Causal interpretation rules for encoding and decoding models in neuroimaging

Sebastian Weichwald[1], **Timm Meyer**[1], **Ozan Özdenizci**[2], **Bernhard Schölkopf**[1], **Tonio Ball**[3], **Moritz Grosse-Wentrup**[1]
[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
{sweichwald, moritzgw}@tuebingen.mpg.de
[2] Sabancı University, Faculty of Engineering and Natural Sciences, Istanbul, Turkey
oozdenizci@sabanciuniv.edu
[3] Bernstein Center Freiburg, University of Freiburg, Freiburg, Germany
tonio.ball@uniklinik-freiburg.de

Abstract

How neural activity gives rise to cognition is arguably one of the most interesting questions in neuroimaging [1]. While causal terminology is often introduced in the interpretation of neuroimaging data, causal inference frameworks are rarely explicitly employed (cf. [3] for an example).

In our recent work we cast widely used analysis methods in a causal framework in order to foster its acceptance in the neuroimaging community [7]. In particular we focus on typical analyses in which variables' *relevance* in encoding and decoding models [4] (also known as generative or discriminative models [5]) with a dependent stimulus/response variable is interpreted. By linking the concept of *relevant* variables to marginal/conditional independence properties we demonstrate that (a) identifying *relevant* variables is indeed a first step towards causal inference; (b) combining encoding and decoding models can yield further insights into the causal structure, which cannot be gleaned from either model alone. We demonstrate the empirical relevance of our findings on EEG data recorded during a visuomotor learning task.

The rigorous theoretical framework of causal inference allows to expound the assumptional underpinnings and limitations of common (intuitive) analyses in this field. Furthermore, it sheds light on problems covered in recent neuroimaging literature such as confounds in multivariate pattern analysis [6] or interpretation of linear encoding and decoding models [2].

References

[1] C. F. Craver. *Explaining the brain*. Oxford University Press, 2007.

[2] 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*, 87:96–110, 2014.

[3] J. A. Mumford and J. D. Ramsey. Bayesian networks for fMRI: A primer. *NeuroImage*, 86:573–582, 2014.

[4] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56:400–410, 2011.

[5] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press, 2002.

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

[7] S. Weichwald, T. Meyer, O. Özdenizci, B. Schölkopf, T. Ball, and M. Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015.