NeuroImage xxx (2017) 1–15

Contents lists available at ScienceDirect

NeuroImage

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

Deconstructing multivariate decoding for the study of brain function

Martin N. Hebart *, Chris I. Baker

Section on Learning and Plasticity, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA

ARTICLE INFO

Keywords:
Multivariate decoding
Multivariate analysis
Multivariate pattern analysis
Encoding
Decoding
fMRI
Prediction

ABSTRACT

Multivariate decoding methods were developed originally as tools to enable accurate predictions in real-world applications. The realization that these methods can also be employed to study brain function has led to their widespread adoption in the neurosciences. However, prior to the rise of multivariate decoding, the study of brain function was firmly embedded in a statistical philosophy grounded on univariate methods of data analysis. In this way, multivariate decoding for brain interpretation grew out of two established frameworks: multivariate decoding for predictions in real-world applications, and classical univariate analysis based on the study and interpretation of brain activation. We argue that this led to two confusions, one reflecting a mixture of multivariate decoding for prediction or interpretation, and the other a mixture of the conceptual and statistical philosophies underlying multivariate decoding and classical univariate analysis. Here we attempt to systematically disambiguate multivariate decoding for the study of brain function from the frameworks it grew out of. After elaborating these confusions and their consequences, we describe six, often unappreciated, differences between classical univariate analysis and multivariate decoding. We then focus on how the common interpretation of what is signal and noise changes in multivariate decoding. Finally, we use four examples to illustrate where these confusions may impact the interpretation of neuroimaging data. We conclude with a discussion of potential strategies to help resolve these confusions in interpreting multivariate decoding results, including the potential departure from multivariate decoding methods for the study of brain function.

1. Introduction

Multivariate decoding[1] has become a central method for the analysis of neuroscientific data. It is being employed commonly in fMRI (Haynes, 2015; Haynes and Rees, 2006; Norman et al., 2006; Tong and Pratte, 2012), but also neurophysiology in non-human primates (Quian Quiroga and Panzeri, 2009) and humans (Contini et al., 2017). The approach grew rapidly in popularity in the neuroimaging community when it became clear that it was not only useful for classification related to real-world applications such as brain-computer interfaces, but also for studying brain function. Now, in many domains classical univariate methods have been replaced by multivariate decoding, in part owing to the higher sensitivity afforded by these techniques (Haynes and Rees, 2006; Norman et al., 2006). In this way, multivariate decoding for brain interpretation grew out two established approaches: multivariate decoding for predictions in real-world applications, and classical univariate analysis for the study of brain function.

In this article, we argue that rather than being part of a consistent and independent statistical framework, multivariate decoding for brain interpretation often reflects a mixture of the philosophies it originated from (Fig. 1A), one activation-based and the other information-based. As a consequence, this mixture of philosophies creates a lot of potential for confusion in the interpretation of results derived from multivariate decoding methods. The aim of this article is

* Corresponding author. Laboratory of Brain and Cognition, National Institute of Mental Health, Building 10 Room 4C108, 10 Center Drive, Bethesda, MD 20814. USA.
E-mail address: martin.hebart@nih.gov (M.N. Hebart).

[1] For the reader unfamiliar with multivariate decoding in neuroimaging, we provide a brief working definition. In the neuroimaging literature, multivariate decoding refers to techniques that jointly analyze multiple measurement channels (e.g. fMRI voxels) to make predictions about variables of interest. For categorical predicted variables, this approach reflects multivariate classification, while for continuous variables it reflects multivariate regression. Multivariate decoding is typically performed using machine learning algorithms, for example support vector machines. One instance of measurements across channels is described as a "pattern" (e.g. a multi-voxel pattern).

http://dx.doi.org/10.1016/j.neuroimage.2017.08.005
Received 1 April 2017; Received in revised form 28 July 2017; Accepted 1 August 2017
Available online xxxx

to provide a systematic understanding of multivariate decoding for the study of brain function and the assumptions and limitations of this approach in the interpretation of multivariate decoding results.

First, we describe the two sources of confusion: i) the mixture of multivariate decoding for prediction and multivariate decoding for interpretation, and ii) the mixture of the statistical and conceptual philosophies underlying classical univariate analysis and multivariate decoding. Next, we illustrate six methodological and interpretational changes that – explicitly or implicitly – are adopted when shifting from classical univariate methods to multivariate decoding. This discussion is important, because it shows how multifaceted the differences between these approaches are and why they have been so difficult to characterize. Moving to a purely multivariate description of data, we then describe how the meaning of signal and noise is different in the statistical frameworks underlying classical univariate analysis and multivariate decoding. Finally, using four illustrative examples we demonstrate how the sources of confusion can affect the interpretation of multivariate decoding results.

Throughout the article, we use functional MRI as an example, where multivariate data are multiple voxels measured at different time points, and where predicted variables are experimental conditions.[2] However, this discussion applies equally to other modalities (e.g. structural MRI, MEG/EEG, connectivity measures) whenever multivariate decoding is used as a method of data analysis. In addition, we focus our discussion of multivariate decoding on multivariate classification, although our arguments may apply equally to multivariate regression in a decoding setting.

2. Two sources of confusion

2.1. Multivariate decoding for prediction vs. interpretation

The first major source of confusion stems from the distinction between multivariate decoding for prediction and multivariate decoding for interpreting brain function (Fig. 1A), which can be illustrated by the results of the 2006 Pittsburgh Brain Activity Interpretation Competition. The purpose of the competition was to use brain activity data measured with fMRI to predict the subjective perception of movie segments according to several criteria including the objects, spatial locations, sounds, and emotions associated with these segments. The winner was determined by who best predicted ratings based on independent fMRI data. According to the competition website and call for submissions, the goals of the competition were “to advance the methodology and assess the state of the science”, and “to advance the understanding of how the brain encodes, represents, and operates on dynamic experience”.[3] The competition received a lot of interest in the community, with multiple participants using multivariate decoding methods including sophisticated machine learning algorithms to carry out predictions (Nature Neuroscience Editorial, 2006). Surprisingly, the winners of the contest were a team of data scientists who acknowledged they did not know much about the brain prior to the competition (Sona et al., 2007). When visualizing the voxels their classifier used for predictions, many of them were contained within the ventricles and other regions typically related to physiological noise. Potentially, the most predictive voxels did not reflect brain activity in response to the ratings, but rather head motion and changes in physiological noise. Thus, one important lesson learned through the competition in 2006 is that the use of multivariate decoding can lead to excellent predictions, but sometimes to not very useful interpretations in terms of brain function. Perhaps for this reason, in 2007 the competition included a separate neuroscience prize for making substantial contributions to the understanding of brain function.

Today, the dichotomy of maximal prediction on the one hand and interpretation of brain function on the other continues to be of importance.[4] *Multivariate decoding for prediction* aims at identifying biomarkers that can be used to carry out predictions about underlying states of the brain. Here, maximal decoding performance is the goal, and success is determined by a model that can decode mental or physiological states from previously unseen data with high accuracy. The most frequently used tools in multivariate decoding are machine learning classifiers or variants thereof, which are often treated as a black box approach to assign labels to available data. Among others, studies employing multivariate decoding for prediction have investigated the prediction of disease status and progression (Ewers et al., 2011; Orrù et al., 2012), the usefulness of neuroimaging for brain computer interfaces in quadriplegic patients (Blankertz et al., 2007), and the feasibility of neuroimaging-based lie detection (Davatzikos et al., 2005; Farah et al., 2014; Peth et al., 2015). In addition, multivariate decoding for prediction has been used for read-out of information from visual cortex during perception (Kay et al., 2008; Miyawaki et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Thirion et al., 2006) and during sleep (Horikawa et al., 2013), and from auditory cortex during speech (Formisano et al., 2008). The source of the information is not necessarily of interest to these approaches, as long as the prediction is successful and can generalize to other relevant datasets.[5]

In contrast, *multivariate decoding for interpretation* aims at a better understanding of the human brain and does not require high predictive accuracy. The reasoning behind this approach is that as soon as a decoding model performs reliably better than chance, this demonstrates that there is structure in the data with respect to the conditions of interest, for example whether the participant was presented with a picture of a car or a chair. From this the researcher typically concludes that a given brain region carries discriminative information[6] about these categories, which may enlighten us about the neural computations carried out in this brain region. Among others, multivariate decoding for interpretation revealed the existence of subcortical effects of binocular rivalry (Haynes et al., 2005), feature binding in primary visual cortex (Seymour et al., 2009), working memory representations in primary visual cortex (Harrison and Tong, 2009), unconscious intentions in frontopolar cortex (Soon et al., 2008), visual search templates in object-selective cortex (Peelen et al., 2009), and reward value representations in parietal cortex (Kahnt et al., 2014). For this approach, variables such as head motion would act as confounds even when they consistently co-occur with the experimental variables.

While this distinction between prediction and interpretation was made explicit early on (Norman et al., 2006), multivariate decoding is

[2] In the following, we use the terms “experimental condition”, “experimental variable” or “independent variable” not in the narrow sense as variables under the experimenter's control (e.g. stimulus A vs. stimulus B), but in a broader sense including so called “quasi-experimental” settings, where the variable is under the environment's control and selected post-hoc by the experimenter (e.g. participant's choice A vs. choice B).

[3] Competition website: http://www.lrdc.pitt.edu/ebc/2006/comp_overview.htm, call for submissions: https://afni.nimh.nih.gov/afni/community/board/read.php?1,51415.

[4] The term *prediction* can have different meanings depending on the context. In inferential statistics, it refers to the existence of a model that can be used to tell how a variable will change in the future. For that reason, any model that describes a statistical dependence between two sets of variables can also be used as a predictive model. In the context of this article, prediction refers to models that are designed with a direct application in mind (such as stock market prediction), and where the reasons for this statistical dependence are only of secondary interest. While not irrelevant, space constraints preclude a discussion of the distinction between *predictive models* that allow predictions of dependent variables given the data without explicit assumptions about the data generation process, and generative models that additionally allow making predictions about the data given the model (Bzdok, 2016; Naselaris et al., 2011).

[5] Knowledge about the source of the information can help during the development of a new predictive model, when it is not yet clear if this source will help generalizing to all relevant cases. Using our example of the Pittsburgh brain interpretation competition, a non-neural source of information can and should be used for predictions if it is present in all relevant datasets.

[6] Our use of the term *information* follows the common use in human neurosciences employing multivariate decoding, i.e. the presence of a statistical dependence in the data that can be read out with the help of machine learning methods and that is believed to be of neuronal origin. This use of the term does not imply that the brain region can communicate this information to another brain region or that it is used in behavior (Williams et al., 2007; de-Wit et al., 2016).