Visualizing the global geometry of population representations of multiple visual object categories with spheres

Andrew David Zaharia (az2522@columbia.edu)
Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Psychology
Columbia University, New York, NY 10027, USA

Alexander Walther (alexander.walther@realeyesit.com)
Realeyes, 2 Riding House St, London W1W 7FA, UK

Nikolaus Kriegeskorte (n.kriegeskorte@columbia.edu)
Mortimer B. Zuckerman Mind Brain Behavior Institute
Departments of Psychology, Neuroscience, and Electrical Engineering
Columbia University, New York, NY 10027, USA

Abstract

Brain computation can be understood as the transformation of representations across stages of processing. The content and format of a representation is reflected in the geometry of stimulus-related response patterns. Here we characterize the representations along the human ventral visual pathway with a new visualization technique called "hypersphere2sphere" (H2S). It takes as input a labeled set of points in a high-dimensional space (the multivariate response space of each cortical region) and fits a hypersphere to represent each category. It visualizes these high-dimensional hyperspheres as a set of spheres in 3D (or circles in 2D), revealing their relative sizes, separations, and overlaps. Using functional magnetic resonance imaging (fMRI), we measured response patterns to 48 images from four categories (faces, bodies, inanimate objects, and scenes). We computed unbiased distance estimates in representational space using crossvalidation. With H2S, we observed the emergence of response pattern clustering, based on category, at the level of the lateral occipital complex. Categories also occupy non-overlapping hyperspheres in face- and place-selective areas, with faces most spread in the former and scenes in the latter. H2S provides a useful perspective on high-dimensional representational geometries that promises new insights on the basis of hemodynamic and electrophysiological brain-activity measurements.

Keywords: ventral visual pathway; representational geometry; visualization; dimensionality reduction; embedding algorithm; multidimensional scaling; human; fMRI

Introduction

To understand brain computation, we need to understand the representations those computations operate on. Brain representations can be characterized in terms of the geometry of the response patterns elicited by a set of stimuli. Each stimulus elicits a response pattern, which corresponds to a point in the high-dimensional space spanned by the neurons (or measurement channels, such as electrode sites or fMRI voxels). To understand the information represented and the format in which it is represented, we must understand the geometry of these points. Linear decoding analyses test whether particular information is amenable to linear readout. Linear decoding can reveal, for example, whether two categories are linearly separable in a brain region's representational space. Representational analyses, such as encoding models, representational similarity analysis, and pattern component modeling (Diedrichsen & Kriegeskorte, 2017), attempt to make detailed predictions about the representations of individual stimuli. Here we assume that the stimuli fall into a number of distributions corresponding to predefined categories. We tackle the challenge of characterizing the global geometry of those distributions in representational space, including the spread of each category and its relationship to the other categories.

Our goal is to visualize each category in terms of its overall spread, and its separation from, and their overlap with, each of the other categories. Therefore we want to abstract from the individual samples of a given category and instead show a single graphical element to represent all category samples as a whole. Instead of showing responses to each individual face, for example, we can show the spread of measured response patterns to faces. Given that our experiments typically sample only tens or hundreds of exemplars from each category, a detailed characterization of the high-dimensional distribution is unrealistic. Here we explore perhaps the simplest possible model of a high-dimensional category-related distribution: a uniform distribution within a hypersphere. We fit a hypersphere to each category in the high-dimensional response space and visualize the categories as a set of spheres in 3D (or circles in 2D). This intuition lends the method its name: hypersphere2sphere (H2S).

Consider the toy example of a 3D response space and a 2D visualization space. The true distributions are three spheres (3D) and the visualization consists of three circles (2D). The H2S visualization, then, is a planar slice defined by the three centers of the spheres. Each sphere is represented by a circle marking the intersection of the sphere and the visualization plane. The resulting 2D set of circles *perfectly* represents the pairwise distances between the sphere centers, the sphere radii, and the overlaps (as measured along the

radii) between the spheres. If there are more categories, then the pairwise distances, radii, and overlaps along the radii are approximately represented. More generally, the visualization can be thought of as a d-dimensional slice through the high dimensional space, where d $(= 2, 3)$ is the dimensionality of the visualization space. If the underlying generating distributions really are uniform within hyperspheres, then up to $d + 1$ categories can be perfectly represented in the d-dimensional visualization space. For other distributions and larger numbers of categories, the visualization conveys an approximation to the overall category-representational geometry.

Unlike multidimensional scaling (MDS) and t-distributed stochastic neighbor embedding (t-SNE, van der Maaten and Hinton (2008)), H2S uses category labels and conveys the overall spread of each category, rather than the location of individual points. This is useful when data for each category is limited, as is often the case in neuroscience. Like MDS, but unlike t-SNE, H2S is distance-preserving; it reduces to MDS when each exemplar is a separate category. In contrast to MDS and t-SNE, H2S is not biased by an imbalance in the numbers of samples across different categories (figure 2e).

We used H2S to visualize the representations of images of faces, bodies, inanimate objects, and scenes in primary visual cortex (V1), the second and third visual areas (V2 and V3), the lateral occipital complex (LOC), the fusiform face area (FFA), and the parahippocampal place area (PPA).

Methods

Hypersphere2sphere is a dimensionality reduction and visualization method that models high dimensional data distributions as uniform D-balls (points uniformly distributed within D-dimensional hyperspheres) and places them as circles in a 2D embedding or spheres in a 3D one. Briefly, the algorithm (figure 1a) follows two steps: (1) A uniform D-ball distribution is fit to the distribution of samples from each category, yielding a center (a D-vector) and a radius (a scalar) for each category. (2) The center and radius parameters of the low-dimensional visualization are optimized to best represent the inferred hyperspheres' center separations, spreads (radii), and overlaps.

For step (1), we use Markov Chain Monte Carlo sampling to estimate the joint posterior of the centers and radii, assuming a uniform D-ball distribution model. Alternatively, one might choose a different center and radius estimator, trading statistical for computational efficiency. In general, the stability and interpretation of the inferred hypersphere will depend on the underlying data distribution and the assumptions motivating the estimator. In low dimensions, the Gaussian differs markedly from the uniform D-ball distribution in terms of its long tail. In high dimensions, however, D-ball and Gaussian distributions converge, concentrating in a thin shell that is well represented by a hypersphere (figure 1b). For step (2), we place the centers using MDS with metric stress (Young & Householder, 1938; Torgerson, 1952) as the optimization criterion. The radii of the visualization spheres are the means of the marginal posteriors of the hypersphere radii.

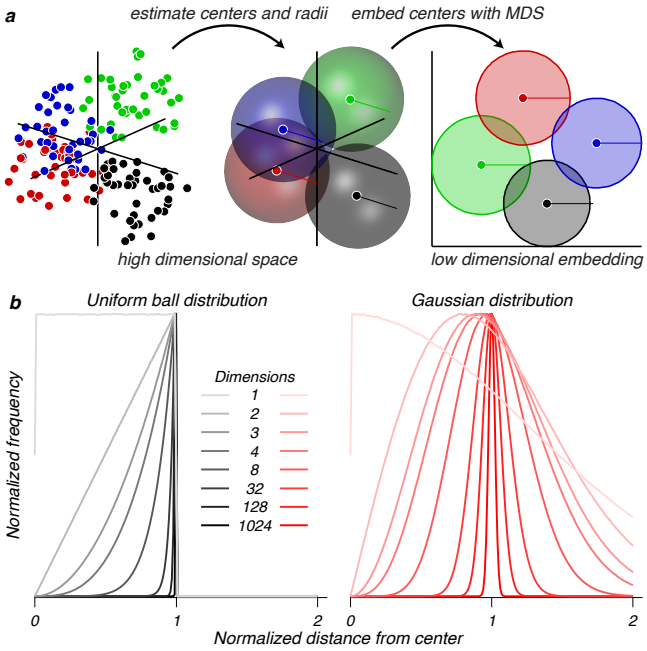

Figure 1: **Hypersphere2sphere intuition and motivation.** (a) The hypersphere2sphere (H2S) algorithm. (b) Histograms of each point's distance from the center of a uniform ball (black) or Gaussian distribution (red) with different dimensionalities. Histograms are normalized so that their maxima match. The distribution types converge in high dimensions to a delta function at 1—the distribution for a uniform unit hypersphere.

Simulations. We tested H2S on simulated data in 3D and 200D, and compared it to other dimensionality-reducing visualization techniques: principal components analysis (PCA), MDS, and t-SNE. For equal radius D-balls touching at one point (figure 2a), all methods give similar results when $D = 3$ (left). When $D = 200$ (right), however, all but H2S exaggerate the spatial separation between the two distributions, and MDS distorts the distribution shapes. For concentric balls with different radii (figure 2b), all visualizations correctly show the smaller (red) distribution surrounded by the larger (black) one. However, the configuration of the distributions changes substantially as the dimensionality grows for all methods except H2S, which stably and correctly represents the geometry regardless of dimensionality. When the smaller inner ball is touching the larger ball at one point (figure 2c), H2S reflects the fact that the two balls are touching at one point in both the 3D and 200D cases, while this fact is completely lost for the other visualizations of the 200D data. Similarly, when two equal-radius hyperspheres intersect (figure 2d), the degree of overlap is correctly reflected, regardless of dimensionality, by H2S, but not by the other visualizations. Finally, while the embedding geometry in MDS and t-SNE is strongly dependent on the relative number of samples for each category, H2S is not biased by having different numbers of samples for each category (figure 2e).

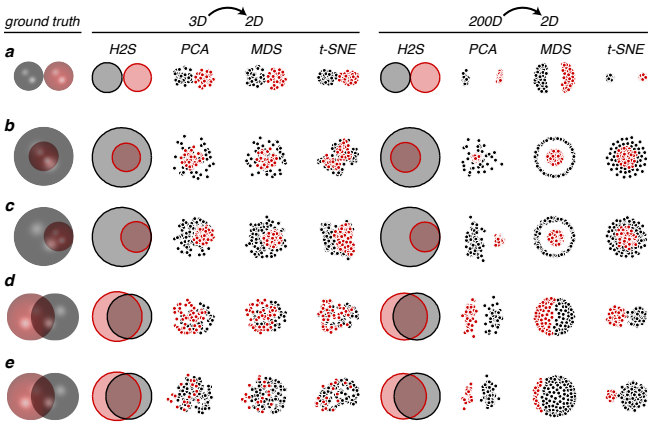

Figure 2: **Comparison of H2S, PCA, MDS, and t-SNE on synthetic examples, in 3D and 200D.**
The leftmost column shows 3D renderings of uniform 3-ball distributions generating the points to the right. The second through fifth columns from the left show two-dimensional embeddings of the three-dimensional distributions, and the four rightmost columns show two-dimensional embeddings of equivalent 200-dimensional distributions. (a) Two touching equal-radius hyperspheres. (b) Two concentric hyperspheres with different radii. (c) One larger hypersphere enclosing a smaller one, touching at one surface point. (d,e) Two intersecting hyperspheres with equal radii and the same number of points (d), or a different number of points (e).

Human imaging experiment. In an fMRI experiment, 24 human subjects were presented with images from visual categories such as animate/inanimate, face/body, and animal/human face (figure 3a), for a total of 48 distinct stimuli (see Walther (2015) and Walther, Diedrichsen, et al. (2016) for details). 320-voxel regions of interest were then defined for six functional visual areas using these and retinotopic mapping stimuli (figure 3b). For each pair of stimuli, the crossnobis distance estimator (Kriegeskorte et al., 2007; Nili et al., 2014; Walther, Nili, et al., 2016) provided an unbiased estimate of the representational distance. That distance matrix can then be visualized using H2S or classical MDS (figure 3c, left and right respectively, in each column).

Results

The H2S visualizations for V1, V2, and V3 on the left of figure 3c show all spheres, and therefore all categories, as highly overlapping. H2S renderings should not be interpreted as fits to the MDS scatters; rather, they are alternative, and likely more accurate, visualizations of high-dimensional geometry with fewer nonlinear distortions than those introduced by MDS. This overlap indicates that each of those three areas do not respond, on a population level, in a manner that clearly separates the stimuli by category. It is apparent, both from their H2S and MDS visualizations, that there is significant spread within each category, indicating that these areas do respond differently to the different stimuli. This matches

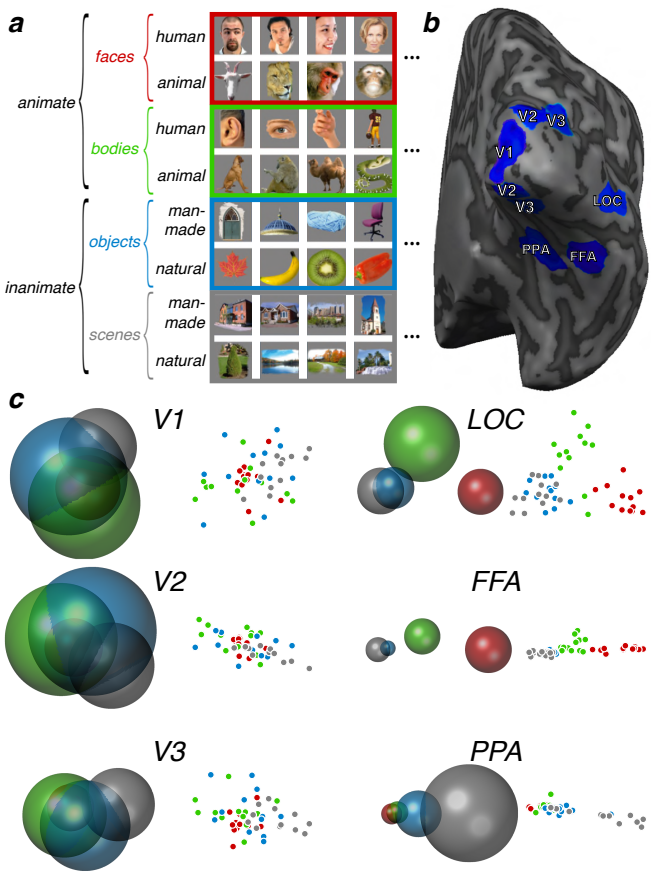

Figure 3: **Category representations in human ventral visual regions.**
(a) Visual stimuli and their categories presented during imaging experiment. (b) Six visual regions in a human brain separately analyzed for their responses to the stimuli. Responses to each stimulus in each area are 320-dimensional: they were measured in a 320-voxel region of interest. (c) H2S (left) and MDS (right) representations of faces (red), bodies (green), objects (blue), and scenes (gray), for the six visual areas in (b). H2S renderings should not be interpreted as fits to the MDS scatters; rather, they are alternative, and likely more accurate, visualizations of high-dimensional geometry. V1, V2, and V3 do not distinguish the different categories well: response distributions are highly overlapping. The LOC separates bodies and faces well while representing objects and scenes more similarly. FFA increases those separations, with faces being particularly well isolated. PPA represents scenes with the most spread—their representation is most different from faces. (a) and (b) and data in (c) are reproduced from Walther (2015) and Walther, Diedrichsen, et al. (2016).

our prior expectations for how these early visual areas should behave: neurons there are known to be selective for local features such as static and moving oriented edges, contours, and textures and are therefore responding to stimulus features that may or may not be predictive of stimulus category.

The more temporal areas on the right of figure 3c represent

visual object categories quite differently. In the LOC, the face and body categories suddenly become quite separable from each other and from objects and scenes, the latter two still overlapping. This makes sense, given that the LOC is thought to signal simple object shape, serving as an early stage of object processing (Grill-Spector et al., 2001). If indeed the LOC is sensitive to overall object shape, then one would expect stimuli with more similar shapes to group together, and have representations more dissimilar from stimuli with very different shapes. This appears to be reflected in the geometry: faces are furthest from objects and scenes, with bodies forming another distinct group. It is surprising then, that objects and scenes are not more strongly separated. This may be due to the choices of images for objects and scenes, but may also point to high-order contextual processing, beginning in the LOC, that gives rise to its sensitivity to stimulus category.

The FFA is a module in the ventral stream defined by its selectivity for faces (Kanwisher et al., 1997); this is supported by the H2S visualization. It shows a representation in which faces are both the largest and most isolated category. The large size reflects the variety of response patterns to different faces which in turn makes the FFA good at discriminating faces. The isolation reflects its ability to discriminate faces from non-face stimuli. The small size and overlapping nature of the object and scene categories point to the area's relative insensitivity to different samples within those categories. The isolation and size of the bodies category suggests that FFA is selective for bodies as well, though to a lesser extent that it is for faces.

The PPA is an even more temporal visual area that responds selectively for "places": images of scenes regardless of the number of objects within the scene (Epstein & Kanwisher, 1998). Again, H2S reflects this selectivity by rendering the place category as a separate sphere. The larger size additionally suggests discriminability of individual scenes. When it was discovered, the PPA was noted to be "weakly" responsive to objects and not at all to faces (Epstein & Kanwisher, 1998)—this is reflected in the relative sizes of the object and face category spheres.

Conclusions

We introduced H2S, a novel method for visualizing the spreads, separations, and overlaps of category-related distributions in high dimensions. Visualizing the representations of natural image categories in human ventral visual regions revealed overlapping categories in the early visual areas V1-3 and strong category separation in LOC, FFA, and PPA. Faces and places had the greatest spread in FFA and PPA, respectively (figure 3c), consistent with their hypothesized functions in the literature. H2S provides a simple and intuitive global picture of the representational geometry of categories that is complementary to visualization techniques, such as MDS and t-SNE, that produce an arrangement of individual exemplars.

References

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, *13*(4), 1–33. doi: 10.1371/journal.pcbi.1005508

Epstein, R. A., & Kanwisher, N. G. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. doi: 10.1038/33402

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10-11), 1409–1422. doi: 10.1016/S0042-6989(01)00073-6

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–11. doi: 10.1098/Rstb.2006.1934

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20600–5. doi: 10.1073/pnas.0705654104

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4). doi: 10.1371/journal.pcbi.1003553

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrica*, *17*(4), 401–419. doi: 10.1007/BF02288916

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(2579-2605), 85. doi: 10.1007/s10479-011-0841-3

Walther, A. (2015). *Beyond brain decoding: Representational distances and geometries*. Unpublished doctoral dissertation, University of Cambridge.

Walther, A., Diedrichsen, J., Mur, M., Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2016). Sudden emergence of categoricality at the lateral-occipital stage of ventral visual processing. *Journal of Vision*, *16*, 407.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. doi: 10.1016/j.neuroimage.2015.12.012

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*(1), 19–22. doi: 10.1007/BF02287916