Causal and anti-causal learning in pattern recognition for neuroimaging

Sebastian Weichwald[1], Bernhard Schölkopf[1], Tonio Ball[2], Moritz Grosse-Wentrup[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
{sweichwald, bs, moritzgw}@tuebingen.mpg.de
[2] Bernstein Center Freiburg, University of Freiburg, Freiburg, Germany
tonio.ball@uniklinik-freiburg.de

***Abstract*—Pattern recognition in neuroimaging distinguishes between two types of models: encoding- and decoding models. This distinction is based on the insight that brain state features, that are found to be relevant in an experimental paradigm, carry a different meaning in encoding- than in decoding models. In this paper, we argue that this distinction is not sufficient: Relevant features in encoding- and decoding models carry a different meaning depending on whether they represent causal- or anti-causal relations. We provide a theoretical justification for this argument and conclude that causal inference is essential for interpretation in neuroimaging.**

I. Introduction

Pattern recognition in neuroimaging aims to provide insights into the neural basis of cognitive processes. Two types of models are used in this endeavor: encoding- and decoding models. Encoding models predict a subject's brain state for a given experimental condition, while decoding models aim to reconstruct experimental conditions from neuroimaging data. This difference has important consequences for the interpretation of brain state features that are found to be relevant in each type of model.

It has been argued that only encoding models can provide a complete functional description of a region of interest [1]. Decoding models, on the other hand, may determine brain state features as relevant that are statistically independent of the experimental condition [2]. While in linear decoding models potential misinterpretations can be avoided by converting them into encoding models [3], this is a substantially more difficult problem for non-linear decoding models. As decoding models are becoming ever more popular in the analysis of neuroimaging data [4], the correct interpretation of such models is of considerable importance.

In this paper, we argue that the distinction between encoding- and decoding models is not sufficient to determine the meaning of relevant features in each type of model: Pattern recognition models need to be further distinguished with respect to whether they learn causal- or anti-causal relations [5]. In general, neuroimaging studies are based on the following causal structure: stimulus $\rightarrow$ brain activity $\rightarrow$ response. We note that more complex experimental paradigms, in which responses again act as stimuli [6], can also be modeled in this way by considering time-resolved variables, e. g. stimulus$[t_0]$ $\rightarrow$ brain activity$[t_1]$ $\rightarrow$ response$[t_2]$. Depending on whether experimental conditions are chosen to represent stimuli or responses, encoding- and decoding models then model causal- or anti-causal relations. In the following, we argue that this has important consequences for the interpretation of relevant features in each type of model. Furthermore, we argue that interpretation of neuroimaging data de facto requires causal inference problems to be solved.

The remainder of this article is organized as follows. In section II we introduce the necessary notation and terminology to formulate our proposed distinction of pattern recognition models in section II-D. Next, we theoretically investigate the interpretability of relevant features in each type of pattern recognition model (sections III-A to III-D) and briefly summarize our findings in section III-E. In section IV we argue that interpreting encoding- and decoding models is only a first step towards solving causal inference problems in the interpretation of neuroimaging data. We close with a conclusion in section V.

978-1-4799-4149-0/14/$31.00

II. Pattern recognition models

A. Notation

By X we denote the brain states represented by d features obtained from neuroimaging data, i. e. $X = \{X_1, ..., X_d\}$; by Y we denote the (usually discrete) experimental conditions. Throughout this paper we use the notations $p(X)$, $p(X|Y)$ and $p(X, Y)$ for (conditional or joint) probability density functions (PDFs). All PDFs are assumed to be known.

Independence is denoted by $X \perp\!\!\!\perp Y$ and conditional independence by $X \perp\!\!\!\perp Y|Z$. Dependence and conditional dependence is denoted by $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y|Z$, respectively. Causal relations in a directed acyclic graph are denoted by $X \rightarrow Y$ [7].

B. Encoding and decoding models

An encoding model $p(X|Y)$ represents how various experimental conditions are encoded in different brain states. We ask "How does the brain state look like given a certain experimental condition?". Examples for encoding models are

the general linear model [8] or the class-conditional mean: $E\{X|Y\}$.

A decoding model $p(Y|X)$ represents how different experimental conditions can be inferred from different brain states [9]. We ask "Which experimental condition is most likely given a certain brain state?". Decoding models are for example obtained using support vector machines or linear regression.

Note that this distinction solely reflects the direction of modeling according to the brain state but neglects any causal relation between brain state and experimental condition that might be known a priori.

C. Causal and anti-causal learning

The brain is constantly exposed to the world's stimuli and processes them, e. g. giving raise to perceptions. As such, stimuli S can only be causes but not effects of brain states X. The brain also constantly generates responses, e. g. movements, that are caused by the brain states. This gives rise to the following causal structure in neuroimaging studies: stimulus $\rightarrow$ brain state $\rightarrow$ response. Note that we are not necessarily able to observe all stimuli that cause a certain brain state or all features of the brain state which are causal for R. The causal structure enables us to distinguish between the following two scenarios:

1) Stimulus-based experiments: In a stimulus-based experiment the experimental conditions Y correspond to stimuli S presented to the subject. In general, we can control the stimulus presentation procedure and are thus able to randomize the presentation of stimuli. An example of a stimulus-based experiment is the randomized presentation of auditory stimuli to either the left or right ear. The causal structure of this setup is given by $S \rightarrow X$, i. e. stimuli cause brain activity.

In this case the encoding model $p(X|Y) = p(X|S)$ represents a causal relation, while the decoding model $p(Y|X) = p(S|X)$ models an anti-causal relation.

2) Response-based experiments: In a response-based experiment the experimental conditions Y represent subjects' responses that we observe. An example of a response-based experiment is the recording of volitional movements of either the left or right hand. The causal structure of this setup is given by $X \rightarrow R$, i. e. brain activity causes responses. Note that in this setting we are not able to control for and randomize the experimental conditions.

In contrast to a stimulus-based experiment, the encoding model $p(X|Y) = p(X|R)$ of a response-based experiment represents an anti-causal relation, while the decoding model $p(Y|X) = p(R|X)$ models a causal relation.

D. Distinction of pattern recognition models

Considering both the distinction of encoding- and decoding models and the distinction of stimulus- and response-based experiments we obtain the following four types of models:

A. Causal encoding models – $p(X|S)$
B. Anti-causal decoding models – $p(S|X)$
C. Anti-causal encoding models – $p(X|R)$
D. Causal decoding models – $p(R|X)$

In the following section we provide theoretical justifications why this distinction needs to be considered before interpreting encoding- or decoding models. As we show, interpretability of relevant features depends on whether the model represents causal or anti-causal relations.

III. Interpretation of relevant features

When interpreting an encoding model $p(X|Y)$, we want to link features relevant for encoding to the experimental condition. Relevant here means that we determine the set of brain state features that the encoding model deems dependent on the experimental condition, i. e. the features X_i for which $p(X_i|Y) \neq p(X_i)$ and hence $X_i \not\perp\!\!\!\perp Y$. The remaining features are independent of Y. One way to do this in practice is to test the class-conditional sample means of each feature for statistically significant differences. Features that, according to this univariate test, significantly vary with Y are considered relevant for the encoding model.

When interpreting a decoding model $p(Y|X)$, we want to determine which features are relevant for decoding the experimental condition. Relevant here means that we determine if a brain state feature or a set of features X_i helps in decoding the experimental condition, i. e. it is tested whether $p(Y|X) \neq p(Y|X \setminus X_i)$ and hence $X_i \not\perp\!\!\!\perp Y|X \setminus X_i$. One way to do this in practice is recursive feature elimination, i. e. permuting or removing X_i from the feature set and testing whether this significantly decreases decoding accuracy [10]. It is common to remove all features that are irrelevant for decoding to reduce dimensionality and obtain the minimal set of features that yields an optimal decoding model. Features of that set are considered relevant for the decoding model. We note that there might be other ways of identifying relevant features of a decoding model which might lead to different conclusions.

For our theoretical arguments we assume that we can identify all relevant features for each type of model. We now show that relevant features in encoding- and decoding models carry a different meaning depending on the causal structure.

A. Causal encoding models

From the encoding model $p(X|Y) = p(X|S)$ of a stimulus-based experiment we obtain the set X_S^{enc} of features that are dependent on S, i. e. for every $X_i \in X_S^{\text{enc}}$ we have $S \not\perp\!\!\!\perp X_i$. We denote the complementary set as $X_0^{\text{enc}} := X \setminus X_S^{\text{enc}}$.

According to Reichenbach's principle [11], the dependency between S and X_S^{enc} implies that $S \rightarrow X_S^{\text{enc}}$, $S \leftarrow X_S^{\text{enc}}$, or $S \leftarrow H \rightarrow X_S^{\text{enc}}$ with H a joint common cause of S and X_S^{enc}. In the stimulus-based setting we can control for and randomize the stimulus. This enables us to rule out the last two cases and conclude $S \rightarrow X_S^{\text{enc}}$, i. e. the features in X_S^{enc} are genuine effects of S [12].

In addition, we have $S \perp\!\!\!\perp X_0^{\text{enc}}$, which allows us to conclude that features in X_0^{enc} are not genuine effects of S.

As such, all relevant features in a causal encoding model are genuine effects of S, while irrelevant features are not effects of S.

B. Anti-causal decoding models

From the decoding model $p(Y|X) = p(S|X)$ of a stimulus-based experiment we obtain the minimal set X_S^{dec} of features that allows to decode the stimulus, i. e. $p(S|X) = p(S|X_S^{\text{dec}})$. It hence holds that $S \perp\!\!\!\perp X_0^{\text{dec}} | X_S^{\text{dec}}$ where $X_0^{\text{dec}} := X \setminus X_S^{\text{dec}}$ is the set of features that do not further improve decoding.

We now describe two counterexamples that show that one can neither conclude that features in X_0^{dec} are not genuine effects of S nor that features in X_S^{dec} are indeed genuine effects of S. First, assume $S \to X_1 \to X_2$. Since $p(S|X_1, X_2) = p(S|X_1)$, i. e. $S \perp\!\!\!\perp X_2 | X_1$, we have $X_2 \in X_0^{\text{dec}}$ although X_2 is actually a genuine effect of S. Second, assume $S \to X_1 \leftarrow X_2$. Since $p(S|X_2, X_1) \neq p(S|X_1)$, i. e. $S \not\perp\!\!\!\perp X_2 | X_1$, we obtain $X_2 \in X_S^{\text{dec}}$ although X_2 is not a genuine effect of S.

This establishes that interpreting anti-causal decoding models in this way has two drawbacks. First, features in X_S^{dec} can only be considered as potential effects of S. Second, genuine effects of S might be missed.

C. Anti-causal encoding models

Form the encoding model $p(X|Y) = p(X|R)$ of a response-based experiment we obtain the set of features that are dependent on R, i. e. for every $X_i \in X_R^{\text{enc}}$ we have $X_i \not\perp\!\!\!\perp R$. We denote the complementary set as $X_0^{\text{enc}} := X \setminus X_R^{\text{enc}}$ (overloading notation).

According to Reichenbach's principle, the dependency between $X_i \in X_R^{\text{enc}}$ and R implies that $X_i \to R$, $X_i \leftarrow R$, or $X_i \leftarrow H \to R$ with H a joint common cause of X_i and R. A priori we know that brain activity $\to$ response. This enables us to rule out the case $X_i \leftarrow R$. As we show next, we can not uniquely determine which of the last two scenarios is the case, i. e. features in X_R^{enc} are potential but not necessarily genuine causes of R.

Consider $X_2 \leftarrow X_1 \to R$: we have $X_1 \not\perp\!\!\!\perp R$ and $X_2 \not\perp\!\!\!\perp R$ and therefore $X_1, X_2 \in X_R^{\text{enc}}$. But note that $X_1 \to R$ while $X_2 \not\to R$, i. e. X_2 is not a cause of R. This shows that features in X_R^{enc} are not necessarily genuine causes of R.

Features in X_0^{enc}, on the other hand, are independent of R and can hence be considered to be no causes of R.

As such, not all relevant features in anti-causal encoding models are genuine causes of R, while irrelevant features are indeed not causal for R.

D. Causal decoding models

From the decoding model $p(Y|X) = p(R|X)$ of a response-based experiment we obtain the minimal set X_R^{dec} of features that allows to decode the response, i. e. $p(R|X) = p(R|X_R^{\text{dec}})$. It hence holds that $S \perp\!\!\!\perp X_0^{\text{dec}} | X_R^{\text{dec}}$ where $X_0^{\text{dec}} := X \setminus X_R^{\text{dec}}$ is the set of features that do not further improve decoding.

We now describe two counterexamples that show that one can neither conclude that features in X_0^{dec} are not genuine causes of R nor that features in X_R^{dec} are genuine causes of R. First, assume $X_2 \to X_1 \to R$. Since $p(R|X_1, X_2) = p(R|X_1)$, i. e. $X_2 \perp\!\!\!\perp R | X_1$, we have $X_2 \in X_0^{\text{dec}}$ although X_2 is a cause of R. Second, assume the graph shown in figure 1 where H is a hidden common cause of X_1, X_2 and R which is not observable as a brain state feature. Since $p(R|X_2, X_1) \neq p(R|X_1)$ and $p(R|X_2, X_1) \neq p(R|X_2)$ we have $X_1, X_2 \in X_R^{\text{dec}}$ although both X_1 and X_2 are not causes of R.

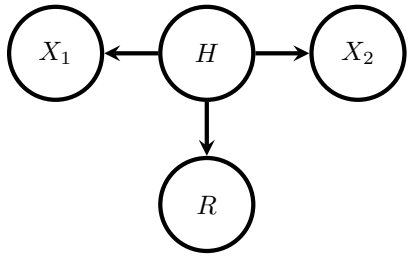

Fig. 1. Causal graph of an exemplary response-based experiment: H is not observable as a brain state feature and hence denotes a hidden common cause of the observed brain state features X_1 and X_2 and the response R.

This establishes that interpreting causal decoding models this way has two drawbacks. First, features in X_R^{dec} are not necessarily causes of R. Second, genuine of R causes might be missed.

E. Subsumption

In the previous sections we showed that the interpretation of relevant features in encoding- and decoding models depends on the underlying causal structure. This justifies our argument that the distinction of encoding- and decoding models is not sufficient. In particular we argued that, without employing further assumptions,

A. causal encoding models $p(X|S)$ allow to identify genuine effects X_S of S.
B. anti-causal decoding models $p(S|X)$ allow to identify some potential effects of S.
C. anti-causal encoding models $p(X|R)$ allow to identify potential causes of R.
D. causal decoding models $p(R|X)$ allow to identify some potential causes of R.

IV. Causal inference in neuroimaging

So far, we have argued that the causal structure of a neuroimaging study, i. e. whether we learn in causal- or anti-causal direction, has to be taken into account when interpreting relevant features in encoding- and decoding models. In particular, we have shown that, with the exception of the causal encoding model, the meaning of relevant features in encoding- and decoding models is ambiguous. In the following, we demonstrate on two examples that such ambiguities can be resolved by means of causal inference [7, 13]. Throughout this section we assume *faithfulness*, i. e. we assume that all observed (conditional) independence relations are implied by the causal structure [13]. In the following examples, we additionally assume *causal sufficiency*, i. e. we assume that there are no hidden confounders.

A. Causal inference in stimulus-based experiments

Consider two brain state features X_1 and X_2 in a stimulus-based experiment with $S \not\perp\!\!\!\perp X_1$, $S \perp\!\!\!\perp X_2$, and $S \not\perp\!\!\!\perp X_2|X_1$.

If we learn an encoding model on this data, we find that $X_1 \in X_S^{\text{enc}}$ and $X_2 \in X_0^{\text{enc}}$. We can thus conclude that X_1 is an effect of S, i. e. $S \rightarrow X_1$ (cf. section III-A). We can not, however, determine the causal relation between X_1 and X_2.

If we learn a decoding model, on the other hand, we find that $X_1, X_2 \in X_S^{\text{dec}}$, i. e. we find both features to be relevant for decoding S, as $S \not\perp\!\!\!\perp X_2|X_1$ and $S \rightarrow X_1$.

Under the assumptions of faithfulness and causal sufficiency, the only causal structure that can give rise to these observations is $S \rightarrow X_1 \leftarrow X_2$, i. e. X_2 is a cause of X_1 [7].

By learning both an encoding- and a decoding model on the same data, and comparing relevant features, we have thus determined the causal relations between the observed variables. An example of this inference procedure, known as the inference rule for potential causation [7], is given in [14].

B. Causal inference in response-based experiments

Consider two brain state features X_1 and X_2 in a response-based experiment with $X_1 \not\perp\!\!\!\perp R$, $X_2 \not\perp\!\!\!\perp R$ and $X_2 \perp\!\!\!\perp R|X_1$.

If we learn an encoding model on this data, we find that $X_1, X_2 \in X_R^{\text{enc}}$ as $X_1, X_2 \not\perp\!\!\!\perp R$. We thus conclude that both X_1 and X_2 are potential but not necessarily genuine causes of R (cf. section III-C).

If we learn a decoding model, on the other hand, we find that only $X_1 \in X_R^{\text{dec}}$, as X_2 does not help for decoding if X_1 is already known due to $X_2 \perp\!\!\!\perp R|X_1$. By only looking at the decoding model, we would only identify X_1 as a potential cause of R.

Taken together, however, the only causal structures that can give rise to these observations, again assuming faithfulness and causal sufficiency, are $X_2 \leftarrow X_1 \rightarrow R$ or $X_2 \rightarrow X_1 \rightarrow R$ [7]. As in both structures $X_1 \rightarrow R$, we can conclude that X_1 is a direct cause of R. The role of X_2, however, remains ambiguous.

By learning both an encoding- and a decoding model on the same data, and comparing relevant features, we have thus again identified a causal relation between observed variables.

V. Conclusion

In the previous section, we have demonstrated on two examples how the combination of encoding- and decoding models can resolve ambiguities that can not be decided when only looking at one type of model. This is due to the fact that relevant features are determined by univariate independence tests in encoding models and by multivariate conditional independence tests in decoding models. Both types of tests provide complementary information on the underlying causal structure.

As we have shown in section IV-B, however, these tests do not always uniquely determine the causal structure of a given set of observed variables. In general, conditional independence tests on all subsets of observed variables may provide further information [7, 13]. An exhaustive description of the causal inference rules based on conditional independence tests is beyond the scope of the present paper.

We conclude by emphasizing that the causal structure, as determined by a priori knowledge and/or causal inference methods, has to be taken into account when interpreting neuroimaging data.

References

[1] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fMRI," *NeuroImage*, vol. 56, no. 2, pp. 400–410, 2011.

[2] M. T. Todd, L. E. Nystrom, and J. D. Cohen, "Confounds in multivariate pattern analysis: Theory and rule representation case study." *NeuroImage*, vol. 77, pp. 157–165, 2013.

[3] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, 2014.

[4] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: a tutorial overview," *NeuroImage*, vol. 45, no. 1, pp. S199–S209, 2009.

[5] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," in *29th International Conference on Machine Learning (ICML 2012)*, 2012, pp. 1255–1262.

[6] M. Gomez-Rodriguez, J. Peters, J. Hill, B. Schölkopf, A. Gharabaghi, and M. Grosse-Wentrup, "Closing the sensorimotor loop: haptic feedback facilitates decoding of motor imagery," *Journal of Neural Engineering*, vol. 8, no. 3, p. 036005, 2011.

[7] J. Pearl, *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

[8] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[9] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Machine Learning*, vol. 57, no. 1-2, pp. 145–175, 2004.

[10] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns," *NeuroImage*, vol. 43, no. 1, pp. 44–58, 2008.

[11] H. Reichenbach, *The Direction of Time*. University of California Press, Berkeley, 1956.

[12] P. W. Holland, "Statistics and causal inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[13] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[14] M. Grosse-Wentrup, B. Schölkopf, and J. Hill, "Causal influence of gamma oscillations on the sensorimotor rhythm," *NeuroImage*, vol. 56, no. 2, pp. 837–842, 2011.