NeuroImage 110 (2015) 48–59

Contents lists available at ScienceDirect

NeuroImage

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

Causal interpretation rules for encoding and decoding models in neuroimaging

Sebastian Weichwald [a,*], Timm Meyer [a], Ozan Özdenizci [b], Bernhard Schölkopf [a], Tonio Ball [c], Moritz Grosse-Wentrup [a]

[a] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[b] Sabancı University, Faculty of Engineering and Natural Sciences, Istanbul, Turkey
[c] Bernstein Center Freiburg, University of Freiburg, Freiburg, Germany

ARTICLE INFO

Article history:
Accepted 14 January 2015
Available online 24 January 2015

Keywords:
Encoding models
Decoding models
Interpretation
Causal inference
Pattern recognition

ABSTRACT

Causal terminology is often introduced in the interpretation of encoding and decoding models trained on neuroimaging data. In this article, we investigate which causal statements are warranted and which ones are not supported by empirical evidence. We argue that the distinction between encoding and decoding models is not sufficient for this purpose: relevant features in encoding and decoding models carry a different meaning in stimulus- and in response-based experimental paradigms.We show that only encoding models in the stimulus-based setting support unambiguous causal interpretations. By combining encoding and decoding models trained on the same data, however, we obtain insights into causal relations beyond those that are implied by each individual model type. We illustrate the empirical relevance of our theoretical findings on EEG data recorded during a visuo-motor learning task.

1. Introduction

The question how neural activity gives rise to cognition is arguably one of the most interesting problems in neuroimaging (Hamann, 2001; Ward, 2003; Atlas et al., 2010). Neuroimaging studies per se, however, only provide insights into neuralcorrelates but not into neural causes of cognition (Ward, 2003; Rees et al., 2002). Nevertheless, causal terminology is often introduced in the interpretation of neuroimaging data. For instance, Hamann writes in a review on the neural mechanisms of emotional memory that "Hippocampal activity in this study was correlated with amygdala activity, supporting the view that the amygdala *enhances* explicit memory by *modulating* activity in the hippocampus" (Hamann, 2001), and Myers et al. note in a study on working memory that "we tested […] whether pre-stimulus alpha oscillations measured with electroencephalography (EEG) *influence* the encoding of items into working memory" (Myers et al., 2014) (our emphasis of causal terminology). The apparent contradiction between the prevalent use of causal terminology and the correlational nature of neuroimaging studies gives rise to the following question: which causal statements are and whichones are not supported by empirical evidence?

We argue that the answer to this question depends on the experimental setting and on the type of model used in the analysis of neuroimaging data. Neuroimaging distinguishes between encoding and decoding models (Naselaris et al., 2011), known in machine learning as generative and discriminative models (Jordan, 2002). Encoding models predict brain states, e. g. BOLD activity measured by fMRI or event-related potentials measured by EEG/MEG, from experimental conditions (Friston et al., 1994, 2003; David et al., 2006). Decoding models use machine learning algorithms to quantify the probability of an experimental condition given a brain state feature vector (Mitchell et al., 2004; Pereira et al., 2009). Several recent publications have addressed the interpretation of encoding and decoding models in neuroimaging, discussing topics such as potential confounds (Todd et al., 2013; Woolgar et al., 2014), the dimensionality of the neural code (Davis et al., 2014), and the relation of linear encoding and decoding models (Haufe et al., 2014). We contribute to this discussion by investigating, for each type of model, which causal statements are warranted and which ones are not supported by empirical evidence. Our investigation is based on the seminal work by Pearl (2000) and Spirtes et al. (2000) on causal inference (cf. (Ramsey et al., 2010; Grosse-Wentrup et al., 2011; Waldorp et al., 2011; Mumford and Ramsey, 2014) for applications of this framework in neuroimaging). We find that the distinction between encoding and decoding models is not sufficient for this investigation. It is further necessary to consider whether models work in causal or anti-causal direction, i. e. whether they model the effect of a cause or the cause of an effect (Schölkopf et al., 2012). To accommodate this distinction, we distinguish between stimulus- and response-

* Corresponding author.
E-mail addresses: sweichwald@tuebingen.mpg.de (S. Weichwald), tmeyer@tuebingen.mpg.de (T. Meyer), oozdenizci@sabanciuniv.edu (O. Özdenizci), bs@tuebingen.mpg.de (B. Schölkopf), tonio.ball@uniklinik-freiburg.de (T. Ball), moritzgw@tuebingen.mpg.de (M. Grosse-Wentrup).

http://dx.doi.org/10.1016/j.neuroimage.2015.01.036
1053-8119/

based paradigms. We then provide causal interpretation rules for each combination of experimental setting (stimulus- or response-based) and model type (encoding or decoding). We find that when considering one model at a time, only encoding models in stimulus-based experimental paradigms support unambiguous causal statements. Also, we demonstrate that by comparing encoding and decoding models trained on the same data, experimentally testable conditions can be identified that provide further insights into causal structure. These results enable us to reinterpret previous work on the relation of encoding and decoding models in a causal framework (Todd et al., 2013; Woolgar et al., 2014; Haufe et al., 2014).

The empirical relevance of our theoretical results is illustrated on EEG data recorded during a visuo-motor learning task. We demonstrate that an encoding model allows us to determine EEG features that are effects of the instruction to rest or to plan a reaching movement, but does not enable us to distinguish between direct and indirect effects. By comparing relevant features in an encoding and a decoding model, we provide empirical evidence that sensorimotor μ- and/or occipital α-rhythms(8–14 Hz) are direct effects, while brain rhythms in higher cortical areas, including precuneus and anterior cingulate cortex, respond to the instruction to plan a reaching movement only as a result of the modulation by other cortical processes.

We note that while we have chosen to illustrate the empirical significance of our results on neuroimaging data, and specifically on EEG recordings, the provided causal interpretation rules apply to any encoding and decoding model trained on experimental data. This provides a guideline to researchers on how (not) to interpret encoding and decoding models when investigating the neural basis of cognition. A preliminary version of this work has been presented in Weichwald et al. (2014).

2. Methods

We begin this section by introducing the causal framework by Pearl (2000) and Spirtes et al. (2000) that our work is based on (Section 2.1) and demonstrate how it leads to testable predictions for the causal statements cited in the introduction (Section 2.2). We then introduce the distinction between causal and anti-causal encoding and decoding models (Section 2.3) and establish a connection between these models and causal inference (Section 2.4). This connection enables us to present the causal interpretation rules for encoding and decoding models in Section 2.5. In Section 2.6, we show that combining an encoding and a decoding model trained on the same data can provide further insights into causal structure. We conclude this section by providing a reinterpretation of previous work on encoding and decoding models in a causal framework (Section 2.7).

2.1. Causal Bayesian Networks

By X we denote the finite set of d random variables representing the brain state features, i. e. $X = \{X_1,\ldots, X_d\}$ While these variables may correspond to any type of independent and identically distributed (iid) samples of d brain state features, it is helpful to consider bandpower features of different EEG channels, trial-averaged BOLD activity at various cortical locations, or mean spike rates of multiple neurons as possible examples. By C we denote the random variable representing the (usually discrete) experimental condition. C stands for a stimulus ($C \equiv S$) or response ($C \equiv R$) variable and it will be made clear when C is restricted to either particular case. For convenience, we denote the set of all random variables by $\hat{X} = \{C, X_1, \ldots, X_d\}$. Throughout this article, we denote marginal, conditional and joint distributions by $P(X)$, $P(X|C)$ and $P(X, C)$, respectively (overloading the notation of P). For our theoretical investigations, we assume that the involved distributions have probability mass or density functions (PMFs or PDFs) with values denoted by $P(x)$, $P(x|c)$ and $P(x, c)$ respectively, again overloading the notation of P while it is always clear from the argument which function is meant. We use the common notations for independence and conditional independence:

$$X \perp\!\!\!\perp C \quad :\Longleftrightarrow \quad P(X|C) = P(X),$$
$$X \perp\!\!\!\perp C|Y \quad :\Longleftrightarrow \quad P(X|C,Y) = P(X|Y).$$

In the framework of Causal Bayesian Networks (CBNs) (Pearl, 2000; Spirtes et al., 2000), a variable X_i is said to be a cause of another variable X_j if the distributions ($PX_j|\mathrm{do}(X_i = x_i)$) are sensitive to x_i (cf. Pearl, 2000,p. 24f.). In this notation, the intervention $\mathrm{do}(X_i = x_i)$ signifies that X_i is externally set to a constant x_i, possibly resulting in a change of the distribution of X_j. The framework of CBNs thus defines cause-effect relations in terms of the impact of external manipulations. This is in contrast to frameworks that define causality in terms of information transfer (Granger, 1969; Roebroeck et al., 2005; Lizier and Prokopenko, 2010).

Causal relations between variables in CBNs are represented by directed acyclic graphs (DAGs). If we find a directed edge $X_i \to X_j$, we call X_i a direct cause of X_j and X_j a direct effect of X_i. In case there is no directed edge but at least one directed path $X_i \dashrightarrow X_j$, we call X_i an indirect cause of X_j and X_j an indirect effect of X_i. Note that the terms *(in-)direct cause/effect* depend on the set $\hat{X}$ of observed variables: consider $\hat{X} = \{C, X_1, X_2\}$ and the causal DAG $C \to X_1 \to X_2$. Then $C \dashrightarrow X_2$ and $C \to X_2$ wrt. $\hat{X}$, while $C \to X_2$ wrt. $\hat{X}' = \{C, X_2\}$. That is, whether a cause or effect is direct or indirect depends on the set of observed brain state features. We omit the considered set of nodes if it is clear from the context.

To establish a link between conditional independences and DAGs the following concepts are required:

- *d-separation*: Disjoint sets of nodes A and B are d-separated by another disjoint set of nodes C if and only if all $a \in A$ and $b \in B$ are d-separated by C. Two nodes $a \neq b$ are d-separated by C if and only if every path between a and b is blocked by C. A path between nodes a and b is blocked by C if and only if there is an intermediate node z on the path such that (i) $z \in C$ and z is a tail-to-tail ($\leftarrow z \rightarrow$) or head-to-tail ($\rightarrow z \rightarrow$) or (ii) z is head-to-head ($\rightarrow z \leftarrow$) and neither z nor any of its descendants is in C.
- *Causal Markov Condition (CMC)*: The CMC expresses the notion that each node in a causal DAG becomes independent of its non-descendants given its direct causes, i. e. that the causal structure implies certain (conditional) independences.
- *Faithfulness*: The faithfulness assumption states that all (conditional) independences between the random variables of a DAG are implied by its causal structure, i. e. there are no more (conditional) independences than those implied by the CMC.

Assuming faithfulness and the causal Markov condition, d-separation is equivalent to conditional independence, i. e. C d-separates A and B if and only if A and B are independent given C (Spirtes et al., 2000). The following three examples are the most relevant instances of d-separation for our following arguments. Firstly, consider the chain $X_0 \to X_1 \to X_2$. Here, X_1 d-separates X_0 and X_2 by blocking the directed path from X_0 to X_2. This implies that $X_0 \perp\!\!\!\perp X_2|X_1$. Secondly, consider the fork $X_0 \leftarrow X_1 \to X_2$. Here, X_1 d-separates X_0 and X_2, as X_1 is a joint cause of X_0 and X_2. This again implies that $X_0 \perp\!\!\!\perp X_2|X_1$. Thirdly, consider the collider $X_0 \to X_1 \leftarrow X_2$. In this case, X_1 does not d-separate X_0 and X_2. As X_1 is a joint effect of X_0 and X_2, it unblocks the previously blocked path between X_0 and X_2, implying that $X_0 \not\perp\!\!\!\perp X_2|X_1$.

The equivalence between d-separation and conditional independence enables us to infer causal relations between variables in $\hat{X}$ from observational data. By identifying conditional independences that hold in our data, and mapping them onto the equivalent d-separations, we gain knowledge about the causal structures that can give rise to our