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