Three Virtues of Similarity-based Multivariate Pattern Analysis: An example from the human object vision pathway

Andrew C. Connolly*[1], M. Ida Gobbini[1,2], and James V. Haxby[1,3]

[1] *Dartmouth College, Hanover, New Hampshire, USA*
[2] *University of Bologna, Bologna, Italy*
[3] *University of Trento, Trento, Italy*

In *Understanding Visual Population Codes*, N. Kriegeskorte, G. Kreiman, Eds. (MIT, Cambridge) 2012. pp. 335–356

Acknowledgments

This investigation was supported by the National Institutes of Health under Ruth L. Kirschstein National Service Research Award F32MH085433 and by National Institute of Mental Health Grant 5R01MH075706.

*Corresponding author, 6207 Moore Hall, Hanover, NH 03755, andrew.c.connolly@dartmouth.edu

Abstract

We present an fMRI investigation of object representation in the human ventral vision pathway highlighting three aspects of similarity analysis that make it especially useful for illuminating the representational content underlying neural activation patterns. First, similarity structures allow for an abstract depiction of representational content in a given brain region. This is demonstrated using hierarchical clustering and multidimensional scaling (MDS) of the dissimilarity matrices defined by our stimulus categories—female and male human faces, dog faces, monkey faces, chairs, shoes, and houses. For example, in ventral temporal (VT) cortex the similarity space was neatly divided into face and non-face regions. Within the face region of the MDS space, male and female human faces were closest to each other, and dog faces were closer to human faces than monkey faces. Within the non-face region of the abstract space, the smaller objects—shoes and chairs—were closer to each other than they were to houses. Second, similarity structures are independent of the data source. Dissimilarities among stimulus categories can be derived from behavioral measures, from stimulus models, or from neural activity patterns in different brain regions and different subjects. The similarity structures from these diverse sources all have the same dimensionality. This source independence allowed for the direct comparison of similarity structures across subjects ($N = 16$) and across three brain regions representing early–, mid–, and late–stages of the object vision pathway. Finally, similarity structures can change shape in well-ordered ways as the source of the dissimilarities changes—helping to illuminate how representational content is transformed along a neural pathway. By comparing similarity spaces from three regions along the ventral visual pathway, we demonstrate how the similarity structure transforms from an organization based on low-level visual features—as reflected by patterns in early visual cortex—to a more categorical representation in late object vision cortex with intermediate organization at the middle stage.

Introduction

Multivariate techniques for the analysis of functional magnetic resonance imaging (fMRI) provide powerful means for better understanding the cognitive functions of the brain. Chief among these are tools for investigating the representational structure underlying brain activation patterns. Recent advances along these lines include the combination of classifier analyses with cognitive models of stimulus spaces for the purpose of testing specific theories of neural representation (Kay, Naselaris, Prenger, & Gallant, 2008; Mitchell, Shinkareva, Carlson, Chang, Malave, Mason, & Just, 2008), and the use of pattern-based similarity analysis to explore structure in areas where no explicit model is assumed (Edelman, Grill-Spector, Kushnir, & Malach, 1998; O'Toole, Jiang, Abdi, & Haxby, 2005; Hanson, Matsuka, & Haxby, 2004; Kriegeskorte, Mur, Ruff, Kiani, Bodurka, Esteky, Tanaka, & Bandettini, 2008; Kriegeskorte, Mur, & Bandettini, 2008). In this chapter, we make a case for similarity analysis in two parts. First, we provide some background about where similarity analysis fits within the fMRI analysis toolkit. Then we report an experiment that demonstrates its usefulness—highlighting three aspects of similarity analysis that make it a useful tool for analyzing representational structure. The first is representational abstractness. Similarity structures—defined as matrices of pair-wise dissimilarities[1]

[1]Throughout we will follow the convention of referring to similarity structures or similarity spaces in the generic sense. However, when referring to the actual numbers used in the analyses we use the term dissimilarity—technically, a number that goes up the more two items differ from each other. The term distance is used for Euclidean and correlation distances.

between experimental conditions—provide an abstract description of the representational space for a set of cognitive states. These descriptions are amenable to visualization techniques such as clustering or multidimensional scaling, enabling investigators to explore abstract cognitive structure. The second aspect is source independence (Kriegeskorte, Mur, Bandettini, 2008). Because the number of stimuli fixes the number of dimensions, similarity structures provide a common second-order high-dimensional space in which to directly compare results from different sources, including behavioral judgments and neural activity patterns in different brain regions or different subjects—even across species (Kriegeskorte, Mur, Ruff, et al., 2008). The third aspect is transmutability. This refers to the idea that similarity spaces can change shape in well-ordered ways as the sources of the measures change. This can be particularly useful in brain research for investigating how representations are transformed as information is processed in a pathway of connected cortical areas. This last aspect is demonstrated in an experiment reported below in which we demonstrate how the similarity space for a set of face and non-face visual objects transforms form early to late stages of the ventral object vision pathway.

Where does similarity analysis fit within the fMRI analysis toolkit?

The set of tools available for fMRI analysis is under constant development. The standard toolkit consists primarily of a set of techniques for univariate statistical analysis on a voxel-by-voxel basis with the general linear model (GLM) as the core component. There are several readily available software packages that implement these methods (e.g., AFNI, Cox, 1996; FSL, Smith, Jenkinson, Woolrich, Beckman, Behrens, et al., 2004; and SPM, Friston, 2006). Although, the majority of new fMRI studies still rely mainly on standard univariate methods, a complementary set of multivariate methods can reveal information in fMRI data that univariate methods are insensitive to (for recent reviews see Norman, Polyn, Detre, & Haxby, 2006; Haynes & Reese, 2006; O'Toole, Jiang, Abdi, Perbard, Dunlop, & Parent, 2007;). Software packages to aid in the implementation of this new set of techniques are being developed—including PyMVPA (Hanke, Halchenko, Sederberg, Hanson, Haxby, & Pollmann, 2009) and the Princeton MVPA Toolbox for MATLAB^{TM} (Detre, Polyn, Moore, Natu, Singer, Cohen, Haxby, & Norman, 2006). The new multivariate toolkit incorporates analysis techniques developed by the machine learning community, including a large variety of state-of-the-art pattern classifiers. In addition to classification techniques, which have been at the heart of the trend toward multivariate fMRI analysis, techniques for exploring representational structure have begun to appear in the literature as well. Of these, two general approaches can be identified: (1) classification with cognitive models, and (2) similarity analysis. We now present a brief overview of the development of this new toolkit for fMRI beginning with a justification for the adoption of multivariate techniques as a necessary complement to standard analysis techniques.

The standard approach

The standard approach to fMRI analysis involves modeling a time-course of experimental conditions and using the general linear model (GLM) to approximate the magnitude of response for each experimental condition at each voxel (Friston, Jezzard, & Turner, 1994), evaluating the significance of activations using univariate statistics. Typically, data are smoothed or blurred

using a spatial filter, which has the dual effect of increasing the signal-to-noise ratio and lessening the problem of multiple comparisons by reducing the resolution of an image, thereby increasing the size of the unit of analysis from the voxel to the "blob"—composed of non-independent, neighboring voxels. To compare results across subjects, individual brain maps are standardized to a common voxel grid by aligning anatomical landmarks and warping the data to fit a common template. To the extent that active blobs in each subject overlap in the standard space, positive effects are reported as significant outcomes in a random-effects analysis (e.g., *t-test*, *ANOVA*).

This standard approach has two considerable drawbacks. The first is due to spatial smoothing, which results in the loss of informative variation in signal strength occurring at spatial frequencies higher than those corresponding to the smoothness of the processed data. The second—more serious—drawback of the standard approach is that it limits the range of experimental questions to those that can be answered by measuring the magnitude of BOLD response for brain regions with volumes greater than one milliliter. As a result, fMRI studies have largely been concerned with mapping particular cognitive functions onto particular brain regions. Many important discoveries have been made about the function of certain regions using this technique—the fusiform face area (FFA, Sergent, Ohta, & MacDonald, 1992; Kanwisher, McDermott, & Chun, 1997) and the parahippocampal place area (PPA, Epstein & Kanwisher, 1998) being two famous examples in the ventral temporal cortex (VT). However, results can be misleading. In the case of the FFA, the highly replicable finding that activity in this region is greatest during perception of faces, led to the strong claim that the region was exclusively involved in face processing and thus a face-processing "module" in the brain (Kanwisher et al., 1997; Kanwisher & Yovel, 2006). A different view emerges, however, when analyzing how patterns of activity—instead of overall magnitude—inside and outside of areas of peak activity can inform the nature of face representation in this area. Haxby et al. (Haxby, Gobbini, Furey, Schouten & Pietrini, 2001) demonstrated how multi-voxel pattern analysis (MVPA) could be used to show that activity associated with viewing faces is not limited to areas of peak activity identified as the FFA. When peak regions were left out of the analysis, patterns consisting of voxels in the surrounding cortex were sufficient to classify face and non-face activity. Thus Haxby et al. provided evidence of a distributed overlapping system for faces and objects in ventral temporal cortex—a stark contrast to the modular view offered by Kanwisher and colleagues. The public debate that followed highlighted differences in ideological stances about brain organization—modular vs. distributed, and it demonstrated the importance of multivariate pattern analysis as an alternative approach.

A take-away point from that debate is that the tools used for fMRI analysis can have a significant impact on the conclusions that can be drawn about the nature of neural representation. The standard approach to fMRI analysis is limited because it emphasizes mapping particular structures to particular functions. If the cognitive brain were organized as a collection of functional modules—each of an appropriate size to match fMRI resolution, then standard analysis methods should be sufficient to find them.[2] The task remaining would be to close the set of experiments that will map each region to its function. Such a scenario, however, is neither plausible nor is it an outcome expected even by the most committed modularists. Even the strongest versions of cognitive modularity envision a role for modules only in the input and output pathways of the sensory and motor systems; the so-called central cognitive systems responsible for abstract

[2]However, we may still need pattern analysis to read out how information is encoded by neural populations *within* each module.

thought, planning, memory, etc., are said to be non-modular integrating information across sensorimotor modalities (Fodor, 1983). Nevertheless, it is likely that there will be more discoveries of specialized regions like the FFA and PPA—there is a growing list (e.g., the extra-striate body area, EBA, Downing, Jiang, Shuman & Kanwisher, 2001). It is also likely that these regions—as appears to be true of the FFA—will not be strictly modular. While a region may be maximally activated by a narrow range of stimuli, this does not prevent it being at the same time differentially tuned to a wider range of stimuli, or embedded in a larger distributed system of representation where sub-maximal activity can be as informative as the peak activity. The standard analytic approach is well suited for locating regions that are involved in certain processing tasks, though it is unsuited for exploring the non-modular, distributed aspects of neural representation such as the representational content of neural population codes (Mur, Bandetini, Kriegeskorte, 2009).

The multivariate approach

Multi-voxel pattern analysis improves upon the standard approach both in terms of better preservation of fMRI signal and in its ability to measure brain activity related to distributed neural representation. MVP analyses typically involve no or minimal smoothing, and the distortion and averaging associated with spatial normalization is avoided because analyses are typically carried out in individual subjects' native brain spaces. More important than the preservation of high spatial frequency image details, though, is the suitability of MVP analysis for investigating distributed activity. There is strong evidence that information in the cortex is represented as neural population codes. Thus a particular cognitive state is not necessarily represented by the on-off state of a single neuron or cluster of neurons, but instead states are characterized as patterns of graded activity over large ensembles of neurons. This principle has been demonstrated numerous times with the use of recording electrodes implanted in the brains of non-human mammals (e.g, Georgopoulos, Schwartz, & Kettner, 1986). Although fMRI does not provide the direct measurement of neural activity that extracellular recording techniques do, it can be used in a way analogous to ensemble electrode recording to investigate neural population codes. Kamitani & Tong (2005) demonstrated this in a study investigating line-orientation representation in human visual cortex. Using pattern classifiers, they were able to differentiate patterns of BOLD activity in visual cortex evoked by viewing square-wave gratings of different orientations. Although it was long known that the mammalian visual cortex contains columns of orientation preference (Hubel & Wiesel, 1968), it was not thought that such fine-grained organization could be detected using fMRI because of its low resolution—all orientations are represented within the cortical space measured in a single voxel. Classification of voxel patterns was nevertheless possible because just as each orientation column has an orientation preference or tuning curve, each voxel also has an idiosyncratic "tuning curve" reflecting the sum of activity for all of the columns it comprises. Thus as a result of non-uniformity in the distribution of orientation biases across voxels, voxel activity patterns may be used to read out columnar activity with the help of multivariate pattern classifiers (for further discussion and debate of this account see: Haynes & Reese, 2005, Kamitani & Sawahata, 2010, Kriegeskorte, Cusack , & Bandettini, 2010, and Op de Beeck, 2010).

The classification of line-orientation response patterns by Kamitani and Tong provides an example of how MVPA can be used to uncover the representation of a single stimulus dimension. This technique can be thought of as testing a single parameter model of neural representation

where the stimuli used are relatively isomorphic to the dimension being tested. Investigation need not be limited to single parameter models and simple stimuli, however. Researchers have also begun to test more complex cognitive models using multivariate classification techniques. These approaches involve the application of a multidimensional model to represent a complex natural stimulus, and in turn associating brain activity patterns with coordinates in the model space. For example, Kay et al. (2008) modeled natural visual scenes using a set of spatial filters simulating the hypothesized representation in V1. They used a large set of natural scene images to build regression equations that predict the activity in each V1 voxel based on the V1 model representation for each natural image. They then used this encoding model to predict fMRI patterns for novel images that were not in the training set, and showed that the measured fMRI patterns for these stimuli matched the predicted patterns with a high degree of accuracy. Using a similar approach in a different domain, Mitchell and colleagues (Mitchell, et al., 2008) classified brain patterns evoked by reading nouns using a model of semantic representation similar to latent semantic analysis (Landauer & Dumais, 1997) to model word meaning.

But what if there is no explicit model to be tested for a given brain region? For example, we have good evidence about the general activity profile of ventral temporal cortex (VT), but it is less clear what types of specific information or cognitive model is represented by the underlying neural population codes. There are well-known areas of peak activity for faces, places, and living-things vs. non-living things, etc., but what is the structure of representation beyond these relatively coarse dissociations? Classification analysis can be informative here as well, although to a limited degree unless we consider further the similarity structure. In classification analysis we can impose a set of category labels, such as faces, shoes, or hairbrushes, and use classifiers to answer the simple question: does the information discriminate among labeled categories? MVPA has an advantage over the standard approach for answering this question because the answer does not depend on finding significant differences in peak activity. Rather, by pooling over many voxels, slight but reliable differences can be aggregated boosting the ability to reliably discriminate among conditions. However, while classification analyses confer greater power to the task of finding brain activity patterns that distinguish among experimental conditions, the technique doesn't elucidate the underlying structure of representations.

Although, classification analysis provides no immediate insight into the structure of representation, the basis of pattern classification—i.e., distances between high-dimensional pattern vectors—can be used to represent the similarity space for a set of stimuli. Analysis of similarity spaces can in turn be used to uncover an inherent representational structure. Similarity structure analysis has a long history in cognitive psychology (e.g., Rips, Shoben, & Smith, 1973; Solis & Arabie, 1979, Connolly, Gleitman, & Thompson-Schill, 2007) as well as in diverse other disciplines including genetics (Eisen, Spellman, Brown, & Bostein, 1998), ecology (Clarke, 1993), and political science (Jakulin, Buntine, La Pira, & Brasher, 2009). As a result of the broad applicability of similarity analysis, numerous methods have been developed for clustering, visualization, and decomposing similarity structures that can be incorporated into the fMRI toolkit with appropriate modifications. Several studies have demonstrated the utility of similarity-based MVP analysis. For example, O'Toole et al. (2005) used multidimensional scaling to represent the similarity structure of eight categories from the Haxby (2001) dataset. Hanson et al. (2004) calculated the similarities for the same dataset as correlations between values in a hidden layer of a neural network classifier. They used hierarchical clustering to provide a visualization of the similarity relations between conditions. More recently, Kriegeskorte et al. (2008) demon-

strated that representing brain activity for multiple conditions as similarity structures not only provides a way to assess the internal organization of representations—exploiting the representational abstractness of similarity spaces, it also provides a second-order representation that can be used to directly compare results across domains—exploiting the source independence of similarity structures. They collected fMRI data from human subjects and extracellular recording data from monkeys, while both monkeys and humans viewed a common set of visual stimuli. Multivariate patterns—voxel patterns in VT cortex in humans and electrode ensemble patterns from inferior temporal (IT) cortex in monkeys—were used to calculate similarity structures for a large set of stimuli. The resulting similarity structures—notable for a strong segregation of living and non-living objects, as well as interesting structure segregating animal faces from animal bodies—correlated between species. Representations in humans and monkeys exhibited the same category structure and correlated similarity structure even within categories. This suggests that the representations of animate objects in human VT and monkey IT cortex have a structure that is preserved over evolution.

Experiment

As mentioned, VT is known to be specialized for the representation of visual objects while less is known about the neural code underlying these representations. The stimulus set for this experiment was designed to investigate face and object representation primarily in VT. The stimuli include several types of faces and non-face objects. We investigate the classification accuracy of these representations in three stages of the object vision pathway representing early-, mid-, and late- stages of processing. We refer throughout to these regions as early visual (EV) cortex, lateral occipital (LO) cortex, and ventral temporal (VT) cortex—see below for details on their delineation. We then show how the similarity structure transforms from EV through LO to VT, illuminating the functional relationships among these regions. We assess also the stability and reliability of similarity structures both within and between subjects.

Subjects and stimuli. Sixteen healthy subjects (8 men) viewed static grayscale pictures of four categories of faces—human female, human male, monkeys, and dogs—and three categories of objects—houses, chairs, and shoes—while undergoing fMRI scanning. One block of each stimulus category was presented in each of eight runs (1536 images, one every 2 s). Images were presented for 500 ms with 2000 ms inter-stimulus intervals. Sixteen images from one category were shown in each block and subjects performed a one-back repetition detection task. Repetitions were different pictures of the same face or object. Blocks were separated by 12 s blank intervals. One block of each stimulus category was presented in each of eight runs.

fMRI image acquisition. Blood oxygen level dependent (BOLD) MRI images were obtained with gradient echo echoplanar imaging using a Siemens Allegra head-only 3T scanner (Siemens, Erlangen, Germany). Functional images were composed of 32 3mm thick axial images (TR = 2000 ms, TE = 30 ms, Flip angle = 90 x 64 x 64 matrix, FOV = 192 mm x 192 mm) that included all of the occipital and temporal lobes and all but the most dorsal parts of the frontal and parietal lobes. 192 volumes were obtained in each of eight runs.

Structural MRI image acquisition. High-resolution T1-weighted images of the entire brain were obtained in each imaging session (MPRAGE, TR = 2500 ms, TE = 4.3 ms, flip angle = 8 x 256 x 256 matrix, FOV = 256 mm x 256 mm, 172 1 mm thick sagittal images).

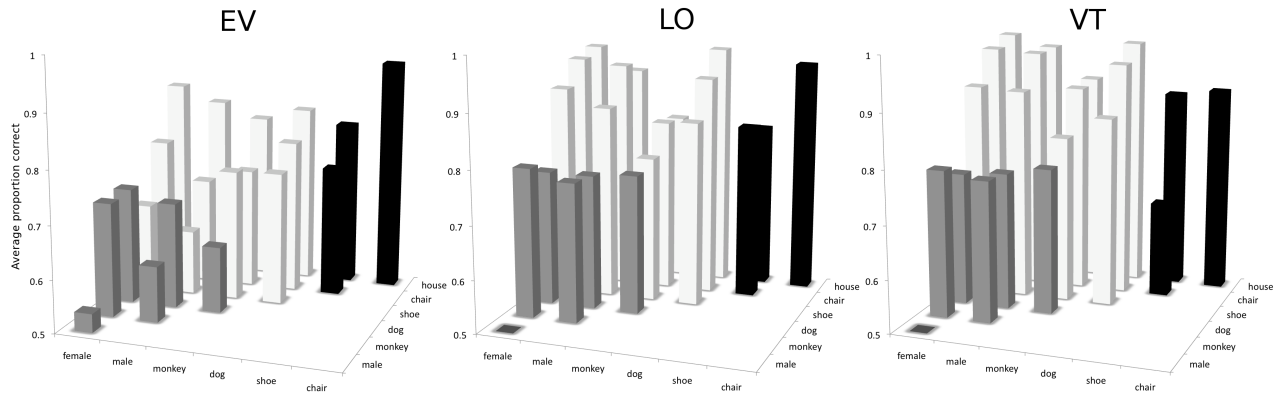

Figure 1: Binary classification accuracy for three brain regions: early visual (EV), lateral occipital (LO), and ventral temporal (VT); gray bars = living–living pairs, white bars = living–non-living pairs, black bars = non-living–non-living pairs. All accuracies are significantly above chance—except for female vs. male human face in all rois—at at least the $p < 0.05$ level via t-test with subjects as the random variable.

Regions of interest. Masks for three regions of interest (ROIs) were hand-drawn for each subject based on structural landmarks in the high-resolution anatomical images. These regions are: (1) early visual cortex (EV), including all of bilateral calcarine sulci, (2) mid-level object vision region—bilateral lateral inferior occipital cortex (LO), and (3) the late object-vision region—bilateral ventral temporal cortex (VT), including the inferior temporal, fusiform, and lingual/parahippocampal gyri. The anterior boundary of LO was set at $y = -70$ mm in Talairach coordinates. The posterior and anterior boundaries of VT were set at $y = -70$ and $y = -30$ mm, respectively. The drawing of masks and all analyses were carried out in subject native space. Although we did not collect data to identify functionally defined regions, we make the following assumptions: (1) EV contains all of V1 and possibly portions of V2; (2) LO contains the object sensitive lateral occipital complex (LOC), EBA, and MT; (3) VT contains the FFA and PPA.

Preprocessing. Preprocessing of the fMRI data included slice-timing correction, volume registration to correct for minor head movements, correction of extreme values (despiking), and mean correction for each run. Eighteen temporally consecutive brain images covering the time period of each stimulus block—taking into account the hemodynamic lag—were averaged on a voxel-by-voxel basis to produce a pattern for each block. These block-mean patterns—one pattern per condition per run—provide the basis for the analyses reported below.

Classification analysis. We tested the discriminability of patterns within our three ROIs using a simple binary nearest-neighbor classification technique with leave-one-out N-fold cross-validation. Template patterns were created for each stimulus class by averaging patterns over seven of the eight experimental runs. Feature selection for each cross-validation fold included application of a threshold mask based on overall visual responsiveness of voxels using an omnibus general linear test (F-statistic, $p < 10^{-6}$) for all conditions modeled separately vs. baseline based on a GLM analysis using the seven training runs. Patterns from the left-out run were then classified based on Pearson correlation (nearest-neighbor) with each template on a strictly pair-wise basis such that chance performance was 0.5. This was repeated for each of eight data folds averaging results across

folds. Classifier performance for each stimulus pair is presented in Figure 1. Vertical bars indicate the mean performance averaged across subjects. All accuracies—except for female *versus* male human faces—were significantly above chance performance at $p < 0.05$ level, using one-sample t-test with subjects as the random variable and 0.5 as the mean for the null-hypothesis; one-tailed; uncorrected for multiple comparisons. These results demonstrate good discriminability between our conditions throughout the three regions we tested with the notable exception of the inability to discriminate between male and female faces in any region.

We measured dissimilarity between conditions using a variation on correlation distance, which conventionally defines dissimilarity between vectors j and k as:

$$(1)\ \delta_{jk} = 1 - r_{jk}$$

where r_{jk} is the Pearson correlation between two patterns. In the analyses that follow, we use a version of this that takes into account the internal reliability of conditions. The formula in (1) assumes that maximum observable correlation between conditions is 1.0. However, given the noisy nature of fMRI data, it is unlikely under optimum circumstances that even the same stimulus will yield perfectly correlated patterns across different observations. Therefore, instead of using 1 as an upper bound, we subtract the average between-condition correlation from the average within-condition correlation calculated using different folds of the data. Specifically we use the formula:

$$(2)\ \delta_{jk} = \sqrt{(r_{jj}^2 + r_{kk}^2)/2} - (r_{jk} + r_{kj})/2$$

where r_{jk} is the average Pearson correlation across data folds between template patterns (based on 7 runs) and hold-out patterns (1 run) for conditions j and k, respectively. Pair-wise dissimilarity patterns are plotted in Figure 2 (top). These dissimilarity matrices are similar to the pair-wise classification results. The correlations between dissimilarity matrices and classification results show high agreement within brain regions ($r = 0.92$, $r = 0.89$, and 0.80 for VT, LO, and EV, respectively).

The depiction of the similarity structure in the top of Figure 2 makes evident an increasingly categorical organization from EV to LO to VT: The dissimilarities for face versus non-face pairs increase (white bars) whereas dissimilarities for face versus face pairs (gray bars) and non-face versus non-face pairs (black bars) decrease. To better appreciate how similarity structure is transformed from one area into the other, it is useful to visualize the data using hierarchical clustering (Figure 2 middle) and multidimensional scaling (MDS, Figure 2 bottom). The tree structures were created using single-linkage nearest-neighbor hierarchical clustering and the MDS plots were created using metric 2-dimensional MDS (Torgerson, 1958). The face versus non-face distinction is captured in the structure of the tree diagram for VT as two main branches and in the x-axis for the MDS solutions for LO and VT. There is less evidence for the face–non-face distinction in EV, where human faces are more similar to shoes than to animal faces. Thus moving from region to region, the face–non-face (alternatively, animate–inanimate) distinction is non-existent in EV, begins to become evident in LO, and is the dominant feature of similarity

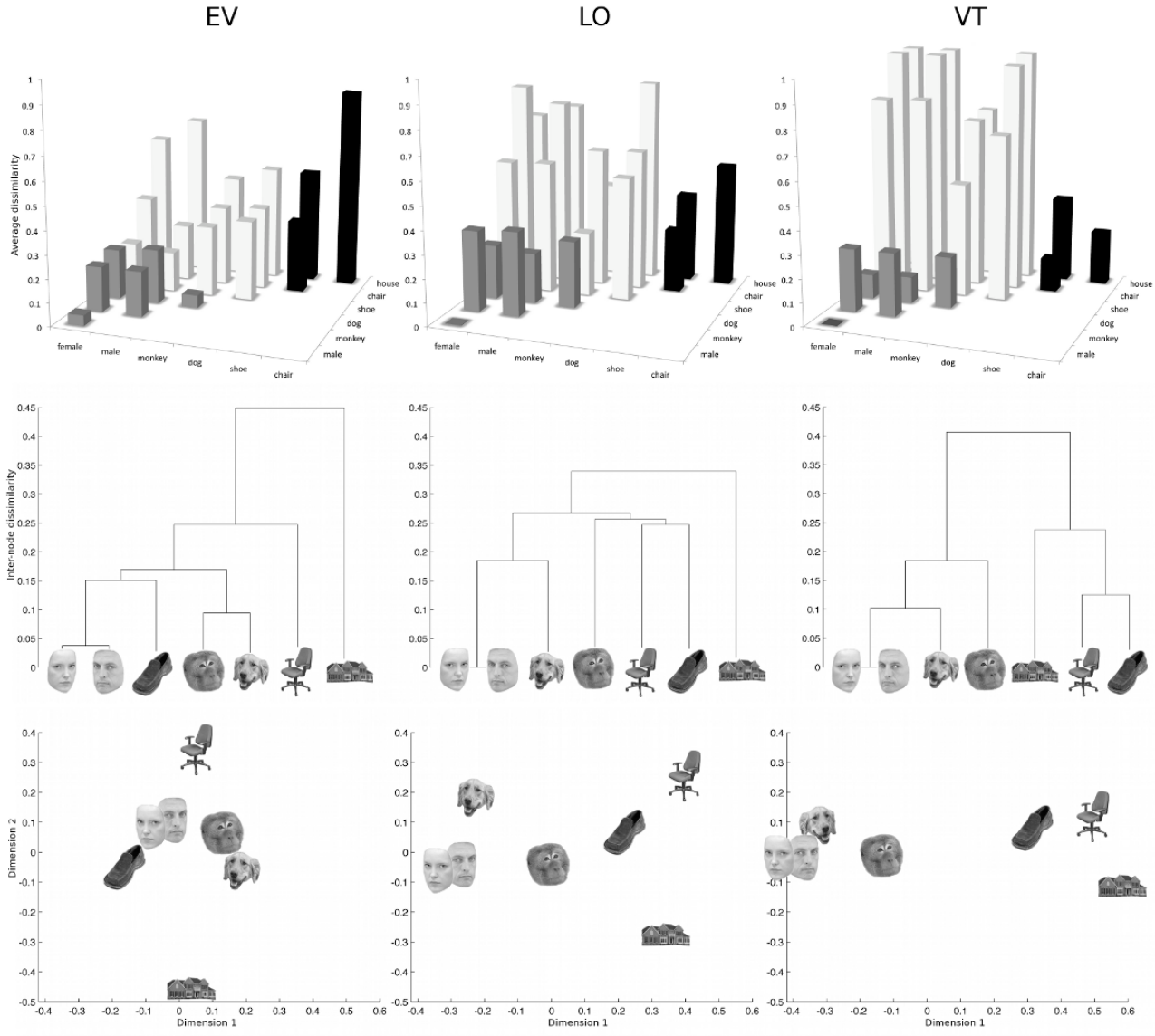

Figure 2: Three depictions of similarity structure for 3 brain regions: ventral temporal (VT), lateral occipital (LO), and early visual (EV) cortex. Top: Pair-wise dissimilarities calculated using the formula in equation (2) and averaged across subjects; gray bars = living–living pairs, white bars = living–non-living pairs, black bars = non-living–non-living pairs. Middle: Hierarchical clustering of pair-wise dissimilarities. Bottom: 2-Dimensional solutions for pair-wise distances using metric MDS. Note: To better visualize continuity between regions, the axes in the MDS solution for EV have been rotated from the original output. Results are based on category-average response patterns (averaged across exemplars within each category). Each category is represented by one exemplar in the figure.

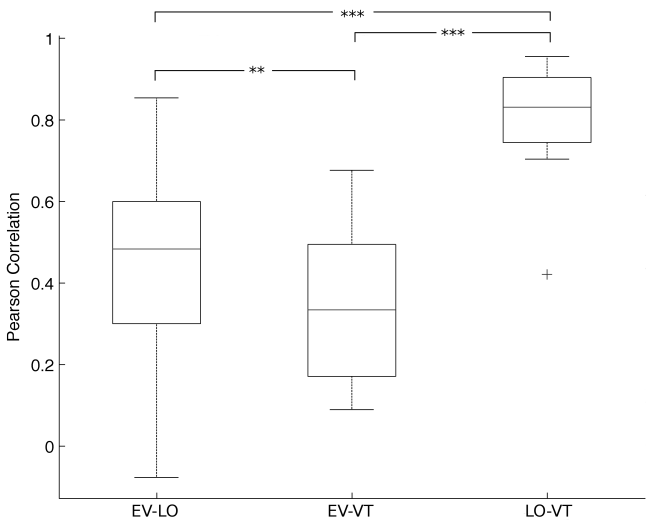

Figure 3: Correlations between similarity structures from different brain regions. Boxplots show the distribution of correlations across subjects. The center line marks the median, the upper and lower edges of the boxes mark the upper and lower quartiles, and the whiskers span the range with outliers marked by '+'. Differences between distributions were tested for significance using paired t-tests. Asterisks indicate p-values: $^{***}p < .001$; $^{**}p < .01$.

structure in VT, consistent with Kriegeskorte et al. (2008). Another interesting dimension is reflected in the y-axes of the MDS solutions bounded at the extremes by houses and chairs. (Note that this dimension is the dimension of maximum variance for the solution in EV. As such, convention is to present it along the x-axis. To make visual comparisons with VT and LO easier, we rotated the solution in EV from its original configuration—not shown—switching the x- and y-axes and reflecting values about the zero point on the y-axis.) The persistence of this dimension reflects the functional continuity between EV and the other regions. It is likely that organization of representations along this dimension reflects similarities in low-level visual attributes, however, the specific nature of those attributes is an open question.

Comparing similarity across brain regions

To explore how similarity space transforms from one brain region to the next, we directly compared similarity structures across regions in two ways. First, for each subject we calculated the correlation between similarity structures from each brain region (Figure 3). Similarity structures from VT correlate more with similarity structures from LO than with EV ($t(15) = 7.10$, $p < .001$), and correlations between EV and LO are higher than those between EV and VT ($t(15) = 2.98$, $p < .01$). This pattern supports the idea that differences in similarity structures reflect successive transformations of representations from early to later stages of the visual processing stream. Second, to gain further insight into the nature of the transformations between regions we used individual differences (three-way) multidimensional scaling (INDSCAL, Carroll and Chang, 1970; Takane, Young, de Leeuw, 1977). Input to the INDSCAL analysis consisted of all 45 dissimilarity matrices—one each from three brain regions for 15 subjects—omitting one subject (S12) whose results were unreliable, see Figure 6. The INDSCAL analysis produces two solution spaces (Figure 4) based on these dissimilarity matrices. The group space (Figure 4 left) represents the best fit for estimated distances between stimuli for all 45 dissimilarity matrices. The individual space (Figure 4 right) represents the weights needed to derive individual distance spaces from the group space. Thus, the estimated distance between two stimuli along a given dimension for an

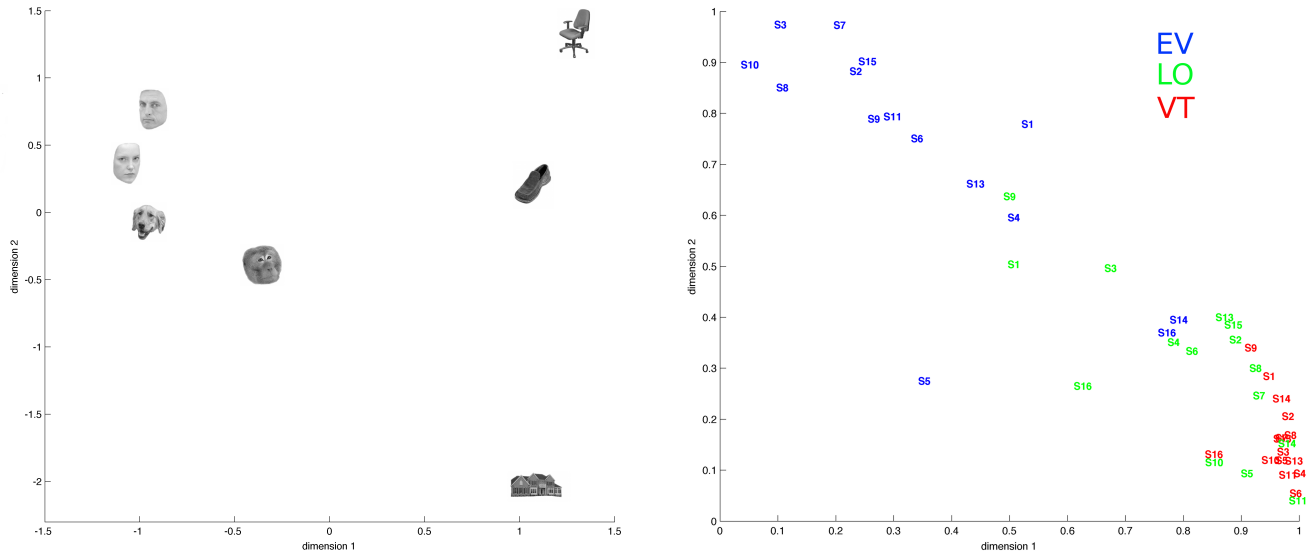

Figure 4: Individual differences multidimensional scaling (INDSCAL, Carroll & Chang, 1977) solution for 45 dissimilarity matrices: three matrices—EV, LO, and VT—from each of 15 subjects (omitting subject 12, see Figure 6). Left: The compromise two-dimensional MDS solution representing the best fit across all 45 matrices. Right: Weights for each of the contributing similarity matrices coded by subject identity (e.g., 'S16' stands for subject number 16) and brain region: red = VT, green = LO, and blue = EV. Weights indicate the degree to which variation along each dimension explains variance in each contributing distance matrix. Variance in distance matrices from VT is mostly explained by the face *versus* non-face distinction captured by dimension 1, whereas variance in distance matrices from EV is explained mostly by dimension 2. Weights for matrices from LO fall between those for EV and VT on both dimensions.

individual is proportional to the distance between those stimuli on that dimension in the group space and the weight for that dimension for that individual. Figure 4 (right) shows how similarity structures from VT load most heavily on the face versus non-face dimension—dimension 1—and not much on dimension 2. Structures from EV load most heavily on dimension 2—presumably reflecting low-level visual similarity—and not much on dimension 1. Finally, structures from LO fall in between EV and VT on both dimensions.

Stability of similarity structures

Another important factor in assessing the validity of the similarity structures is their stability. To test within-subject stability we calculated the correlation between similarity structures for split-halves of each subject's data correlating structures for odd and even runs for each ROI. This was done for a range of mask sizes from 50 to 1000 voxels, where possible, to assess stability as a function of pattern size. Voxels were added to the analysis in order of highest to lowest visual responsiveness based on an F-statistic from an omnibus general linear test for all conditions vs. rest based on data from all runs. Subjects were dropped from the analysis as the number of voxels exceeded their mask size. The results of the split-half analysis are displayed on the

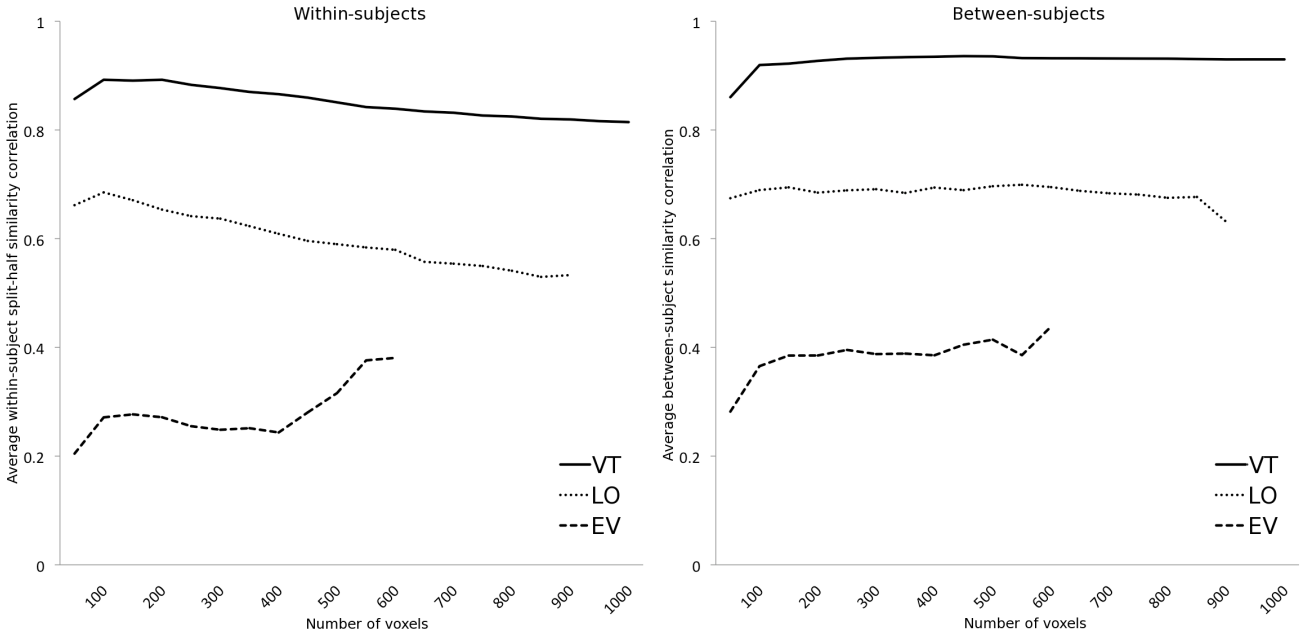

Figure 5: Stability for similarity structures as a function of the number of voxels in three brain regions: VT, solid lines; LO, dotted lines; and EV, dashed lines. Left: Within-subject stability calculated as the correlation between similarity structures from odd and even runs for each subject. Curves represent the mean of these correlations across subjects. Right: Between-subject stability calculated as the correlation between similarity structures from different subjects (based on all 8 runs). Curves represent the average correlation across all pairs of subjects. Both left and right: Curves represent means calculated using 15 out of the 16 subjects because one subject (S12) did not have replicable similarity structures—see Figure 6.

left side of Figure 5. One subject's split-half correlations were well below the average for the group, especially in VT. Figure 6 shows the relative positions of all subjects with respect to split-half correlation—averaged across all voxel-size masks—for VT and LO. Because the similarity structures were not replicable in this subject, the curves in Figure 5 were calculated as averages using the 15 remaining subjects. VT produced the highest split-half correlation for similarity structures, followed by LO, followed by EV. This pattern is repeated when comparing similarity structures between subjects. The right side of Figure 5 shows the average correlation between similarity structures across all pairs of subjects ($N = 15$) as a function of brain region and number of voxels. Here, the similarity structure for each subject was calculated using the average patterns across all 8 runs. Note that the between-subject stability of the similarity structure is slightly higher than the within-subject stability—this is not surprising because the between-subject estimate was based on all of the data, whereas the within-subject estimate was based on split-halves. This effect disappears when the between-subject stability is estimated using half the data, resulting in roughly equivalent values for between versus within (data not shown).

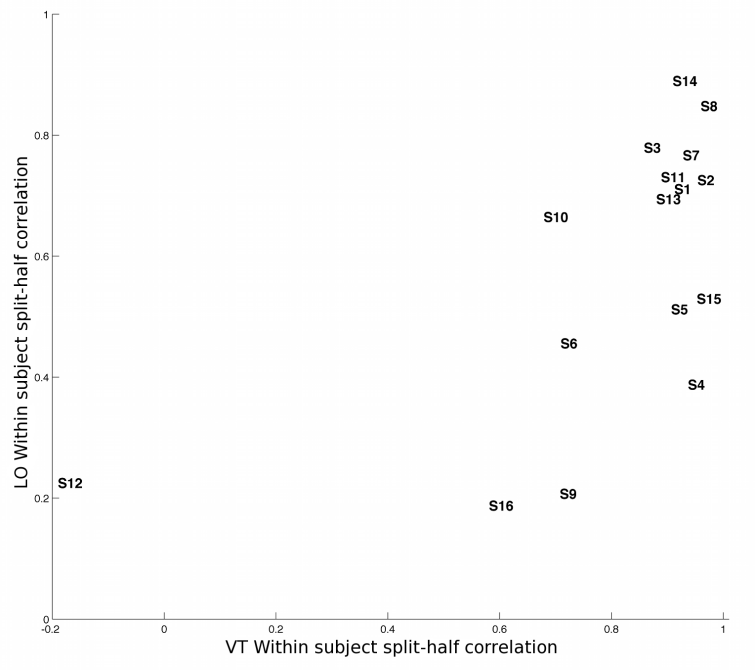

Figure 6: Split-half correlations of similarity structures in VT and LO for 16 subjects. Each point represents a single subject (e.g., 'S16' stands for subject number 16) with the value of the average split-half correlation for that subject in VT on the x-axis and for LO on the y-axis. Averages were calculated across voxel-size masks from 50 to 1000 voxels with 50 voxel increments (see Figure 5). Similarity structures were highly replicable in 15 out of 16 subjects.

Discussion

The primary feature of the similarity spaces in LO and VT is the distinction between faces and objects. This finding is in line with other studies that have found strong distinctions along the lines of animate versus inanimate boundaries (e.g. Hanson, Matsuka, & Haxby, 2004, O'Toole, Jiang, Abdi, & Haxby, 2005, Kriegeskorte, et al., 2008, Mahon, et al., 2009, Caramazza & Shelton, 1998). Within faces, there was no discernible difference between male and female faces, and dog faces were more similar to human faces than were monkey faces. The reason why dog faces should be more similar to human faces than monkey faces is a matter of speculation. Perhaps, this reflects greater familiarity of dog faces compared to monkey faces, or a greater perceived expressiveness for dogs than monkeys. Possibly the effect reflects something idiosyncratic to our stimuli—further experimentation would be needed to know for sure. Among inanimate objects, shoes and chairs were more similar to each other than to houses.

A notable aspect of the similarity structure—especially in VT—is remarkably high replicability both within and between subjects—providing strong evidence for a common code in VT for the representation of faces and objects. Replicability and between-subject correspondence were less for similarity structures in EV. Thus, similarity structures in EV were more variable as reflected by both within and between-subject comparisons. We assume activations in EV to be driven by low-level visual features. As such, variation in low-level visual features within experimental conditions may have contributed significant noise to the similarity calculations which involved averaging across stimuli within each category. Stimuli within each condition consisted of many different exemplars, each with slightly different characteristics—outline contours, points-of-view, etc.—all of which may have been a source of low-level visual variation. In addition, subjects were not required to fixate a central point and were thus free to move their eyes while viewing the stimuli, possibly introducing more noise into the analysis.

Although similarity structures in EV were less consistent than those in LO and VT, there was a reliable shape to the similarity spaces in EV as captured by dimension 2 in the INDSCAL analysis, providing evidence that activation corresponding to low-level visual attributes was captured in the structure of the similarity spaces in EV. The orderly translation of weights from dimension 2 to dimension 1 moving from EV to LO to VT provided direct evidence of the functional relationships between these regions in that the representational content can be seen to transform incrementally from early to mid to late stages of the object vision pathway.

Conclusion

These experimental results highlight the virtues of similarity-based MVP analysis. The abstract nature of similarity structures allowed for the depiction of cognitive structure in terms of the relationships among stimuli. These relationships were explored using cluster analysis and multidimensional scaling helping to reveal the underlying structure of representation in three brain regions. Rather than the engagement or disengagement of specific areas as a function of experimental conditions, similarity analysis reveals cognitive states represented as distributed patterns in larger regions of cortex. The source independence afforded by the fixed dimensional space of similarity structure allowed for the direct comparison of structures across subjects revealing a highly replicable common code for faces and objects in VT. This ability to compare similarity structures across subjects satisfies an important desideratum in cross-subject MVP analysis. Finally, the comparison of similarity structures across functionally connected brain regions—representing three stages of the object vision pathway—reveals how representations are transformed from early visual cortex through to VT.

Similarity-based analyses provide a powerful means for uncovering the structure of cognitive representations present in fMRI data. Yet, we have only scratched the surface of its full potential. There are numerous available techniques for analyzing similarity structure that have been developed outside the realm of fMRI that could be incorporated into the fMRI toolbox for the further exploration of cognitive representation. Additional techniques include clustering algorithms (e.g., Spectral Clustering, Ng, Jordan, & Weiss, 2002) and additional methods for analyzing individual differences (e.g., INDCLUS, Carroll & Arabie, 1983; DISTATIS, Abdi, Dunlop, & Williams, 2009). Methods for individual differences multidimensional scaling and clustering provide means for directly assessing the relative contributions of various dimensions that shape similarity spaces from different domains—including, brain regions, individual subjects, or alternate measures like similarity judgments.

References

Hervé Abdi, Joseph P Dunlop, and Lynne J Williams. How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (distatis). *Neuroimage*, 45(1):89–95, Mar 2009.

A. Caramazza and J. R. Shelton. Domain-specific knowledge systems in the brain the animate-inanimate distinction. *J Cogn Neurosci*, 10(1):1–34, Jan 1998.

J. D. Carroll and P. Arabie. Indclus: An individual differences generalization of the adclus model and the mapclus algorithm. *Psychometrika*, 48:157–169, 1983.

K. R. Clarke. Non-parametric multivariate analyses of changes in community structure. *Austral Ecology*, 18:117–143, 1993.

Andrew C Connolly, Lila R Gleitman, and Sharon L Thompson-Schill. Effect of congenital blindness on the semantic representation of some everyday concepts. *Proc Natl Acad Sci U S A*, 104(20):8241–8246, May 2007.

R W Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res*, 29(3):162–73, Jun 1996.

G. Detre, S.M. Polyn, C.D. Moore, V.S. Natu, B.D. Singer, J.D. Cohen, J.V. Haxby, and K.A. Norman. The multi-voxel pattern analysis (mvpa) toolbox. In *Poster presented at the Annual Meeting of the Organization for Human Brain Mapping*, 2006.

P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, Sep 2001.

S Edelman, K Grill-Spector, T Kushnir, and R Malach. Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, 26(4):309–321, DEC 1998.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998.

R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, Apr 1998.

J. A. Fodor. *Modularity of Mind: An Essay on Faculty Psychology.* Cambridge, Mass.: MIT Press., 1983.

K.J. Friston, P. Jezzard, and R. Turner. Analysis for functional mri time-series. *Human Brain Mapping*, 1:153–171, 1994.

A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, Sep 1986.

Michael Hanke, Yaroslav O Halchenko, Per B Sederberg, Stephen José Hanson, James V Haxby, and Stefan Pollmann. PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, 7(1):37–53, 2009.

Stephen José Hanson, Toshihiko Matsuka, and James V Haxby. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage*, 23(1):156–66, Sep 2004.

J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, Sep 2001.

John-Dylan Haynes and Geraint Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci*, 8(5):686–691, May 2005.

John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, 7(7):523–534, Jul 2006.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J Physiol*, 195(1):215–243, Mar 1968.

A. Jakulin, W. Buntine, T.M. La Pira, and H. Brasher. Analyzing the u.s. senate in 2003: Similarities, clusters, and blocs. *Political Analysis*, 17:291–310, 2009.

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nat Neurosci*, 8(5):679–685, May 2005.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–4311, Jun 1997.

Nancy Kanwisher and Galit Yovel. The fusiform face area: A cortical region specialized for the perception of faces. *Philos Trans R Soc Lond B Biol Sci*, 361(1476):2109–2128, Dec 2006.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, Mar 2008.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci*, 2:4, 2008.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–41, Dec 2008.

T. K. Landauer and S.T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

Bradford Z Mahon, Stefano Anzellotti, Jens Schwarzbach, Massimiliano Zampini, and Alfonso Caramazza. Category-specific organization in the human brain does not require visual experience. *Neuron*, 63(3):397–405, Aug 2009.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, May 2008.

Marieke Mur, Peter A Bandettini, and Nikolaus Kriegeskorte. Revealing representational content with pattern-information fmri–an introductory guide. *Soc Cogn Affect Neurosci*, 4(1):101–109, Mar 2009.

A. Ng, M. Jordan, and Y. Weiss. *Advances in Neural Information Processing Systems 14*, chapter On spectral clustering: analysis and an algorithm. MIT Press, 2002.

Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends Cogn Sci*, 10(9):424–430, Sep 2006.

Alice J O'Toole, Fang Jiang, Hervé Abdi, and James V Haxby. Partially distributed representations of objects and faces in ventral temporal cortex. *J Cogn Neurosci*, 17(4):580–90, Apr 2005.

Alice J O'Toole, Fang Jiang, Hervé Abdi, Nils Pénard, Joseph P Dunlop, and Marc A Parent. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J Cogn Neurosci*, 19(11):1735–1752, Nov 2007.

L.J. Rips, E.J. Shoben, and E.E. Smith. Semantic distance and verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 14:665–681, 1973.

J. Sergent, S. Ohta, and B. MacDonald. Functional neuroanatomy of face and object processing. a positron emission tomography study. *Brain*, 115 Pt 1:15–36, Feb 1992.

Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy E J Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, Rami K Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J Michael Brady, and Paul M Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23 Suppl 1:S208–19, 2004.

S. D. Soli and P. Arabie. Auditory versus phonetic accounts of observed confusions between consonant phonemes. *J Acoust Soc Am*, 66(1):46–59, Jul 1979.

Y. Takane, Forrest W. Young, and J. de Leeuw. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42:441–451, 1977.

W. S. Torgerson. *Theory and methods of scaling.* NewYork: Wiley, 1958.