

Shared high-dimensional latent structure in the neural and mental representations of objects

Raj Magesh Gauthaman (rgautha1@jh.edu)

Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Michael F. Bonner (mfbonner@jhu.edu)

Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Abstract

Recent work has demonstrated that visual cortex representations of natural scenes are high-dimensional, with a power-law spectrum of stimulus-related variance. However, the statistical structure of the mental representations underlying visual behavior remains unknown — is there a limited subset of latent dimensions that fully captures human behavior on a visual task? Here, we investigate the dimensionality of visual object representations in the human mind and brain by analyzing behavioral and fMRI responses from the large-scale THINGS-data collection using spectral decomposition methods. First, we find that neural representations of objects have a high-dimensional power-law structure throughout visual cortex, replicating previous findings for natural scenes. Next, we show that mental representations of objects, inferred directly from human similarity judgments, have an underlying power-law covariance spectrum, consistent with the power-law structure observed in neural representations of these stimuli. Finally, we show that the dimensionality of shared mental and neural representations increases systematically over stages of visual processing from V1 to hV4 to LOC. Our results suggest that a shared high-dimensional latent structure underlies both mental and neural representations of objects.

Keywords: dimensionality; THINGS-data; mental object representations; fMRI; visual cortex; covariance spectra

Introduction

The neural population codes for natural images in human and mouse visual cortex exhibits high-dimensional statistical structure, with reliable stimulus-related variance distributed over thousands of latent dimensions according to a power law (Gauthaman, Ménard, & Bonner, 2024; Stringer, Pachitariu, Steinmetz, Carandini, & Harris, 2019). This universal representational signature is observed throughout visual cortex and is even shared across individuals (Gauthaman et al., 2024). What can we learn about the statistical structure of *mental* representations underlying visual behaviors using similar spectral analyses? Recent work has used tools from machine learning to interpret the latent dimensions underlying large-scale human object similarity judgments (Hebart, Zheng, Pereira, & Baker, 2020). Here, we directly investigate the statistical structure of mental representations by re-analyzing these behavioral data using a spectral approach and comparing them to visual cortex representations of the same stimuli (Hebart et al., 2023).

Methods

Behavioral and neural datasets. We analyze behavioral and fMRI responses from the large-scale THINGS-data collection (Hebart et al., 2023). In the behavioral experiment (Figure 1A, top), online participants performed a triplet odd-one-out task on object images and a 66-dimensional mental representational space was inferred from responses on

$\approx 4.7 \times 10^6$ trials using Sparse Positive Similarity Embedding (SPoSE) (Hebart et al., 2020). In a separate neuroimaging experiment (Figure 1A, bottom), 3 participants viewed 8,640 images sampled uniformly across 720 object categories from the THINGS database (Hebart et al., 2019) while fMRI BOLD responses were recorded. Here, we focus on activations in visual regions V1 through hV4 and object-selective lateral occipital complex (LOC).

Spectral analyses. We compute the *mental* covariance spectrum by applying principal component analysis (PCA) to the mental representations of object categories extracted from behavioral responses (720 categories \times 66 dimensions). To estimate neural and shared covariance spectra, we use cross-decomposition (Figure 1C, equations), which measures the 8-fold cross-validated spectrum of reliable variance shared between two representations $X \in \mathbb{R}^{N \times P_x}$ and $Y \in \mathbb{R}^{N \times P_y}$ of the same N stimuli (Gauthaman et al., 2024). First, we measure *neural* spectra of reliable stimulus-related variance shared across participants in the fMRI data (8,640 images \times P voxels). Then, we average neural responses across the 12 images within each category (720 categories \times P voxels) and use cross-decomposition to compute the spectra of reliable category-related variance *shared* between these neural and mental representations of object categories.

Results & Discussion

First, we analyze the covariance structure of the 66-dimensional *mental* object representation inferred from behavioral data using SPoSE. These dimensions are sparse, positive and interpretable but have significant pairwise correlations (Figure 1B, inset). Their covariance spectrum obeys a power law over all available dimensions (Figure 1B); we anticipate that relaxing the sparsity and positivity constraints and using a cross-validated spectral approach to measure reliable signal at high ranks will extend this spectrum of mental object representations well beyond 66 latent dimensions.

Next, we estimate the latent structure of *neural* object representations shared across individuals by measuring cross-subject covariance spectra (Figure 1C, top). Throughout visual cortex, we find reliable stimulus-related signal over $\approx 10^3$ dimensions, limited by the number of voxels in each region of interest. This replicates recent findings of universal high-dimensional structure in the cortical representations of natural scenes (Gauthaman et al., 2024).

Finally, we measure the dimensionality of category-related variance *shared* between mental and neural object representations (Figure 1C, bottom). Over the stages of visual processing from V1 to hV4 to LOC, this dimensionality increases up to the limit of detectability in hV4 and LOC.

Conclusion. Our results suggest that high-level cortical object representations measured using fMRI and mental object representations derived from human behavior share the same underlying high-dimensional statistical structure.

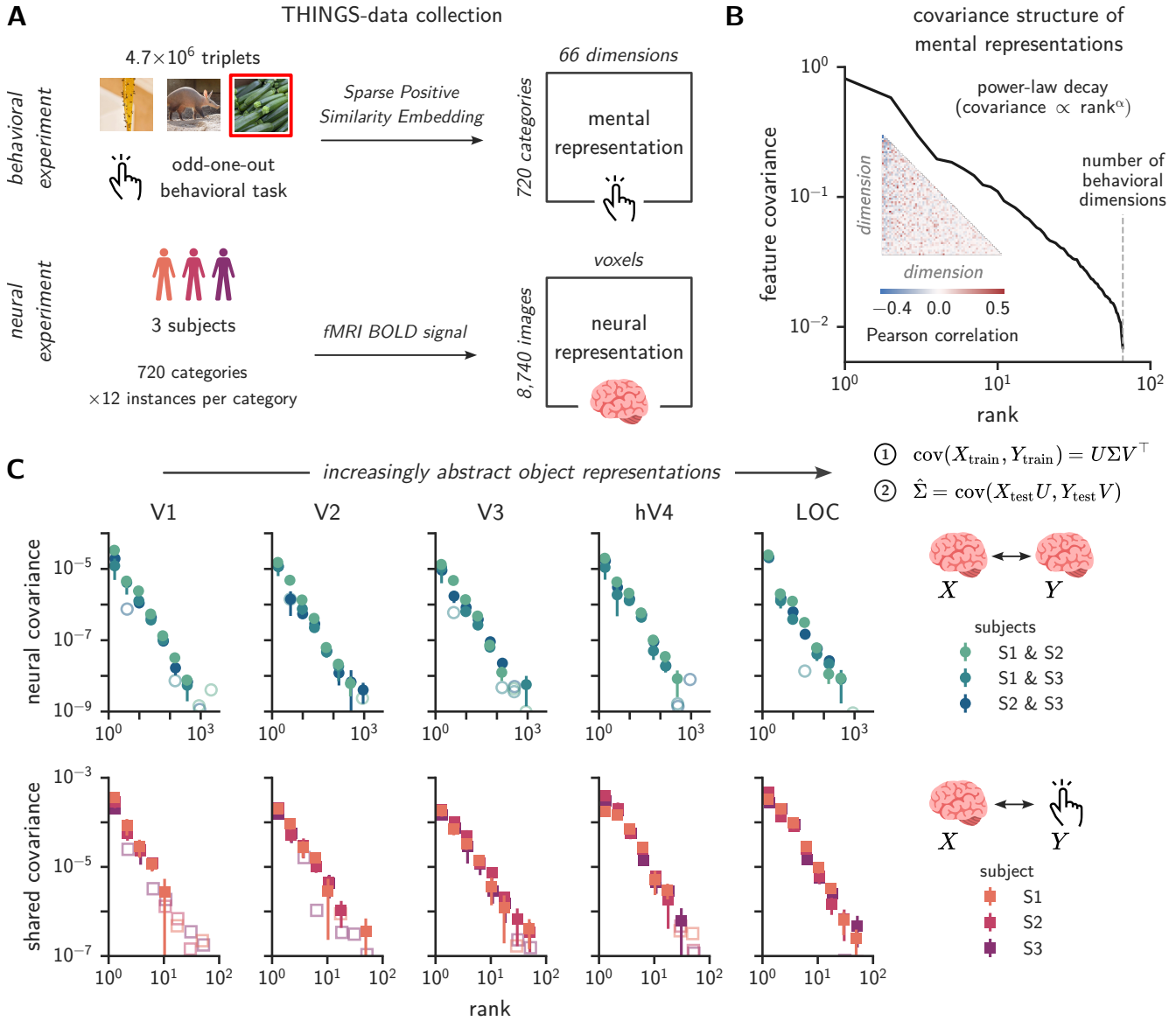


Figure 1: **(A) Behavioral and neural datasets.** (Top) 66-dimensional mental object representational space inferred from a triplet odd-one-out behavioral task using Sparse Positive Similarity Embedding (SPoSE) (Hebart et al., 2020, 2023). An example trial of the task is depicted with images from the THINGS database (Hebart et al., 2019). (Bottom) Cortical activations from an fMRI experiment where 3 participants viewed 8,640 images (12 images from each of 720 object categories). **(B) Covariance structure of mental representations.** Power law covariance spectrum of behavior-derived representations from PCA. (Inset) Pairwise correlations between all 66 behavior-derived feature dimensions. **(C) Neural and shared covariance spectra.** (Equations) Cross-decomposition, a spectral method that estimates reliable stimulus-related variance shared between any two datasets X and Y (see Methods) (Gauthaman et al., 2024) (Top) Neural object representations shared across subjects are high-dimensional throughout visual cortex. (Bottom) Shared mental and neural representations increase in dimensionality from V1 through hV4 to LOC. All spectra are averaged within bins of exponentially increasing width to increase the signal-to-noise ratio at high ranks. Error bars denote standard deviations across 8 folds of cross-validation. Open symbols denote bins of data that are not significantly above chance ($p > 0.001$, permutation tests with shuffled test stimuli, $n = 5000$).

Acknowledgments

PHY-2309135 to the Kavli Institute for Theoretical Physics.

This research was supported in part by a Johns Hopkins Catalyst Award to MFB, Institute for Data Intensive Engineering and Science Seed Funding to MFB and BM, and grant NSF

References

- Gauthaman, R. M., Ménard, B., & Bonner, M. F. (2024). *Universal scale-free representations in human visual cortex*. arXiv. Retrieved from <https://arxiv.org/abs/2409.06843> doi: 10.48550/ARXIV.2409.06843
- Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., ... Baker, C. I. (2023, February). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12. Retrieved from <http://dx.doi.org/10.7554/eLife.82580> doi: 10.7554/eLife.82580
- Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019, October). Things: A database of 1, 854 object concepts and more than 26, 000 naturalistic object images. *PLOS ONE*, 14(10), e0223792. Retrieved from <http://dx.doi.org/10.1371/journal.pone.0223792> doi: 10.1371/journal.pone.0223792
- Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020, October). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11), 1173–1185. Retrieved from <http://dx.doi.org/10.1038/s41562-020-00951-3> doi: 10.1038/s41562-020-00951-3
- Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019, June). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765), 361–365. Retrieved from <http://dx.doi.org/10.1038/s41586-019-1346-5> doi: 10.1038/s41586-019-1346-5