ℓ_1-norm regularized

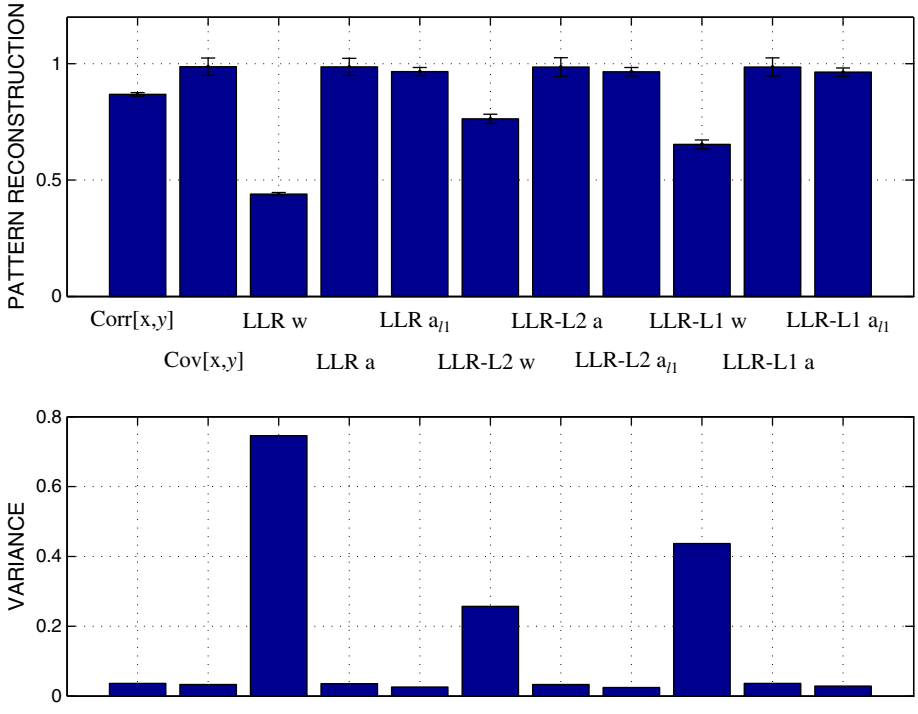

Fig. 4. Pattern reconstruction performance and variance of weight vector entries in the classification setting.

pattern estimates is slightly lower than for the unregularized (this relation reverses, if less samples are used), none of the differences between the seven weight vectors are actually significant. Statistical significance of the difference of two correlations is measured here by transforming correlation coefficients into normal distributed quantities using the Fisher transformation, and performing a two-sample z-test. Compared to all pattern estimates, a significantly lower reconstruction accuracy of $r = 0.87$ is observed for the mass-univariate correlation Corr[$\mathbf{x}(n)$, $\mathbf{y}(n)$]. The three filter weight vectors show the lowest reconstruction accuracies of $r = 0.44$ for OLS, $r = 0.65$ for LLR-L1 and $r = 0.76$ for LLR-L2. These are all significantly different from each other and from the rest. In line with theoretical considerations stating that filters depend on the noise structure of the data in contrast to patterns, all three filters are much less stable across experiments than their corresponding patterns, as well as the two mass-univariate measures (see lower panel of Fig. 4).

EEG data

Fig. 5 depicts the spatial extraction filters and activation patterns corresponding to the five SSD components with strongest alpha-band peaks, as well as to the locations of those dipolar brain electrical sources best explaining the estimated patterns or filters. We observe that all filters are characterized by a high-frequency spatial structure. Strongest weights are generally observed at central electrodes. In contrast, activation patterns are much smoother, and cover more diverse areas involving also occipital electrodes.

As mentioned earlier, performing inverse source reconstruction from filter weight vectors is conceptually wrong, which becomes also evident from our results. Precisely, we observe that filter weight vectors cannot be well explained by single dipolar brain sources, the correlation of the filter and the EEG scalp potential generated by the best-fitting dipole lying only in the range from $r = 0.27$ to $r = 0.56$. As a result of the neurophysiologically-implausible high-frequent spatial structure of the filters, all dipolar sources are localized in the most superficial brain areas, and would be located in the skull or even skin compartments, if those were included in the search space.

Applying inverse source reconstruction to activation patterns yields meaningful results. We observe that all patterns are almost perfectly explained by a single dipolar electrical brain source, with correlation scores ranging from $r = 0.96$ to $r = 0.98$. Specifically, we find two lateralized central sources, two lateralized occipital sources, and one deep occipital source. These findings are consistent with the literature on so-called mu-rhythm oscillations in the motor system, as well as on alpha-oscillations in the visual system (see, e.g., Niedermeyer & Da Silva, 2005).

fMRI data

Fig. 6 shows an example of filters and patterns obtained from simultaneous recordings of spontaneous activity in the anesthetized macaque monkey. Experimental details are described in Murayama et al. (2010). Filters were estimated using temporal kernel CCA, see Eq. (10), and patterns were obtained using Eq. (6). The filters $\mathbf{w}_x(\tau)$, $\mathbf{w}_y$ (the canonical directions) for both temporally embedded electrophysiological spectrograms and fMRI, respectively, are plotted in the right panels. Both show high weights at the relevant locations of the input space: The fMRI filter $\mathbf{w}_y$ exhibits large positive coefficients around the

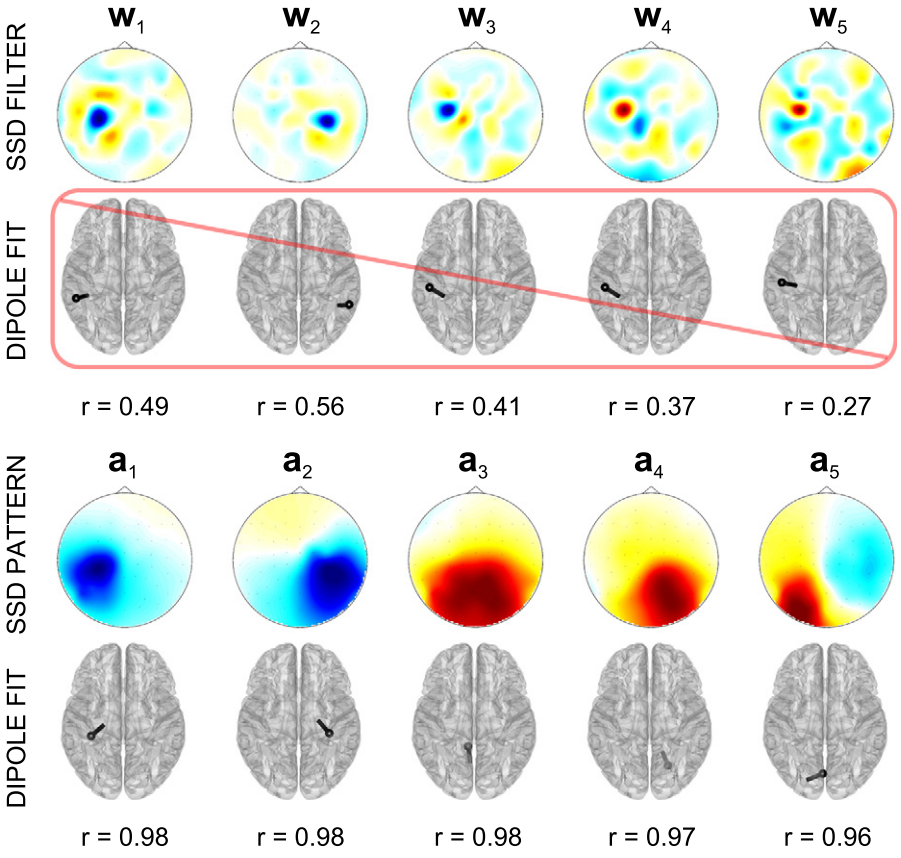

Fig. 5. Spatio-spectral decomposition (SSD) activation patterns and extraction filters of alpha-band EEG components during rest, along with their corresponding approximations by a single dipolar electrical sources in the brain. All activation patterns can be almost perfectly explained ($r > 0.96$) by single dipolar sources representing generators of central mu-rhythms in the motor cortex, as well as alpha-rhythm generators in the visual cortex. Unlike patterns, filters do not reflect the physical process of EEG generation. Therefore, applying dipole fits, as done here for demonstration purposes, is wrong, which is indicated by the red warning sign. This is also evidenced by the poor approximation ($r < 0.56$). Notably, all filters exhibit largest weights at sensors over the central areas, although some of the corresponding extracted EEG components (3–5) originate from occipital areas.

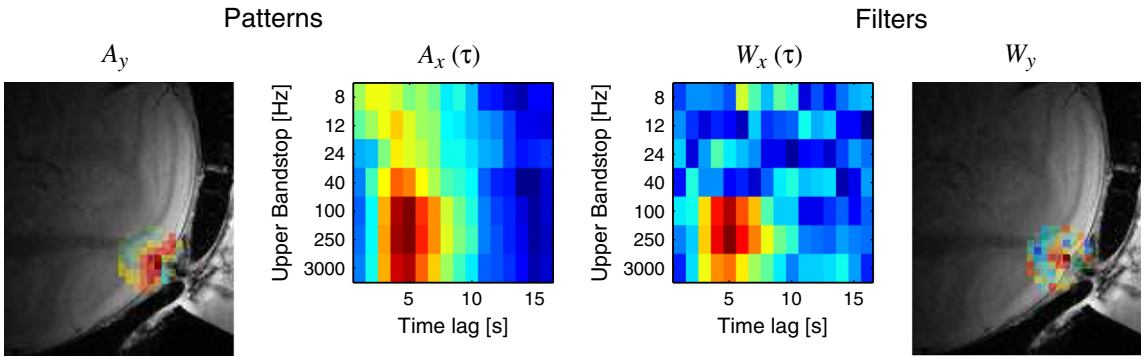

Fig. 6. Filters and patterns for simultaneous recordings of high-resolution fMRI and intracranially recorded neural spectrograms; filters were computed by tkCCA according to Eq. (10), patterns according to Eq. (6); filters $\mathbf{w}_x(\tau)$, $\mathbf{w}_y$ are not the same as $\mathbf{a}_x(\tau)$, $\mathbf{a}_y$. The patterns $\mathbf{a}_x(\tau)$, $\mathbf{a}_y$ reflect the correlation structure inherent to the fMRI data.

recording electrode and the convolutive filter $\mathbf{w}_x(\tau)$ for the electrophysiological spectrogram has the highest weights at time lags of approximately 5 s and at frequencies in the high gamma range. These are the time–frequency features of neural oscillations that are known to be maximally correlated with the fMRI signal (Goense & Logothetis, 2008; Logothetis et al., 2001; Niessing et al., 2005). Note however that the filter structure looks somewhat noisy. In particular the fMRI filter does not show the smooth structure that one would expect given the spatiotemporal point spread function reported in other studies investigating the hemodynamic response (Sirotin et al., 2009). In contrast the structure of the corresponding 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 $a_x(\tau)$ reflect much more clearly the physiology of the neurovascular response that has been reported in other studies. The temporal profile shows a clear peak at 5 s and a later undershoot at about 15 s. The frequency profile still indicates that the strongest hemodynamic response is in the high gamma range but the overall structure now is much clearer than in the case of the filter $\mathbf{w}_x(\tau)$.

Discussion

Backward models can be made interpretable

We have demonstrated that extraction filters of backward models may exhibit large weights at channels not at all picking up the signals-of-interest, as well as small weights at channels containing the signal. Such "misleading" weights are by no means indications of suboptimal model estimation. Rather, they are needed to "filter away" noise and thereby to extract the signal with high SNR. In our simulation, we obtained one additional patch of spurious activity in most extraction filters as a result of the particular choice of a specific noise (distractor) component with specific spatial mixing coefficients. In general, however, filters may deviate arbitrarily much from the true patterns depending on the noise covariance structure. Thus, one can easily also construct examples in which filters have zero weights at most task-relevant channels, or even large weights the sign of which opposes the sign of the true activation.

We derived a transformation by which extraction filters of any linear backward model can be turned into activation patterns of a corresponding forward model. By this means, backward models can eventually be made interpretable.

Whether or not the resulting patterns correspond to those of "true" signals contained in the data, however, depends on the accuracy with which these signals are extracted in the initial backward modeling step. Suboptimal factor extraction will naturally lead to suboptimal pattern reconstruction, since residual noise contributions in a factor estimate will cause noise-related channels to light up also in the corresponding pattern. The problem of "optimal" factor extraction under various conditions, however, is its own field of research treated in an extensive existing body on linear decomposition methods and their underlying assumptions (for some overview see Comon & Jutten, 2010; Golub & Van Loan, 1996; Haufe, 2011). In light of these considerations, activation patterns (no matter whether they are estimated by fitting a forward model, or indirectly using Eq. (6)) should be evaluated only in conjunction with their corresponding factors. In supervised scenarios, for example, an unbiased estimate of the decoding accuracy achieved by an extracted factor may give a good indication about whether its corresponding pattern should be interpreted at all.

In our simulations, where we had sufficient amounts of data to robustly estimate the involved backward models and empirical covariance matrices, the patterns estimated using Eqs. (6) and (9) (LASSO) reflect the spatial structure of the simulated factors well, in contrast to the original filters. Patterns were moreover found to be much more stable than filters. This is explained by the dependence of filters on the noise covariance, which was simulated to vary to some extent here. In practice, we frequently observe that even for datasets recorded from the same subject under the same paradigm during the same session, considerably different filters are obtained, while the corresponding estimated patterns remain relatively stable. Thus, statistical processing such as averaging is better justified for (appropriately normalized) patterns than for filters. Moreover, since the mapping from filters to patterns is bijective, extraction filters $\mathbf{W}$ to a given set of activation patterns $\mathbf{A}$ may be obtained by rearranging Eq. (6) for $\mathbf{W}$. Doing so may be particularly useful if $\mathbf{A}$ is known, e.g., has been derived from a computational model, or pre-estimated from different data. In both cases, the resulting $\mathbf{W}$ will be the filter which optimally extracts the sources with pattern $\mathbf{A}$ given the covariance structure of the (new) data. In the first case, $\mathbf{W}$ can be thought of as a *beamformer*. In the latter case, $\mathbf{W}$ can be a better estimate than a filter obtained directly from the original data, if we assume that both datasets share the same signal sources, but contain different noise sources.

Correlated brain processes

An ubiquitous phenomenon in real data is correlations either between brain processes or between target variables. Most analyses discussed here are incapable of separating multiple collinear components. This includes weight vectors of forward models, prediction accuracies achieved by encoding or decoding models, prediction accuracies achieved by searchlight approaches (see Relationship between searchlights and patterns section), and topographic maps of the univariate correlation/covariance of each channel with a target variable. For these methods, the activation maps related to a specific brain process will typically also highlight channels related to other processes, if these processes are correlated to the process under study. However,

the interpretability of these activation maps is not compromised by the multicollinearity issue, since correlated components are actually heavily statistically dependent.

Technically, of course, empirical correlations in data may arise for various reasons. First, two or more brain processes might be "naturally" co-activated through the way brain processing is organized. Second, "artificial" correlations of two or more functionally unrelated brain processes may be induced by means of correlations of respective external stimuli triggering activity of these processes, which could be an indication of improper experimental design. Unfortunately, however, natural and artificial correlations cannot be distinguished in practice by means of data analysis.

Importantly, the crucial qualitative distinction between weight vectors of forward and backward models remains even in the presence of correlated components. While for all models it holds that correlated components cannot be disentangled, backward models may additionally give significant large weights also to channels lacking any form of statistical dependence to any brain process of interest.

Practical implications for neuroimaging studies

Our analysis of real EEG and fMRI data revealed qualitative differences between the extraction filters and their activation pattern counterparts. While for EEG data this was expected as a results of volume conduction in the head (see The relevance of linear models for EEG data section), our results indicate that overlap of the activation patterns of, e.g., the signal of interest and other spatially correlated distractor signals, occurs also for fMRI data. Thus, the distinction between patterns and filters is crucial for the interpretation of fMRI decoding models, and the considered way of estimating patterns is of practical relevance for achieving interpretability on such data.

Importantly, since forward and backward models are dual to each other, the question whether a model should be regarded as a forward or a backward model entirely depends on which variables are independent, i.e., the quantities one wants to make inference about. We call these variables "data", while any experimentally controlled variables are called "target". Interpretability of weight vectors always requires a forward model of the data. While in this paper, we assume that the neuroimaging recordings are independent variables, one might in other contexts such as optogenetics (Williams & Deisseroth, 2013) actually experimentally control brain activity, and analyze its effects on, e.g., behavioral measurements. In this case, the neuroimaging data takes the role of the target variable, and a forward model of the behavioral data is needed to achieve interpretability of the model parameters with respect to these behavioral measurements.

Regularization does not make backward models better interpretable

If the number of parameters of a model is large compared to the amount of available data, the parameter assignment best explaining the data might not generalize well, i.e., explain new data not as well as the data used for fitting the model. This indicates that the relevant aspects of the data have not been captured by the model due to a lack of data. In order to improve the generalization performance, constraints on the simplicity of the model parameters are usually imposed, which is called *regularization*.

Sparsity of extraction filters does not imply sparsity of activation patterns

Ideally, such constraints should encode prior assumptions on the parameters' distribution, which may be application-specific. In neuroimaging, such assumptions may, for example, refer to the spatial structure of the extracted factors, and formalize the preference for certain brain regions being estimated as active, or the belief that the brain activation map has a certain sparsity structure or is sparse in general. Whether and how easily these assumptions can be integrated in the modeling, however, depends on the type of model used. For forward models, constraints on the structure of the extracted factors may be directly imposed on the mixing coefficients **A**. In backward modeling, however, we have no direct access to **A**. Imposing a certain structure on the demixing coefficients **W**, however, does not at all translate into imposing that exact structure on the factors extracted by **W**. The *effective* assumption imposed on the factors by estimating penalized backward models is hard to assess, and may be different depending on structural features of the data covariance matrix. In particular, sparse filters may actually extract factors contributing to many channels, while non-sparse filters may extract factors which are only expressed in a single channel. A consequence of the results of this paper might therefore be to impose "physiologically-motivated" structural constraints on $\mathbf{A} = \Sigma_x \mathbf{W} \Sigma_s^{-1}$ rather than on **W** in backward modeling approaches.

Regularization is indispensable in case of few data

Importantly, the above considerations do not indicate that regularization of backward models is inappropriate, nor do they imply that the choice of an "improper" regularizer will necessarily spoil the estimation of the corresponding pattern or even of the extracted factor. On the contrary, the pattern approximation quality does not depend on the structure of the filters, but only on the accuracy with which the underlying factor is estimated. To warrant good reconstruction of $\mathbf{s}(n)$, regularized backward modeling is often helpful, as long as the amount of regularization is adjusted in a statistically sound way (see, e.g., Lemm et al., 2011). Note that, in a similar way as for filter estimation, the estimation of corresponding patterns might benefit from regularization, which could be employed within the regression framework outlined in Eq. (8).

Relationship between searchlights and patterns

Complementary to our approach there are a number of other methods to overcome the problem of interpretability of multivariate classifiers in supervised neuroimaging analyses. A particularly successful solution is the searchlight approach (e.g., Chen et al., 2011; Kriegeskorte et al., 2006), which provides a tradeoff between estimation accuracy and localizability of neural sources. In the searchlight approach, a backward model is estimated within a small volume of brain tissue centered around each voxel successively instead of fitting a single backward model to the entire brain volume. The accuracy with which brain states can be estimated ("decoded") in a particular searchlight is interpreted — not the parameters of the single backward models. While the parameters of a backward model are generally not interpretable, the accuracies are, in the sense that they indicate the presence of class-specific information somewhere in the region. Searchlights and the approach presented here serve a similar purpose. Theoretically, the approach presented here is applicable to the searchlights, too. Averaging across all searchlights then will result in a smoothed version of the pattern **A**.

Generalizations

The considered approach for estimating patterns can be generalized in various ways. We here assumed linear forward and backward models, although nonlinear models are conceivable as well. If nonlinear models are involved, an analytic transformation as for the linear/linear case will likely be obtained only in very special cases. However, for any combination of (linear or nonlinear) forward and backward models with the same number of latent factors, K, a regression approach as outlined in the end of the Interpreting results of backward modeling: obtaining activation patterns from extraction filters section may be adopted. Depending on the type of forward model used, the estimation here may either lead to an analytic solution, or require numerical optimization. A general framework for parameter interpretation, which includes the linear/linear case considered here as a special case, is the feature importance ranking measure (FIRM) by Zien et al. (2009). FIRM is discussed in detail in Appendix D.

Activation patterns obtained from multivariate OLS decoding are equivalent to a mass-univariate analysis

A particularly simple expression for the activation patterns is obtained for decoding approaches fitting the data linearly onto a set of external target variables via ordinary least-squares regression. Notably, this also comprises the LDA classifier, which, in the binary case, can be formulated as an OLS regression with a particular choice of the target variable $\mathbf{y}(n)$ (see Appendix C).

In OLS decoding, both the extraction filters and their corresponding activation pattern estimates may be derived analytically. Denoting the targets by $\mathbf{y}(n)$, and assuming zero mean targets $\mathbf{y}(n)$ and observations $\mathbf{x}(n)$, the filters are given by

$$\begin{aligned}\mathbf{W} &= arg\min_{\widetilde{\mathbf{W}}} \sum_n \left(\widetilde{\mathbf{W}}^\top \mathbf{x}(n) - \mathbf{y}(n)\right)^2 \\ &= \Sigma_x^{-1} Cov[\mathbf{x}(n), \mathbf{y}(n)].\end{aligned} \tag{11}$$

Inserting into Eq. (6) yields the pattern estimate

$$\begin{aligned}\mathbf{A} &= \Sigma_x \mathbf{W} \Sigma_{\hat{s}}^{-1} = \Sigma_x \Sigma_x^{-1} Cov[\mathbf{x}(n), \mathbf{y}(n)] \Sigma_{\hat{s}}^{-1} \\ &= Cov[\mathbf{x}(n), \mathbf{y}(n)] \Sigma_{\hat{s}}^{-1},\end{aligned} \tag{12}$$

which, for uncorrelated factors, and for $K = 1$, is proportional to the mass-univariate covariance between channel readings and target variables. Thus, while OLS regression/LDA classification of course does improve the estimation of a target variable, it does not increase the interpretability of the results upon what can be inferred from mass-univariate analysis. This equivalence, however, does not hold exactly for decoding approaches involving regularization or non-quadratic loss functions.

Note, however, that there are cases when the target variable is not known and thus mass-univariate analyses cannot be applied directly. In unsupervised analyses for instance, the source variable is unknown. While it is possible to estimate $\mathbf{A}$ and $\mathbf{s}(n)$ simultaneously, it has computational advantages to first obtain an estimate of $\mathbf{s}(t)$ by fitting a backward model and then apply the proposed method to obtain $\mathbf{A}$.

Mass-univariate analysis: correlation vs. covariance

The dependence on the SNR makes it generally difficult to assess the activation patterns of the signal components from mass-univariate correlation/class-separability maps. However, there are also situations in which it is reasonable to look at correlation measures instead of plain covariance. For example, EEG recordings may contain a weak neuronal signal, which is highly correlated to the target variable, while at the same time there may be artifactual activity of much larger amplitude. If the artifacts, however, also exhibit a nonzero (possibly small or even insignificant) correlation with the target variable (that is, the uncorrelatedness of signal and noise components assumed in this paper is violated), the mass-univariate covariance pattern might show stronger peaks at channels picking up those artifacts than at channels containing the neuronal signal-of-interest. While this is technically not a wrong result, one may want to highlight only those channels in which the task-related signal is both strong and highly correlated with the task, which can be done by looking at correlation/class separability maps, or by performing searchlight analyses (Haynes et al., 2007; Kriegeskorte et al., 2006).

The relevance of linear models for EEG data

Traditionally, the factors $\mathbf{s}$ of the linear forward model are thought of as variables aggregating and isolating the problem-specific information, while the corresponding activation patterns $\mathbf{a}$ model the expression of each factor in each channel. In the particular case of EEG (and magnetoencephalography, MEG) data, the linear model moreover accurately describes the physical process of data generation. Electroencephalography, for example, measures neuronal electrical activity in the brain indirectly as surface potentials on the scalp. Due to volume conduction in the head, the electrical signals are transformed on their way from brain sources to EEG channels. For the frequencies below 1 kHz (which are of highest interest in EEG analysis), this transformation is mainly an instantaneous spatial blurring. Since contributions from different brain areas add up linearly at the channel level, the entire physical process of signal propagation can be accurately modeled by the linear forward model

$$\mathbf{x}(n) = \mathbf{L}\mathbf{s}_v(n) + \epsilon(n), \tag{13}$$

where $\mathbf{s}_v(n)$ is the electrical activity at $N_v \gg K$ voxel in the brain and $\mathbf{L}$ is the *leadfield* matrix modeling the propagation of electrical activity from brain voxels to EEG sensors (e.g., Baillet et al., 2001). Methods estimating the parameters of a linear forward model $\mathbf{x}(n) = +\varepsilon(n)$ on EEG data may therefore be re-parametrized using $\mathbf{L}$ and $\mathbf{s}_v(n)$. By decomposing $\mathbf{A}$ into $\mathbf{A} = \mathbf{LP}$, Eq. (13) is recovered with $\mathbf{s}_v(n) = \mathbf{P}\mathbf{s}(n)$. Here, $\mathbf{P}$ is a $3N_v \times K$ matrix of current sources densities indicating the brain voxels, which generate the activity of each of the individual K latent factors. Consequently, it is possible to interpret the extracted factors $s_k(n)$ as estimates of the neural activity of specific local (or distributed) brain networks, while the corresponding mixing coefficients $\mathbf{a}_k$ describe the field spread of these brain electrical sources or networks in the particular head. The process of estimating the factorization $\mathbf{A} = \mathbf{LP}$[4] for given $\mathbf{A}$ and $\mathbf{L}$ is called EEG *inverse source reconstruction*, and is usually based on prior assumptions on the spatial structure of the estimated current densities $\mathbf{P}$. The Spatio-spectral decomposition of EEG data and EEG data sections contain an example, in which mixing patterns $\mathbf{a}$ are used to estimate the actual locations of the generating electrical brain sources under the assumption that each latent factor's activity originates only from a single brain voxel.

As a result of the fact that a linear forward model holds for raw EEG data it follows that linear modeling of nonlinearly preprocessed EEG data cannot be used to recover the true underlying brain sources (see, e.g., Dähne et al., 2014).

Conclusions

We have shown that the parameters of multivariate backward/decoding models (called extraction filters) cannot be interpreted in terms of the brain activity of interest alone, because they depend on all noise components in the data, too. In the neuroimaging context, this implies that no neurophysiological conclusions may be drawn from the parameters of such models. Moreover – in contrast to what may be a widespread intuition – the interpretability of the parameters cannot be improved by means of regularization (e.g., sparsification) for such models. However, as we pointed out, there is a simple procedure for transforming backward models into forward models. The parameters (called activation patterns) of forward models allow the exact desired interpretation. We demonstrated on simulated data, as well as on real fMRI and EEG data, that the analysis of extraction filters may lead to severe misinterpretation in practice, while the proposed way of analyzing activation patterns resolves the problem. Our results are not restricted to the neuroimaging context, but hold for any application, in which parameters of backward, e.g., regression or classification models are typically interpreted as properties of the extracted signals. Yet, the data analysis methods covered in this paper are of course only a tiny fraction of what is being used in practice. Generally, we encourage authors to test sophisticated methods in realistic simulations to better

[4] In practice, each pattern $\mathbf{a}_k$ is usually approximated by $\mathbf{a}_k = \mathbf{L}\mathbf{p}_k + \varepsilon_k$, where $[\mathbf{p}_1, \mathbf{p}_K] = \mathbf{P}$.

understand their properties before applying them to real data, a point also made by Haufe et al. (2012) in the context of causal estimation.

Acknowledgments

We thank Eike A. Schmidt for providing the EEG data, and Marius Kloft, Klaus-Robert Müller, Jukka-Pekka Kauppi, Aapo Hyvärinen, Nikolaus Kriegeskorte and an anonymous reviewer for valuable discussions. We acknowledge financial support by the German Bundesministerium für Bildung und Forschung (Grant Nos. 01GQ0850/851, 16SV5839, 01GQ0411, 01GQ1001B and 01GQ1001C), the Deutsche Forschungsgemeinschaft (GRK1589/1), the WCU (World Class University) program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0), and the Brain Korea 21 PLUS Program through the National Research Foundation of Korea funded by the Ministry of Education.

Appendix A. Proof of Theorem 1 (activation patterns of the corresponding forward model)

Assuming the existence of a corresponding forward model (see Appendix B for a proof of the existence), we can insert Eq. (3) into Eq. (5) to obtain

$$\begin{aligned}\hat{\mathbf{s}}(n) &\overset{(5)}{=} \mathbf{W}^\top \mathbf{x}(n)\\ &\overset{(3)}{=} \mathbf{W}^\top(\mathbf{A}\hat{\mathbf{s}}(n)+\epsilon(n))\\ &= \mathbf{W}^\top\mathbf{A}\hat{\mathbf{s}}(n)+\mathbf{W}^\top\epsilon(n).\end{aligned}$$

Multiplying that equation with $\hat{\mathbf{s}}(n)^\top$ from the right, and taking the expected value over samples yields

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{s}}(n)\hat{\mathbf{s}}(n)^\top\right]_n &= \mathbb{E}\left[\mathbf{W}^\top\mathbf{A}\hat{\mathbf{s}}(n)\hat{\mathbf{s}}(n)^\top\right]_n + \mathbb{E}\left[\mathbf{W}^\top\epsilon(n)\hat{\mathbf{s}}(n)^\top\right]_n\\ &= \mathbf{W}^\top\mathbf{A}\mathbb{E}\left[\hat{\mathbf{s}}(n)\hat{\mathbf{s}}(n)^\top\right]_n + \mathbf{W}^\top\mathbb{E}\left[\epsilon(n)\hat{\mathbf{s}}(n)^\top\right]_n\\ &\overset{(4)}{=} \mathbf{W}^\top\mathbf{A}\mathbb{E}\left[\hat{\mathbf{s}}(n)\hat{\mathbf{s}}(n)^\top\right]_n.\end{aligned}$$

Since the matrix $E\left[\hat{\mathbf{s}}(n)\hat{\mathbf{s}}(n)^\top\right]_n$ has full rank (due to the factors in $\hat{\mathbf{s}}(n)$ being linearly independent), we can conclude that

$$\mathbf{W}^\top\mathbf{A} = \mathbf{I}. \tag{A.1}$$

Similarly, by inserting Eq. (5) into Eq. (3), we obtain

$$\mathbf{x}(n) \overset{(3)}{=} \mathbf{A}\hat{\mathbf{s}}(n)+\epsilon(n) \overset{(5)}{=} \mathbf{A}\mathbf{W}^\top\mathbf{x}(n)+\epsilon(n).$$

Rearranging for $\epsilon(n)$ yields

$$\epsilon(n) = \mathbf{x}(n)-\mathbf{A}\mathbf{W}^\top\mathbf{x}(n) = \left(\mathbf{I}-\mathbf{A}\mathbf{W}^\top\right)\mathbf{x}(n)\ ,$$

which, when multiplied with $\mathbf{W}^\top$ from the left leads to

$$\begin{aligned}\mathbf{W}^\top-\epsilon(n) &= \mathbf{W}^\top\left(\mathbf{I}-\mathbf{A}\mathbf{W}^\top\right)\mathbf{x}(n) = \left(\mathbf{W}^\top-\mathbf{W}^\top\mathbf{A}\mathbf{W}^\top\right)\mathbf{x}(n)\\ &\overset{(A.1)}{=} \left(\mathbf{W}^\top-\mathbf{W}^\top\right)\mathbf{x}(n) = 0.\end{aligned} \tag{A.2}$$

Furthermore, it follows from Eqs. (3) and (4) that $\Sigma_{\mathbf{x}} = \mathbf{A}\Sigma_{\hat{\mathbf{s}}}\mathbf{A}^\top + \Sigma_\epsilon$, which leads to

$$\begin{aligned}\Sigma_{\mathbf{x}}\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1} &= \left(\mathbf{A}\Sigma_{\mathbf{s}}\mathbf{A}^\top+\Sigma_\epsilon\right)\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1} = \mathbf{A}\Sigma_{\mathbf{s}}\mathbf{A}^\top\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1}+\Sigma_\epsilon\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1}\\ &\overset{(A.1)}{=} \mathbf{A}\Sigma_{\hat{\mathbf{s}}}\Sigma_{\hat{\mathbf{s}}}^{-1}+\Sigma_\epsilon\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1} \overset{(A.2)}{=} \mathbf{A}+0\Sigma_{\hat{\mathbf{s}}}^{-1} = \mathbf{A}.\end{aligned}$$

This concludes the proof.

Appendix B. Existence of a corresponding forward model for a given backward model

Here we show that the definition of $\mathbf{A}$ given in Eq. (6) provides a forward model that corresponds to the backward model in Eq. (5). Due to the assumption that the factors $\hat{\mathbf{s}}$ are linearly independent, their covariance matrix $\Sigma_{\hat{\mathbf{s}}}$ is invertible, such that the definition of $\mathbf{A}$ in Eq. (6) is well-defined.

It remains to show that the noise term

$$\epsilon(n) := \mathbf{x}(n)-\mathbf{A}\hat{\mathbf{s}}(n) \tag{B.1}$$

satisfies condition (4) of being uncorrelated with the factors. Using

$$\begin{aligned}\epsilon(n) &\overset{(B.1)}{=} \mathbf{x}(n)-\mathbf{A}\hat{\mathbf{s}}(n)\\ &\overset{(6)}{=} \mathbf{x}(n)-\Sigma_{\mathbf{x}}\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1}\hat{\mathbf{s}}(n)\end{aligned}$$

we can conclude that

$$\begin{aligned}\mathbb{E}\left[\epsilon(n)\hat{\mathbf{s}}(n)^\top\right]_n &= \mathbb{E}\left[\mathbf{x}(n)\hat{\mathbf{s}}(n)^\top\right]_n-\mathbb{E}\left[\Sigma_{\mathbf{x}}\mathbf{W}\Sigma_{\hat{\mathbf{s}}}^{-1}\hat{\mathbf{s}}(n)\hat{\mathbf{s}}(n)^\top\right]_n\\ &= \mathbb{E}\left[\mathbf{x}(n)\hat{\mathbf{s}}(n)^\top\right]_n-\mathbb{E}\left[\mathbf{x}(n)\mathbf{x}(n)^\top\mathbf{W}\right]_n\\ &\overset{(5)}{=} \mathbb{E}\left[\mathbf{x}(n)\hat{\mathbf{s}}(n)^\top\right]_n-\mathbb{E}\left[\mathbf{x}(n)\hat{\mathbf{s}}(n)^\top\right]_n\\ &= 0.\end{aligned}$$

Appendix C. Example: linear discriminant analysis

As an example of how extraction filters depend on the spatial structure of the noise, consider the case of linear discriminant analysis (LDA) classification (Fisher, 1936), which is the Bayes optimal classification rule for Gaussian distributed classes with equal covariance matrices. Under these assumptions (and provided that the inverse of the empirical covariance matrix exists), optimal separation of two classes is given by evaluating $\mathbf{w}^\top\mathbf{x}(n) < c$ for some constant c, where $\mathbf{w} = \Sigma_x^{-1}(\mu_{\mathbf{x}}^+ - \mu_{\mathbf{x}}^-)$, and where $\mu_{\mathbf{x}}^{+/-} \in \mathbb{R}^M$ and $\Sigma_x \in \mathbb{R}^{M\times M}$ denote the classwise means and the common covariance of the observed samples in channel space. Now it is obvious that, while the discriminating factor $\hat{\mathbf{s}}(n) = \mathbf{w}^\top\mathbf{x}(n)$ is only one-dimensional, the filter $\mathbf{w}$ depends on the covariance structure of the entire M-dimensional dataset through Σ_x^{-1}, and thus on all noise sources as well as on all other latent factors.

Applying Eq. (6) to the LDA filter, we note that the resulting pattern simplifies to $\hat{\mathbf{a}} = \Sigma_x\mathbf{w}Var[\hat{\mathbf{s}}(n)]\Sigma_x\Sigma_x^{-1}(\mu_{\mathbf{x}}^+-\mu_{\mathbf{x}}^-) = \mu_{\mathbf{x}}^+-\mu_{\mathbf{x}}^-$, i.e., is proportional to the difference of the means of the two classes. This simple mass-univariate structure is explained by the fact that LDA can be regarded as a special case of OLS regression (cf., Activation patterns obtained from multivariate OLS decoding are equivalent to a mass-univariate analysis section), where the target is a univariate binary variable indicating the class membership. By setting $\tilde{y}(n) = 1/N^+$ for samples of the positive class, and $\tilde{y}(n) = -1/N^-$ for samples of the negative class, where N^+ and N^- are the numbers of samples per class with $N^+ + N^- = N$, the OLS filter $\mathbf{w}_{\mathrm{OLS}} = \Sigma_x^{-1}Cov[\mathbf{x}(n),\tilde{y}(n)] = \Sigma_x^{-1}(\mu_{\mathbf{x}}^+-\mu_{\mathbf{x}}^-)$ coincides with the LDA weight vector.

To see how regularization changes the estimated patterns consider the case of regularized LDA (RLDA), which is an instance of *Ridge regression*. The Ridge regression estimator of $\mathbf{w}$ is defined as $\mathbf{w}_{\mathrm{Ridge}} = argmin_{e\mathbf{w}}\sum_n\left(\tilde{\mathbf{w}}^\top\mathbf{x}(n)-\tilde{y}(n)\right)^2+\lambda\|\tilde{\mathbf{w}}\|^2 = (\Sigma_x+\lambda\mathbf{I})^{-1}(\mu_{\mathbf{x}}^+-\mu_{\mathbf{x}}^-)$, while the corresponding pattern estimate is $\hat{\mathrm{a}} = \Sigma_x(\Sigma_x+\lambda\mathbf{I})^{-1}(\mu_{\mathbf{x}}^+-\mu_{\mathbf{x}}^-)Var[\hat{\mathbf{s}}(n)]$. That is, for $\lambda\to\infty$ the RLDA pattern becomes proportional to the LDA pattern (i.e., class-mean difference) "smoothed" by the empirical data covariance.

Appendix D. Relation between patterns and the feature importance ranking measure

The feature importance ranking measure (FIRM) (Zien et al., 2009) was first proposed in the context of bioinformatics data (Rätsch et al., 2006) and later generalized to arbitrary data sets. The advantage of FIRM is that it is universal in the sense that it is defined for any algorithm (independent of loss function, regularizer or type of data) and it is objective in the sense that is invariant with respect to correlations between features or scaling of features. Given some learning algorithm with a scoring function $\rho(\mathbf{x}(n))$ FIRM for a feature $f(\mathbf{x}(n))$ is based on the *conditional expected score* q_f, i.e., the expectation of the score $\rho(\mathbf{x}(n))$ conditioned on the feature $f(\mathbf{x}(n))$ had a certain value t

$$q_f(t) = \mathbb{E}[\rho(\mathbf{x}(n))|f(\mathbf{x}(n)) = t]. \tag{D.1}$$

In the linear foward modeling setting $\rho(\mathbf{x}(n)) = \mathbf{w}^{\mathrm{T}}\mathbf{x}(n)$ and the features $f(\mathbf{x}(n))$ are the individual dimensions, that is the entries of $\mathbf{x}(n)$ corresponding to fMRI voxels or EEG electrodes. For each feature $f(\mathbf{x}(n))$, $qf(t)$ is the output of $\rho(\mathbf{x}(n))$ given that the feature $f(\mathbf{x}(n))$ has the value t. FIRM is now defined as the standard deviation of $q_f(t)$

$$\mathrm{FIRM} := \sqrt{Var\left(q_f(t)\right)}. \tag{D.2}$$

In other words FIRM measures how the output of a source estimation algorithm or a stimulus estimator changes when a given feature changes. For the case considered here, linear scoring functions $\rho(\mathbf{x}(n)) = \mathbf{w}^{\mathrm{T}}\mathbf{x}(n) = \hat{\mathbf{s}}(n)$, FIRM can be computed as

$$\mathrm{FIRM} := \mathbf{D}^{-1}\Sigma_x\left(n\hat{\Sigma}_x\right)^{-1}Cov[\mathbf{x}(n), \hat{\mathbf{s}}(n)] \tag{D.3}$$

where $\mathbf{D}$ is a diagonal matrix containing the standard deviations of the measured neuroimaging data $\mathbf{x}(n)$, Σ_x is the true covariance matrix of $\mathbf{x}(n)$, n is the number of samples and $\hat{\Sigma}_x$ is the empirical estimate of the covariance matrix of $\mathbf{x}(n)$. With enough data the empirical covariance matrix converges to the true covariance matrix and the terms cancel out. Assuming that the features have been normalized to have unit variance and assuming decorrelated sources (or univariate sources), FIRM is equivalent to the pattern obtained by Eq. (7). As FIRM is defined for arbitrary models, including non-linear models and discrete data, Eq. (D.2) can be applied to obtain meaningful patterns for these cases.

References

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

Bießmann, F., Meinecke, F.C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N., Müller, K.-R., 2009. Temporal kernel canonical correlation analysis and its application in multimodal neuronal data analysis. Mach. Learn. 79 (1–2), 5–27.

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

Bießmann, F., Dähne, S., Meinecke, F.C., Blankertz, B., Görgen, K., Haufe, S., 2012a. On the interpretability of linear multivariate neuroimaging analyses: filters, patterns and their relationship. Proceedings of the 2nd NIPS Workshop on Machine Learning and Interpretation in Neuroimaging (To appear).

Bießmann, F., Murayama, Y., Logothetis, N.K., Müller, K.-R., Meinecke, F.C., 2012b. Improved decoding of neural activity from fMRI signals using non-separable spatio-temporal deconvolutions. NeuroImage 61 (4), 1031–1042.

Blankertz, B., Curio, G., Müller, K.-R., 2002. Classifying single trial EEG: towards brain computer interfacing. In: Diettrich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Inf. Proc. Systems (NIPS 01), vol. 14, pp. 157–164.

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

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

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

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44 (1), 112–122 (Jan).

Chen, Y., Namburi, P., Elliott, L.T., Heinzle, J., Soon, C.S., Chee, M.W., Haynes, J.-D., 2011. Cortical surface-based searchlight decoding. NeuroImage 56 (2), 582–592.

Comon, P., 1994. Independent component analysis, a new concept? Signal Process. 36, 287–314.

Comon, P., Jutten, C., 2010. Handbook of Blind Source Separation: Independent Component Analysis and Applications. Independent Component Analysis and Applications SeriesElsevier Science.

Dähne, S., Meinecke, F.C., Haufe, S., Höhne, J., Tangermann, M., Müller, K.-R., Nikulin, V.V., 2014. SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. NeuroImage 86, 111–122.

Dolce, G., Waldeier, H., 1974. Spectral and multivariate analysis of EEG changes during mental activity in man. Electroencephalogr. Clin. Neurophysiol. 36 (6), 577–584 (Jun).

Donchin, E., Heffley, E.F., 1978. Multivariate Analysis of Event-related Potential Data: A Tutorial Review. Multidisciplinary Perspectives in Event-related Brain Potential Research. 555–572.

Dornhege, G., del R. Millán, J., Hinterberger, T., McFarland, D., Müller, K.-R. (Eds.), 2007. Toward Brain–Computer Interfacing. MIT Press, Cambridge, MA.

Fisher, R., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54 (1), 313–327.

Friston, K.J., 2009. Modalities, modes, and models in functional neuroimaging. Science 326 (5951), 399–403.

Friston, K., Jezzard, P., Turner, R., 1994. Analysis of functional MRI time-series. Hum. Brain Mapp. 1, 153–171.

Goense, J.B.M., Logothetis, N.K., 2008. Neurophysiology of the bold fMRI signal in awake monkeys. Curr. Biol. 18 (9), 631–640 (May).

Golub, G., Van Loan, C., 1996. Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Gramfort, A., Strohmeier, D., Haueisen, J., Hamalainen, M.S., Kowalski, M., 2013. Time–frequency mixed-norm estimates: sparse M/EEG imaging with non-stationary source activations. NeuroImage 70, 410–422 (Apr).

Haufe, S., 2011. Towards EEG Source Connectivity Analysis. (Ph.D. thesis) Berlin Institute of Technology.

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

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

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

Haufe, S., Tomioka, R., Dickhaus, T., Sannelli, C., Blankertz, B., Nolte, G., Müller, K.-R., 2011. Large-scale EEG/MEG Source Localization with Spatial Flexibility. 851–859.

Haufe, S., Nikulin, V.V., Müller, K.-R., Nolte, G., 2012. A critical assessment of connectivity measures for EEG data: a simulation study. NeuroImage 64, 120–133.

Haynes, J.D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. 17 (4), 323–328 (Feb).

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. 10 (3), 626–634.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. Adaptive and Learning Systems for Signal Processing, Communications and Control Series.Wiley.

Hyvärinen, A., Hurri, J., Hoyer, P., 2009. Natural Image Statistics. Computational Imaging and Vision, 39. Springer London, Limited.

Koles, Z.J., Lind, J.C., Soong, A.C., 1995. Spatio-temporal decomposition of the EEG: a general approach to the isolation and localization of sources. Electroencephalogr. Clin. Neurophysiol. 95 (4), 219–230 (Oct).

Kragel, P.A., Carter, R.M., Huettel, S.A., 2012. What makes a pattern? Matching decoding methods to data in multivariate pattern analysis. Front. Neurosci. 6 (162).

Kriegeskorte, N., 2011. Pattern-information analysis: from stimulus decoding to computational-model testing. NeuroImage 56 (2), 411–421.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103 (10), 3863.

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

Logothetis, N.K., Pauls, J., Augath, M.A., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412 (6843), 150–157.

Luck, S., 2005. An Introduction to the Event-related Potential Technique. Cognitive NeuroscienceMIT Press.

Misaki, M., Kim, Y., Bandettini, P., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage 53 (1), 103–118.

Murayama, Y., Bießmann, F., Meinecke, F.C., Müller, K.-R., Augath, M.A., Oeltermann, A., Logothetis, N.K., 2010. Relationship between neural and hemodynamic signals during spontaneous activity studied with temporal kernel CCA. Magn. Reson. Imaging 28 (8), 1095–1103 (Jan).

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56 (2), 400–410 (May).

Niedermeyer, E., Da Silva, F., 2005. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Doody's All Reviewed Collection.Lippincott Williams & Wilkins.

Niessing, J., Ebisch, B., Schmidt, K.E., Niessing, M., Singer, W., Galuske, R.A.W., 2005. Hemodynamic signals correlate tightly with synchronized gamma oscillations. Science 309 (5736), 948–951 (Aug).

Nikulin, V.V., Nolte, G., Curio, G., 2011. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. NeuroImage 55 (4), 1528–1535 (Apr).
Nolte, G., Dassios, G., 2005. Analytic expansion of the EEG lead field for realistic volume conductors. Phys. Med. Biol. 50, 3807–3823.
Nolte, G., Meinecke, F.C., Ziehe, A., Müller, K.-R., 2006. Identifying interactions in mixed and noisy complex systems. Phys. Rev. E. 73, 051913.
Parra, L., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., Sajda, P., 2003. Single-trial detection in EEG and MEG: keeping it linear. Neurocomputing 52–54, 177–183 (Jun).
Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. NeuroImage 28 (2), 326–341.
Parra, L., Christoforou, C., Gerson, A., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M., Sajda, P., 2008. Spatiotemporal linear decoding of brain state. IEEE Signal Process. Mag. 25 (1), 107–115.
Pereda, E., Quiroga, R.Q., Bhattacharya, J., 2005. Nonlinear multivariate analysis of neurophysiological signals. Prog. Neurobiol. 77 (1), 1–37.
Rätsch, G., Sonnenburg, S., Schäfer, C., 2006. Learning interpretable SVMS for biological sequence classification. BMC Bioinforma. 7 (Suppl. 1), S9.
Schmidt, R.O., 1986. Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas. Propag. 43, 276–280.
Schmidt, E.A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., Kincses, W.E., 2009. Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving. Accid. Anal. Prev. 41, 1087–1093.
Shirer, W.R., Ryali, S., Rykhlevskaia, E., Menon, V., Greicius, M.D., 2012. Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cereb. Cortex 22 (1), 158–165 (Jan).
Sirotin, Y.B., Hillman, E.M.C., Bordier, C., Das, A., 2009. Spatiotemporal precision and hemodynamic mechanism of optical point spreads in alert primates. PNAS 106 (43), 18390–18395 (Oct).
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288.
Vega-Hernández, M., Martínez-Montes, E., Sánchez-Bornot, J., Lage-Castellanos, A., Valdés-Sosa, P.A., 2008. Penalized least squares methods for solving the EEG inverse problem. Stat. Sin. 18, 1535–1551.
von Bünau, P., Meinecke, F.C., Király, F., Müller, K.-R., 2009. Finding stationary subspaces in multivariate time series. Phys. Rev. Lett. 103, 214101.
Williams, S.C., Deisseroth, K., 2013. Optogenetics. Proc. Natl. Acad. Sci. U. S. A. 110 (41), 16287 (Oct).
Wolpaw, J.R., Wolpaw, E.W. (Eds.), 2012. Brain–Computer Interfaces: Principles and Practice. Oxford University Press.
Zien, A., Krämer, N., Sonnenburg, S., Rätsch, G., 2009. The feature importance ranking measure, In: Buntine, W., Grobelnik, M., Mladenic, D., Shawe-Taylor, J. (Eds.), Proceedings of the European Conference on Machine Learning, xxix edition. Vol. 5782/2009 of Lecture Notes in Artificial Intelligence. Springer, Berlin/Heidelberg, pp. 694–709.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Methodol. 67, 301–320.