eNeuro

Commentary | Cognition and Behavior

Inverted Encoding Models Assay Population-Level Stimulus Representations, Not Single-Unit Neural Tuning

Thomas C. Sprague[1], Kirsten C. S. Adam[2], Joshua J. Foster[2], Masih Rahmati[1], David W. Sutterer[2] and Vy A. Vo[3]

[1]*Department of Psychology, New York University, New York, NY 10003*
[2]*Department of Psychology and Institute for Mind and Biology, University of Chicago, Chicago, IL 60637*
[3]*Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA 92093*

DOI: 10.1523/ENEURO.0098-18.2018

Received: 15 March 2018

Revised: 26 April 2018

Accepted: 3 May 2018

Published: 11 May 2018

Author contributions: TCS wrote the first draft, KCSA, JJF, MR, DWS and VAV edited the manuscript. All authors jointly developed ideas expressed in this commentary.

Funding: http://doi.org/10.13039/100000053HHS | NIH | National Eye Institute (NEI)
F32-EY028438
R01-EY016407

Funding: http://doi.org/10.13039/100000001National Science Foundation (NSF)
Graduate Research Fellowship

Funding: http://doi.org/10.13039/100000025HHS | NIH | National Institute of Mental Health (NIMH)
2R01-MH087214-06A1

Conflict of Interest: Authors declare no conflict of interest.

Supported by NEI F32-EY028438 (TCS); NSF Graduate Student Fellowship (VAV); NEI R01-EY016407 (MR); and NIMH 2R01-MH087214-06A1 (KCSA, JJF, and DWS).

K.C.S.A., J.J.F., M.R., D.W.S. and V.A.V. are contributed equally and listed alphabetically.

Correspondence: Thomas C. Sprague, 6 Washington Place, New York, NY, 10003; E-mail: tsprague@nyu.edu

Cite as: eNeuro 2018; 10.1523/ENEURO.0098-18.2018

Title: Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning

Abbreviated title: Assaying population-level stimulus representations

Thomas C. Sprague[1*], Kirsten C. S. Adam[2+], Joshua J. Foster[2+], Masih Rahmati[1+], David W. Sutterer[2+], and Vy A. Vo[3+]

[1]Department of Psychology, New York University, New York, NY, 10003
[2]Department of Psychology and Institute for Mind and Biology, University of Chicago, Chicago, IL, 60637
[3]Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA, 92093
+ authors contributed equally and are listed alphabetically

Author contributions: TCS wrote the first draft, KCSA, JJF, MR, DWS and VAV edited the manuscript. All authors jointly developed ideas expressed in this commentary.

Correspondence:
* 6 Washington Place, New York, NY, 10003; tsprague@nyu.edu (corresponding author)

Number of Figures: 1
Number of Tables: 0
Number of Multimedia: 0

Number of words, abstract: n/a
Number of words, significance statement: 120/120
Number of words, introduction: n/a
Number of words, discussion: n/a
Total number of words: 1947

Acknowledgments: We wish to thank Clayton Curtis, Edward Ester, and John Serences for comments on early drafts of the manuscript and useful discussions.

Conflict of interest: the authors declare no conflict of interest

Funding sources: Supported by NEI F32-EY028438 (TCS); NSF Graduate Student Fellowship (VAV); NEI R01-EY016407 (MR); and NIMH 2R01-MH087214-06A1 (KCSA, JJF, and DWS).

SIGNIFICANCE STATEMENT (120/120 words)**:** Inverted encoding models (IEMs) are a powerful tool for reconstructing population-level stimulus representations from aggregate measurements of neural activity (e.g., fMRI or EEG). In a recent report, Liu et al. (2018) tested whether IEMs can provide information about the underlying tuning of *single units*. Here, we argue that using stimulus reconstructions to infer properties of single neurons, such as neural tuning bandwidth, is an ill-posed problem with no unambiguous solution. Instead of interpreting results from these methods as evidence about single-unit tuning, we emphasize the utility of these methods for assaying *population-level* stimulus representations. These can be compared across task conditions to better constrain theories of large-scale neural information processing across experimental manipulations, such as changing sensory input or attention.

MAIN TEXT

Neuroscience methods range astronomically in scale. In some experiments, we record subthreshold membrane potentials in individual neurons, while in others we measure aggregate responses of thousands of neurons at the millimeter scale. A central goal in neuroscience is to bridge insights across all scales to understand the core computations underlying cognition (Churchland and Sejnowski, 1988). However, inferential problems arise when moving across scales: single-unit response properties cannot be inferred from fMRI activation in single voxels, subthreshold membrane potential cannot be inferred from extracellular spike rate, and the state of single ion channels cannot be inferred from intracellular recordings. These are all examples of an *inverse problem* in which an observation at a larger scale is consistent with an enormous number of possible observations at a smaller scale.

Recent analytical advances have circumvented challenges inherent in inverse problems by instead transforming aggregate signals from their native 'measurement' space (e.g., activation pattern across fMRI voxels) into a model-based 'information space' (e.g., activity level of modeled information channels). To make this inference possible, aggregate neural signals (fMRI voxel activation or EEG electrode activity) are modeled as a combination of feature-selective information channels, each with defined sensitivity profiles consistent with the single-unit literature (e.g., experimenter-defined tuning to a particular orientation, *Fig. 1A*; Brouwer and Heeger, 2009, 2011). When an aggregate neural signal is described with such an *encoding model*, it is possible to invert this model to infer the activity of each channel given a new pattern of neural activity (hence, these methods are often called 'inverted encoding models', IEM; Sprague et al., 2015). Importantly, rather than attempt to solve the inverse problem (how do single-units respond?), this method makes simplifying assumptions that enable transformation of one population-level measurement (aggregate neural signals in voxel- or

electrode space) into another (stimulus representations in 'channel space'). These reconstructed 'channel response functions' enable visualization, quantification, and comparison of population-level stimulus representations across manipulations of task conditions (Brouwer and Heeger, 2011, 2013; Foster et al., 2017; Garcia et al., 2013; Scolari et al., 2012; Sprague and Serences, 2013).

Recently, Liu et al. (2018) examined whether an IEM applied to fMRI data can be used to unambiguously infer the underlying response properties of single units. To this end, they manipulated the contrast of orientated gratings, because contrast only affects the amplitude of single-unit orientation tuning functions, but not their tuning width (e.g. Sclar and Freeman, 1982). The authors reasoned that, if the width of single-unit tuning functions do not change with stimulus contrast, and if population-level feature reconstructions derived from aggregate neural signals can be used to make meaningful inferences about single-unit tuning, then manipulating contrast should not change the width of population-level channel-response functions.

To test this prediction, the authors used an IEM to reconstruct representations of grating orientations for two different contrast levels. The authors modeled voxel responses as a sum of neural channels tuned to different orientations based on known visual response properties (*Fig. 1A*). After extracting activation patterns from visual cortex, the authors split data from each contrast condition into a training set, used to estimate how each modeled neural channel contributes to each voxel (*Fig. 1B*), and a testing set, which was used in conjunction with the best-fit model from the training set to compute channel response functions (*Fig. 1C*).

The authors found that reconstructed channel response functions in visual cortex were 'broader' for low-contrast gratings than for high-contrast gratings (Liu et al. Fig. 2), which they suggest could be interpreted as evidence that single-unit orientation tuning width depends on stimulus contrast. However, because this observation conflicts with demonstrations from single-unit physiology that orientation tuning is contrast-invariant, Liu et al. (2018) sought to resolve this discrepancy using simulations.

The authors simulated cortical fMRI data under different conditions to assess how changes in single-unit responses might be reflected in reconstructed channel response functions. Each simulated voxel's response was modeled as a noisy weighted sum of orientation-tuned neurons, each with a different orientation preference (Liu et al., Fig. 3). Across runs of their simulations, the authors manipulated simulated response properties, like orientation tuning width of constituent model neurons and signal-to-noise ratio (SNR) of the voxel response. The authors found that by decreasing the response amplitude of each simulated neuron (thus, decreasing SNR) without changing the tuning width, they could almost exactly reproduce the broadening in the width of the channel response function when stimulus contrast was decreased (Liu et al., Fig. 4). Interestingly, they also found that

changes in modeled neural tuning width could alter the width of channel response functions. However, because such broadening is consistent with either a change in SNR or a change in neural tuning width, the authors conclude that it remains impossible to conclusively infer how changes in channel response functions relate to changes in neural tuning. Since it is plausible that low-contrast stimuli evoke weak, noisy responses relative to high-contrast stimuli, the authors argue this is a more parsimonious explanation for their observed data than overturning well-characterized results from the animal physiology literature and inferring that single-unit tuning properties change with contrast. Accordingly, the authors concluded that "changes in channel response functions do not necessarily reflect changes in underlying neural selectivity" (Liu et al., 2018, pg 404).

This report makes an important contribution in its dissection of how model-based analysis methods can be sensitive to features of the data that might vary across conditions (e.g., SNR), and clearly demonstrates that changes in population-level channel response functions cannot and should not be used to infer changes in unit-level neural tuning properties. However, we would like to emphasize that this is not the intended purpose of the IEM approach, which is designed to assess *population-level* stimulus-representations. Any inferences made about single-unit tuning from channel response functions are plagued by the same pitfalls encountered when attempting reverse inference about single-unit neural signals from aggregate measurements.

These issues are not unique to the IEM technique. For example, they also complicate interpretation of results from popular voxel receptive field (vRF) techniques. In these experiments, stimuli traverse the entire visual display while experimenters measure fMRI responses. Then, they fit a receptive field model that best describes how each voxel responds given the visual stimulus (Dumoulin and Wandell, 2008; Wandell and Winawer, 2015). Recent studies have demonstrated that changing task demands (e.g., locus of spatial attention) can change the shape and preferred position of vRFs (Kay et al., 2015; Klein et al., 2014; Sheremata and Silver, 2015; Sprague and Serences, 2013; Vo et al., 2017). While it is tempting to infer that single-neuron RFs change accordingly, it could instead be the case that each neuron maintains a stable RF, but different neurons are subject to different amounts of response gain, altering the voxel-level spatial sensitivity profile measured with these techniques. Moreover, because aggregate measurements like fMRI pool over neurons of different types (excitatory vs. inhibitory), selectivity widths (narrow vs. broad), and cortical layers (e.g., Layer IV vs. Layer II/III), the ability to make inferences about single-unit encoding properties is further limited.

Liu et al.'s (2018) report also highlights that it is important to consider how an encoding model is estimated when comparing channel response functions across conditions. In their work, Liu et al. estimated separate encoding models for each contrast condition (*Fig. 1B*). But because SNR likely differed between conditions, the observed differences between reconstructions may result from

differences in the training sets (i.e. different model fits), or from differences in the testing sets (i.e. different reconstructed activation patterns), or from a combination of the two. More generally, this training scheme can pose a problem for researchers who wish to minimize the effect of known SNR differences between their conditions to study some other variable (e.g. the effect of attention) since it is not possible to unambiguously attribute changes in reconstructed channel response functions to changes in the quality of the model fit or the quality of the representation supported by the population activity pattern, which can both differ between conditions. This problem is roughly akin to reporting a change in a ratio, which can result in changes in the numerator, denominator, or both. One way that others have mitigated this issue is by estimating an encoding model (*Fig. 1B*) using an unbiased (equal numbers of trials from each relevant condition) or neutral (entirely separate task used solely for model estimation) set of data. They then apply that single 'fixed' encoding model to test data from multiple stimulus conditions to reconstruct stimulus representations from each condition. This implementation has the advantage that researchers can avoid problems with comparing channel outputs from different IEMs, so the only difference between conditions is the data used for stimulus reconstruction (*Fig. 1C*). We note that even with such a procedure the central result in Liu et al (2018) could remain true: reconstructions under a fixed encoding model could still broaden with lower contrast. But, as discussed above, this would reflect a change in the quality of the population-level representation rather than provide unambiguous evidence for a change in underlying tuning of individual units. When interpreting results from IEM analyses, it is always critical to consider how the model was estimated.

It would be a mistake to conclude from Liu et al. (2018) that the IEM technique is not useful in the context of its intended purpose: to assay properties of large-scale, population-level neural representations. The quality of these large-scale representations surely depends on myriad factors occurring at the single-unit level. It remains a fascinating question to evaluate how single measurement units, at either the neural or voxel level, change their response properties across visual and task manipulations, but the goal of the IEM approach is to assay the net effect of all these modulations on the superordinate population-level representation. Moreover, few behaviors are guided by single neurons in isolation, and so assaying the joint activity of many neurons, and the resulting population-level representations, is necessary to gain insight into the neural underpinnings of cognition (Graf et al., 2011; Jazayeri and Movshon, 2006; Ma et al., 2006). Indeed, IEMs have been used to assay the time course of covert attention (Foster et al., 2017), understand the consequences of attentional manipulations within WM (Rahmati et al., 2018; Sprague et al., 2016), evaluate how allocation of attention impacts the representation of irrelevant visual stimuli across the visual field (Sprague et al., 2018; Sprague and Serences, 2013; Vo et al., 2017), and probe the influence of top-down expectations on sensory stimulus representations (Kok et al., 2017; Myers et al., 2015).

We do not believe aggregate neural signals will ever be useful for unambiguously inferring single-unit response properties, including feature tuning. However, we see a bright future for collaborative efforts across labs studying similar questions in different model systems, such as human and macaque. When experiments are well-matched between species, both aggregate measurements in humans and single-unit responses in model systems can be used to inform our understanding of neural coding across different cognitive states. In bridging different levels of analysis, Liu et al. (2018) add to the growing literature using data-driven simulations to better understand the relationship between tuning properties and population-level feature representations (Kay et al., 2015; Sprague and Serences, 2013; Vo et al., 2017). Most importantly, their report underscores the importance of avoiding inferences about signal properties, such as single-unit neural feature tuning, that are fundamentally inaccessible via fMRI or EEG, even when using state-of-the-art acquisition and analysis techniques. We hope that future studies take these issues into account when interpreting findings from model-based analyses applied to aggregate measurement tools like fMRI and EEG. Finally, we remain optimistic that the IEM technique, when applied carefully and interpreted appropriately, will continue to reveal how experimental manipulations impact population-level representations of information.

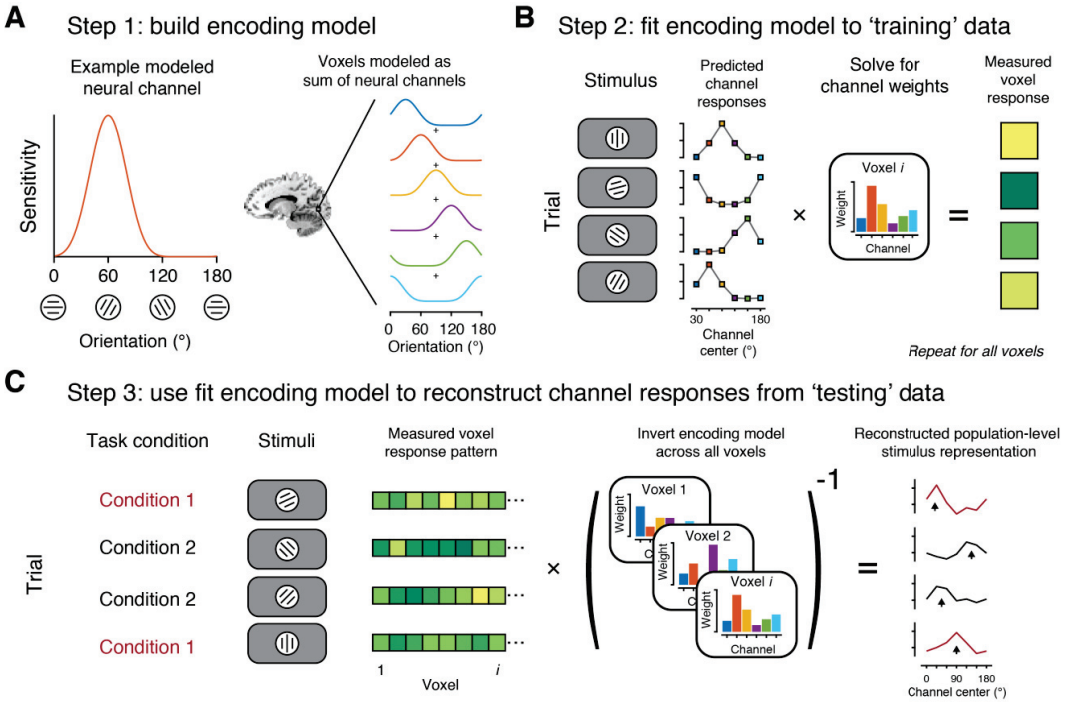

Figure 1. Inverted Encoding Model (IEM): use neural tuning as an assumption to estimate population-level representations

A. The IEM framework assumes that aggregate neural responses (e.g., voxels) can be modeled as a combination of feature-selective information channels (i.e., orientation-selective neural populations). Tuning properties of modeled information channels are experimenter-defined, and often based on findings in the single-unit physiology literature. *B.* Once an encoding model (A) is defined, it can be used to predict how each information channel should respond to each stimulus in the experiment. These predicted channel responses are used to fit the encoding model to each voxel's activation across all trials in a 'training' dataset, often balanced across experimental conditions, or derived from a separate 'localizer' or 'mapping' task. *C.* By inverting the encoding models estimated across all voxels (typically, within an independently-defined region), new activation patterns can be used to compute the response of each modeled neural information channel. This step transforms activation patterns from 'measurement space' (1 number per measurement dimension, e.g., voxel) to 'information space' (1 number per modeled information channel, (A)). These computed 'channel response functions' can be aligned based on the known stimulus feature value on each trial (black arrowheads), and quantified and compared across conditions (e.g., manipulations of stimulus contrast, spatial attention, etc), especially when a fixed encoding model is used for reconstruction (as schematized here). Cartoon data shown throughout figure.

206
207 **REFERENCES:**

208 Brouwer G, Heeger D (2011) Cross-orientation suppression in human visual cortex. J Neurophysiol
209 106:2108–2119.

210 Brouwer G, Heeger D (2009) Decoding and Reconstructing Color from Responses in Human Visual
211 Cortex. J Neurosci 29:13992–14003.

212 Brouwer GJ, Heeger DJ (2013) Categorical clustering of the neural representation of color. J Neurosci
213 33:15454–65.

214 Churchland PS, Sejnowski TJ (1988) Perspectives on cognitive neuroscience. Science 242:741–5.

215 Dumoulin S, Wandell B (2008) Population receptive field estimates in human visual cortex.
216 Neuroimage 39:647–660.

217 Foster JJ, Sutterer DW, Serences JT, Vogel EK, Awh E (2017) Alpha-Band Oscillations Enable
218 Spatially and Temporally Resolved Tracking of Covert Spatial Attention. Psychol Sci 28:929–941.

219 Garcia J, Srinivasan R, Serences J (2013) Near-Real-Time Feature-Selective Modulations in Human
220 Cortex. Curr Biol 23:515–522.

221 Graf ABA, Kohn A, Jazayeri M, Movshon JA (2011) Decoding the activity of neuronal populations in
222 macaque primary visual cortex. Nat Neurosci 14:239–245.

223 Jazayeri M, Movshon JA (2006) Optimal representation of sensory information by neural populations.
224 Nat Neurosci 9:690–696.

225 Kay KN, Weiner KS, Grill-Spector K (2015) Attention reduces spatial uncertainty in human ventral
226 temporal cortex. Curr Biol.

227 Klein BP, Harvey BM, Dumoulin SO (2014) Attraction of position preference by spatial attention
228 throughout human visual cortex. Neuron 84:227–37.

229 Kok P, Mostert P, de Lange FP (2017) Prior expectations induce prestimulus sensory templates. Proc
230 Natl Acad Sci U S A 114:10473–10478.

231 Liu T, Cable D, Gardner JL (2018) Inverted Encoding Models of Human Population Response
232 Conflate Noise and Neural Tuning Width. J Neurosci 38:398–408.

233 Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population
234 codes. Nat Neurosci 9:1432–8.

235 Myers NE, Rohenkohl G, Wyart V, Woolrich MW, Nobre AC, Stokes MG (2015) Testing sensory
236 evidence against mnemonic templates. Elife 4:e09000.

Rahmati M, Saber GT, Curtis CE (2018) Population Dynamics of Early Visual Cortex during Working Memory. J Cogn Neurosci 30:219–233.

Sclar G, Freeman RD (1982) Orientation selectivity in the cat's striate cortex is invariant with stimulus contrast. Exp brain Res 46:457–61.

Scolari M, Byers A, Serences JT (2012) Optimal Deployment of Attentional Gain during Fine Discriminations. J Neurosci 32:1–11.

Sheremata SL, Silver MA (2015) Hemisphere-Dependent Attentional Modulation of Human Parietal Visual Field Representations. J Neurosci 35:508–517.

Sprague TC, Ester EF, Serences JT (2016) Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron 91:694–707.

Sprague TC, Itthipuripat S, Vo VA, Serences JT (2018) Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. J Neurophysiol jn.00059.2018.

Sprague TC, Saproo S, Serences JT (2015) Visual attention mitigates information loss in small- and large-scale neural codes. Trends Cogn Sci 19:215–26.

Sprague TC, Serences JT (2013) Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat Neurosci 16:1879–87.

Vo VA, Sprague TC, Serences JT (2017) Spatial Tuning Shifts Increase the Discriminability and Fidelity of Population Codes in Visual Cortex. J Neurosci 37:3386–3401.

Wandell BA, Winawer J (2015) Computational neuroimaging and population receptive fields. Trends Cogn Sci 19:349–357.