NeuroImage 159 (2017) 449–458

Contents lists available at ScienceDirect

NeuroImage

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

Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis

Sarah Alizadeh [a,b,d,f], Hamidreza Jamalabadi [a,b,d,f], Monika Schönauer [a,b,c], Christian Leibold [b,e], Steffen Gais [a,b,c,*]

[a] Medical Psychology and Behavioral Neurobiology, University of Tübingen, Silcherstr. 5, 72076 Tübingen, Germany
[b] Bernstein Center for Computational Neuroscience, Ludwig-Maximilians-Universität München, Großhadernerstr. 2, 82152 Planegg-Martinsried, Germany
[c] Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 München, Germany
[d] IMPRS for Cognitive and Systems Neuroscience, University of Tübingen, Österbergerstr. 3, 72074 Tübingen, Germany
[e] Department of Biology II, Ludwig-Maximilians-Universität München, Großhadernerstr. 2, 82152 Planegg-Martinsried, Germany
[f] Department of Psychiatry, Division for Translational Psychiatry, University of Tübingen, Calwerstr. 14, 72076 Tübingen, Germany

ARTICLE INFO

Keywords:
Multivariate pattern analysis
Neuroimaging
Permutation statistics
Stimulus-related confounds
Concept-response curve

ABSTRACT

Multivariate pattern analysis (MVPA) methods are now widely used in life-science research. They have great potential but their complexity also bears unexpected pitfalls. In this paper, we explore the possibilities that arise from the high sensitivity of MVPA for stimulus-related differences, which may confound estimations of class differences during decoding of cognitive concepts. We propose a method that takes advantage of concept-unrelated grouping factors, uses blocked permutation tests, and gradually manipulates the proportion of concept-related information in data while the stimulus-related, concept-irrelevant factors are held constant. This results in a concept-response curve, which shows the relative contribution of these two components, i.e. how much of the decoding performance is specific to higher-order category processing and to lower order stimulus processing. It also allows separating stimulus-related from concept-related neuronal processing, which cannot be achieved experimentally. We applied our method to three different EEG data sets with different levels of stimulus-related confound to decode concepts of digits vs. letters, faces vs. houses, and animals vs. fruits based on event-related potentials at the single trial level. We show that exemplar-specific differences between stimuli can drive classification accuracy to above chance levels even in the absence of conceptual information. By looking into time-resolved windows of brain activity, concept-response curves can help characterize the time-course of lower-level and higher-level neural information processing and detect the corresponding temporal and spatial signatures of the corresponding cognitive processes. In particular, our results show that perceptual information is decoded earlier in time than conceptual information specific to processing digits and letters. In addition, compared to the stimulus-level predictive sites, concept-related topographies are spread more widely and, at later time points, reach the frontal cortex. Thus, our proposed method yields insights into cognitive processing as well as corresponding brain responses.

1. Introduction

Advances in electrophysiological, genetic, and neuroimaging methods generate ever growing volumes of data. These massively multivariate data sets require methods of analysis which go beyond traditional statistical ANOVA-based approaches (Haynes and Rees, 2006; O'Toole et al., 2007; Tong and Pratte, 2012). Particularly machine learning methods have seen growing adoption in the life sciences because they can be used to analyze high-dimensional data with great sensitivity (Norman et al., 2006; Haxby et al., 2014). In neuroimaging, multivariate pattern analysis (MVPA) has made it possible not only to investigate differences in brain regional activity during the performance of a task, but also to decode perceptual and mental representations as well as conceptual and semantic information (Kamitani and Tong, 2005; Kay et al., 2008; Mitchell et al., 2008; Schwarzlose et al., 2008; Rissman et al., 2010; Simanova et al., 2014).

The complexity of multivariate analysis, however, leads to unexpected problems (Todd et al., 2013; Woolgar et al., 2014; Haynes, 2015;

* Corresponding author. Institute for Medical Psychology and Behavioral Neurobiology, University of Tübingen, Silcherstr. 5, 72076 Tübingen, Germany.
E-mail address: steffen.gais@uni-tuebingen.de (S. Gais).

http://dx.doi.org/10.1016/j.neuroimage.2017.07.058
Received 16 December 2016; Received in revised form 26 May 2017; Accepted 28 July 2017
Available online 29 July 2017

Jamalabadi et al., 2016). Here, we will explore the consequences of the high sensitivity of MVPA for differences found between subgroups of trials in cognitive experiments. In classical analyses, two conditions with identical means are considered identical. Differences between trials (caused by different stimuli, subjects, etc.) usually average out on the dependent variable and therefore do not influence the group average. The multivariate nature of MVPA, however, allows differences to accumulate over dimensions (Fan and Fan, 2008; Jamalabadi et al., 2016). Any differences between individual elements of the categories will be used by MVPA to distinguish between categories, even if the categories themselves have identical centroids. For example, if concept-related features are the intended focus of study, different combinations of low-level, stimulus-specific features like orientation, shape, color, etc. can drive decoding although there is no overall average difference in these features between both concepts (Haynes and Rees, 2006). In fact, MVPA is sensitive to both the effect of interest and to any other confounding factors that drive a difference between conditions (Todd et al., 2013; Woolgar et al., 2014). Thus, if a data set consists of groups of trials that differ in some stimulus-specific features, MVPA can detect differences that might then be mistakenly attributed to the concept under investigation. In other words, the classifier can use stimulus-specific rather than category-specific features to decode data, effectively predicting stimuli instead of conceptual categories. Therefore, the present paper explores a method to determine the degree to which classification performance is specific to higher order category processing and to lower order stimulus processing.

Consider the following neuro-cognitive experiment, in which the concepts of animate and inanimate objects are to be distinguished based on electrical brain activity. 40 pictures each of six different types of animals (e.g. cow, bear, dog, frog, …) and tools (e.g. knife, scissors, hammer, saw, …) are presented to subjects, with the aim to decode the two conceptual categories from event-related EEG. Since different types of stimuli have features that distinguish them from the other types, the classifier will detect brain responses to individual stimuli based on combinations of their physical features alone (e.g. cows and frogs differ in size, shape and color). As we will show below, these differences between stimulus types will contribute to classification even in the absence of an actual effect of the superordinate concept. We will investigate the relative contribution of these two components, i.e. how much of the decoding performance originates from concept-related information and how much is caused by stimulus differences.

In the following, we will consider the concept-related information as the factor of interest (primary effect) and all the other contributing, concept-irrelevant factors as the nuisance effects. By relabeling the data, we can manipulate the relative contribution of concept (animate, inanimate) and stimulus (cow, frog, knife, scissors, …) to determine the presence of the effect of interest when nuisance effects are controlled for. The basic idea resembles that of a dose-response curve, in that we systematically vary the amount of concept-related information in the training data set of the classifier to assess how classification performance changes with varying levels of conceptual information. When the effect of concept-related information is completely counterbalanced, decoding performance originates solely from concept-irrelevant nuisance effects, which constitutes our null hypothesis for statistical testing. We will apply this method here in several examples, showing how to separate high-level cognitive concepts from low-level stimulus processing. In particular, we will show how this method can be used to describe the detailed time-course of cognitive concept processing. However, we believe that the basic method can find application in many similar problems.

2. Method & results

Suppose that an experiment has the aim to decode conceptual information (e.g. the semantic category) from brain activity. Different exemplars of each category are presented to the subjects and the brain response is recorded. For the sake of simplicity, and without loss of generality, we assume that there are two semantic categories $\mathfrak{A}$ and $\mathfrak{B}$. Each category consists of stimuli coming from $j = 1, 2, \ldots, k$ subclasses (see Fig. 1A). For instance, in our example of animals and tools, there are six subclasses per category (cow, bear, dog, frog, …for animals and knife, scissors, hammer, saw, …for tools). We assume that each stimulus is presented n times, resulting in $k \times n$ trials per category. We consider all of the n trials that belong to the j th subclass as one block of data and denote it with $\mathfrak{A}_j$ or $\mathfrak{B}_j$. Therefore, each category consists of k blocks and can be

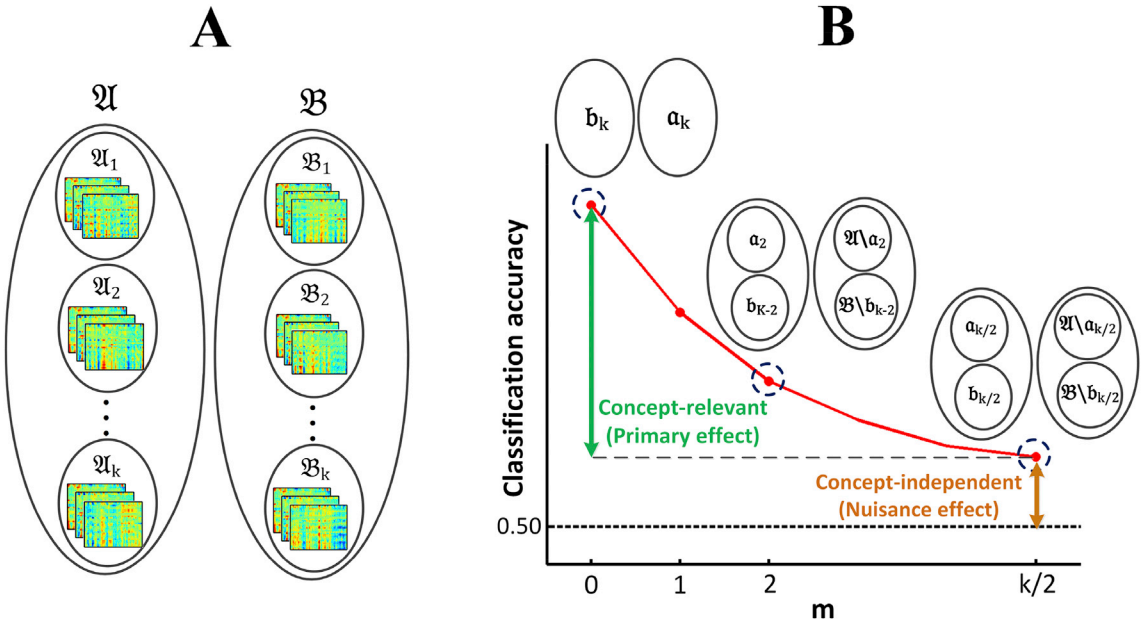

Fig. 1. Example of a concept-response curve. (A) Structure of data with two experimental conditions ($\mathfrak{A}$ and $\mathfrak{B}$, e.g. animate and inanimate objects) and k blocks of data per condition. Each block consists of all trials that belong to one subclass (e.g. frogs, cows, hammers, scissors, …). (B) By changing the number of blocks m in set S_1 belonging to category $\mathfrak{A}$ from 0 to $k/2$, we can change the amount of concept-relevant information distinguishing between sets S_1 and S_2. Each point of the curve is derived from the classification of S_1 versus S_2. $\mathfrak{a}_m$ and $\mathfrak{b}_m$ represent m-block subsets of $\mathfrak{A}$ and $\mathfrak{B}$, respectively. $\mathfrak{A}\backslash\mathfrak{a}_m$ denotes the set of blocks in $\mathfrak{A}$ but not in $\mathfrak{a}_m$ (similar for $\mathfrak{B}\backslash\mathfrak{b}_m$).