Framed RSA: honoring representational geometry and regional-mean response preferences

JohnMark Taylor (jt3295@columbia.edu)
Columbia University

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Columbia University

Abstract

Representational similarity analysis (RSA) characterizes the geometry of neural activity patterns elicited by different stimuli while discarding their regional-mean activity and the location or orientation of the patterns in multivariate response space. Regional-mean activation analysis serves the complementary purpose of comparing the average population response to different stimuli. Here we introduce a novel method, *framed RSA*, which honors both the geometry and the regional-mean preferences in evaluating model-predicted representations. To achieve this, we augment the stimulus patterns with two reference patterns: the zero-point (origin) and a uniform constant pattern, enabling RSA to incorporate information about the global location, orientation, and mean activation of neural population codes. Framed RSA improves accuracy for both brain region identification (using fMRI data from the Natural Scenes Dataset) and deep neural network layer identification relative to existing RSA approaches. Framed RSA thus combines the strengths of two complementary and traditionally separate analysis approaches, and improves power for model-comparative inference.

Keywords: representational similarity analysis; fMRI; deep neural networks; statistical methods

Introduction

RSA (Kriegeskorte et al., 2008) compares two systems based on their pairwise dissimilarities among a set of stimuli, discarding mean stimulus activations. While this abstraction is useful, mean activation informs the functional role of a brain region and the ease of downstream readout (Prince et al., 2024), and in DNNs the thresholds imposed by nonlinear activation functions entail that responses with the same geometry but differing mean activations can vary in their downstream effects.

To combine these traditionally separate approaches, we introduce a new RSA variant, *Framed RSA*, allowing systems to be compared based both on their geometry and their mean activation profiles within a single framework. In Framed RSA, in addition to the stimulus patterns we add two further patterns in multivariate response space, consisting of the zero-point (origin) and a uniform constant pattern (Figure 1). These reference patterns give a "frame" for measuring the position and orientation of the patterns in multivariate response space, making the resulting RDM sensitive to these facets of the geometry, and importantly the regional-mean activations of the stimuli. We show that Framed RSA improves accuracy on brain region and DNN layer identification relative to existing RSA approaches, demonstrating its value for model-comparative inference.

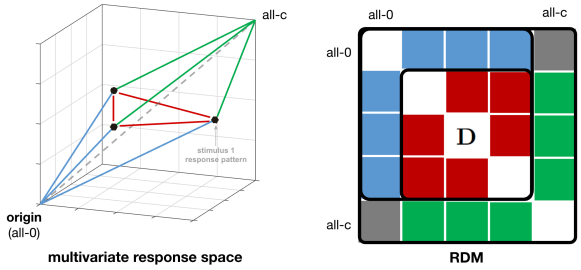

Figure 1: Framed RSA includes distances to two "framing" patterns, adding location, orientation, and mean activation information to the RDM.

Methodology

In Framed RSA, in addition to measuring the pairwise dissimilarity among the patterns elicited by different stimuli, we also measure the dissimilarity between each stimulus pattern and two "framing" patterns: the all-zeros (origin) pattern **z**, and a uniform constant pattern **c** (e.g., [1 1 1 … 1 1 1]). Adding these distances to the resulting RDM gives it sensitivity to the mean activations elicited by

each stimulus. If **x** is an arbitrary stimulus pattern, its squared Euclidean distance (with similar logic applying to other distance metrics) to **z** and **c** are:

$$d_{xz} = ||\mathbf{x} - \mathbf{z}||^2 = ||\mathbf{x}||^2 + ||\mathbf{z}||^2 - 2\mathbf{x}\cdot\mathbf{z} = ||\mathbf{x}||^2$$
$$d_{xc} = ||\mathbf{x} - \mathbf{c}||^2 = ||\mathbf{x}||^2 + ||\mathbf{c}||^2 - 2\mathbf{x}\cdot\mathbf{c}$$

Since $\mathbf{x}\cdot\mathbf{c}$ is simply the sum of the entries of **x** times a fixed constant, the entries of an RDM that includes **z** and **c** can thus be linearly recombined to recover the relative mean activations of the stimulus patterns. Furthermore, the resulting RDM becomes sensitive to translations of the stimulus patterns, as well as to any rotations that change the distances between the stimulus patterns and the framing patterns **z** and **c**, enabling the use of RSA in cases when these parameters are of interest.

To validate Framed RSA and test whether it yields improved model-comparative power in a scenario with a known ground-truth, we tested its performance in brain region identification, and in DNN layer identification. Broadly, this approach tests how often a metric classifies different instances of the "same" processing stage as more similar than instances of different stages (e.g., whether responses from area V1 from different subjects are more similar to each other than responses to V2).

For brain region identification, we used open-source data from the Natural Scenes Dataset (Allen et al., 2022), using data from 14 visually responsive ROIs and eight different subjects. For each method being compared, we computed the RDMs for the ROIs of seven of the eight subjects and those of a left-out test subject, computed the pairwise similarity between each of the test subject's RDMs and the mean RDMs of the training subjects, and tallied the rate at which each test RDM was more similar to the same-region training RDM than to any of the other training RDMs. We compared four conditions: 1) standard RSA using just the stimulus patterns, 2) framed RSA including the stimulus patterns and the all-zeros pattern **z**, 3) framed RSA including the stimulus patterns, the all-zeros pattern **z**, and the constant pattern **c**, and 4) taking the Pearson correlation between the mean stimulus responses of each region, allowing us to test how well mean activation alone can discriminate regions. For #1-3, crossnobis distance was used as the dissimilarity metric, and whitened correlation was used as the RDM comparator. For #3, **c** was tuned to have the same norm as the mean norm of the stimulus patterns. This pipeline was run on repeated samples with varying numbers of stimuli. Finally, to examine whether Framed RSA makes particular pairs of ROIs less confusable, we visualized the difference between the confusion matrices for standard RSA (#1) and Framed RSA (#3).

DNN layer identification followed the same logic, instead using 10 AlexNet (Krizhevsky et al., 2012) instances trained on object recognition from different random seeds, and using layer (e.g., conv1) instead of brain region. To increase the difficulty of the task, isotropic noise was added to the channels at either a low level (noise variance equal to the signal variance in the layer) or a high level (noise variance 15x the signal variance).

Results

Framed RSA improves brain region identification relative to standard RSA (Fig 2A), with the addition of the all-zeros (**z**) pattern and uniform constant (**c**) pattern each providing a further boost. The profile of mean activations is less accurate than Framed RSA with few stimuli, but more accurate with many stimuli. Framed RSA especially improves discrimination among ROIs defined by their mean activation preferences (e.g., to faces or to locations), validating its sensitivity to this information (Fig 2B). For DNN layer identification (Fig 2C), framed RSA outperforms both typical RSA and the profile of mean activations at both low and high levels of noise.

In sum, Framed RSA offers a conceptually simple way to combine RSA and mean activation analysis, improving the power of model-comparative inference.

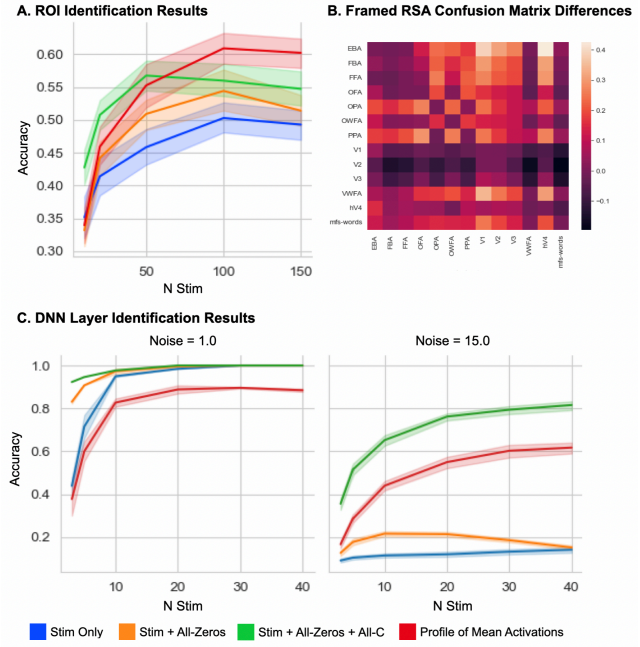

Figure 2: Brain region identification accuracy (**A**), heatmap of confusion matrix differences between framed

and standard RSA (**B**), and DNN layer identification accuracy (**C**).

Acknowledgements

This research was supported by a National Institute of Health Grant (1F32EY033654) to J.T.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience, 25(1), 116-126.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2, 249.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Prince, J. S., Alvarez, G. A., & Konkle, T. (2024) Representation with a capital'R': measuring functional alignment with causal perturbation. In UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models.