#OHBM2020
@JoramSoch

Decoding behavioral responses from fMRI without learning behavioral responses from fMRI

Abstract #3720
Poster #1702

Joram Soch[1,2,9,•], John-Dylan Haynes[1–8]

[1] Bernstein Center for Computational Neuroscience, [2] Berlin Center for Advanced Neuroimaging, [3] Berlin School of Mind and Brain, Berlin, Germany
[4] Clinic for Neurology, Charité – Universitätsmedizin Berlin, Germany, [5] Department of Psychology, Humboldt-Universität zu Berlin, Germany
[6] EXC NeuroCure, Charité Berlin, [7] EXC Science of Intelligence, TU Berlin, [8] CRC Volition and Cognitive Control, TU Dresden, Germany
[9] German Center for Neurodegenerative Diseases, Göttingen, Germany, • BCCN Berlin, Philippstraße 13, Haus 6, 10115 Berlin, Germany;
e-mail address: **joram.soch@bccn-berlin.de**.

Introduction

The data acquired during a functional magnetic resonance imaging (fMRI) experiment can usually be categorized into experimental design **X** (e.g. experimental conditions, modulator variables), physiological data **Y** (i.e. the measured hemodynamic signals) and behavioral data **Z** (e.g. button presses, stimulus ratings).

In multivariate pattern analysis (MVPA) of fMRI data [1,2], button presses are typically decoded by training a classifier to distinguish the recorded responses Z based on the measured data Y (*conventional response decoding*, CRD).

Here we show that this can be achieved without constructing an explicit mapping from fMRI signals to behavioral data. In fact, button presses can be equally well decoded when first reconstructing the experimental design X from measured data Y and then predicting behavioral responses Z from the reconstructed design X (*neurobehavioral decoding*, NBD).

Theory

Decoding from the design [3]: First, we estimate a model of the behavioral data, given the experimental design: **Z = g(X)**. In the case of discrete experimental conditions (indexed by **i**) and response options (indexed by **j**), this mapping is simply given by the condition-to-response transition probabilities (see Figure 2B):

$$\hat{g}: \hat{p}_{ij} = \Pr(z_j = 1 | x_i = 1)$$

Conventional response decoding: Second, we establish a mapping from measured signals to behavioral data: **Z = h(Y)**. This is used to predict button presses from fMRI signals using cross-validation:

$$\hat{Z}_2 = \hat{h}(Y_2), \;\; \hat{h} \leftarrow Y_1, Z_1$$

Neurobehavioral decoding: Finally, we establish a mapping from measured signals to experimental design: **X = f⁻¹(Y)**. Together with the behavioral model **Z = g(X)**, this allows to predict button presses on a route through the experimental conditions (see Figure 1A):

$$\hat{Z}_2 = \hat{g}(\hat{f}^{-1}(Y_2)), \;\; \hat{f}^{-1} \leftarrow Y_1, X_1$$

All these models are estimated in a cross-validated fashion, using leave-one-out cross-validation over fMRI recording sessions (see Figure 1B). Behavioral models (**g**) were estimated with trial-wise linear regression

$$\hat{g}(X) = X\hat{P}, \;\; \hat{P} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z$$

and decoding analyses (**h, f⁻¹**) were performed with trial-wise logistic regression

$$\hat{h}(Y): \log\mathrm{odds}(z_j = 1) = Y\hat{w}_Z^{(j)}$$
$$\hat{f}^{-1}(Y): \log\mathrm{odds}(x_i = 1) = Y\hat{w}_X^{(i)}$$

Because behavioral responses (**Z**) and experimental conditions (**X**) are best decoded from different parts of the brain, the approaches had to be made comparable. For each approach, the most informative searchlight was selected based on within-sample decoding accuracy and results are reported as out-of-sample decoding accuracies (see Figure 1B):

$$\mathrm{DA}(Z, \hat{Z}) = \frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{q}[\hat{z}_{ij} = \max(\hat{z}_i)] \cdot z_{ij}$$

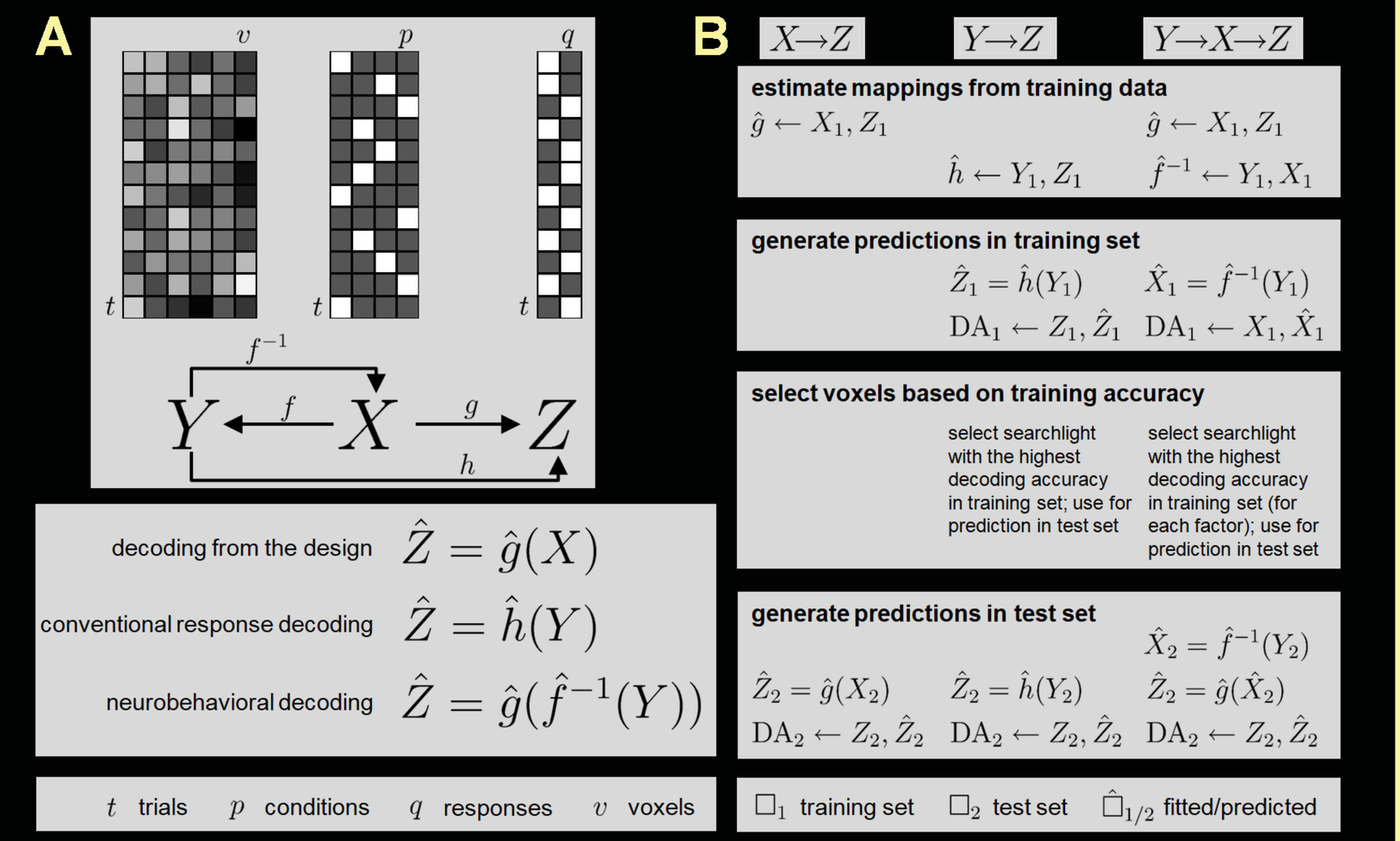
Experiment

We analyzed the example data set [6] of *The Decoding Toolbox* (TDT) [4,5]. This experiment (see Figure 2A) used five experimental dimensions (cue, stimulus color and direction, color and direction requiring left button press) and one behavioral dimension (left vs. right button press).

In CRD, button presses were directly decoded from the fMRI signal. In NBD, stimulus color and direction were decoded from the fMRI signal; these were then combined with cue and response mapping to yield a reconstructed design; this was then combined with the estimated condition-to-response mapping to yield decoded responses.

Overall, we obtain statistically indistinguishable performance of CRD and NBD in both subjects (see Figure 2B). This was irrespective of the searchlight radius (3 or 6 mm) used for decoding analyses.

Figure 1. *Neurobehavioral decoding: theory and analysis logic.*



(A) The analyses operate on the trial-by-voxel matrix Y, the trial-by-condition matrix X and the trial-by-response matrix Z. **(B)** We compare the performances of decoding from the design (X→Z), conventional response decoding (Y→Z) and neurobehavioral decoding (Y→X→Z).

Figure 2. *Neurobehavioral decoding: experiment and decoding results.*

(A) The experimental design varied cue, stimulus color, stimulus direction and stimulus-response mapping. **(B)** Two subjects performed the task with varying behavioral accuracy. CRD (green) and NBD (blue) did not differ significantly (error bars ≙ 90% binomial CIs, n = 256).

Discussion

In this proof-of-concept study, we have demonstrated that behavioral responses can be decoded without training on neurophysiological data measured during behavioral responses, but rather indirectly by taking a detour via the experimental design. This is particularly interesting, because CRD is commonly seen as a sanity check, the decoding accuracy of which should not be exceeded by other analyses. It is also worth noting that in our example, just one response dimension (left vs. right), but two design dimensions (color and direction) had to be decoded. We hypothesize that decoding the design from the data acts as a feature reduction mechanism which helps NBD predicting behavior using the psychologically most meaningful factors. In the future, we want to validate this finding in a larger cohort [7,8] and extend it to continuous behavioral measures such as stimulus ratings and reaction times [9,10].

References

(1) Haynes JD & Rees G (2006). Neuroimaging: decoding mental states from brain activity in humans. Nature Reviews Neuroscience, vol. 7, iss. 7, pp. 523-534.
(2) Haynes JD (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. Neuron, vol. 87, iss.,2, pp. 257-270.
(3) Görgen K et al. (2018). The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. NeuroImage, vol. 180, pp. 19-30.
(4) Görgen K et al. (2012). The Decoding Toolbox (TDT): A new fMRI analysis package for SPM and Matlab. OHBM 2012, Poster #378MT; URL: https://f1000research.com/posters/1092032.
(5) Hebart MN et al. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Frontiers in Neuroinformatics, vol. 8, art. 88.
(6) Hebart MN & Görgen K (2015). Example dataset for The Decoding Toolbox (TDT); URL: https://sites.google.com/site/tdtdecodingtoolbox/.
(7) Reverberi C, Görgen K, Haynes JD (2012). Compositionality of rule representations in human prefrontal cortex. Cerebral Cortex, vol. 22, iss.6, pp. 1237-1246.
(8) Reverberi C, Görgen K, Haynes JD (2012). Distributed representations of rule identity and rule order in human frontal cortex and striatum. JNeurosci, vol. 32, iss.48, pp. 17420-17430.
(9) Todd MT, Nystrom LE, Cohen JD (2013). Confounds in multivariate pattern analysis: theory and rule representation case study. NeuroImage, vol. 77, pp. 157-165.
(10) Woolgar A, Golland P, Bode S (2014). Coping with confounds in multivoxel pattern analysis: what should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. NeuroImage, vol. 98, pp. 506-512.