Defining voxelsets using rich psychological models

Greg Detre

August 28, 2006

Contents

1 Acknowledgments

I'm extremely grateful to Sabine Kastner for permission to use her delayed saccade data, which proved invaluable as a benchmark dataset. Kevin DeSimone, Kevin Weiner, Keith Schneider and other members of the Kastner lab have been unfailingly generous with their time in helping me understand its intricacies.

Meetings with Jim Haxby regularly helped clarify my thinking, and motivated many of the issues discussed here.

Many other members of the Norman, Cohen and Haxby labs have offered support and assistance in a variety of ways. I'm especially indebted to Ehren Newman, Sean Polyn and Chris Moore for some weapons-grade brainstorming, their levity and their keystrokes.

Ken Norman's countless insights, ever-present support, sage advice and merriment have brought me this far.

And finally, I wouldn't like to imagine how any of this would have felt without Sara Szczepanski's happy solace.

2 Abstract

In this paper, we describe a novel method for evaluating groups of voxels at a time based on their 'similarity structure'. This algorithm can be used as a feature selection algorithm for MVPA-style classification (Norman et al., 2006). We demonstrate improved performance relative to an ANOVA, and a further benefit for the multi-voxel version, using a delayed saccade visual attention and working memory dataset (DeSimone et al., submitted).

We also describe a novel analysis of a free recall paradigm based on Sahakyan and Kelley (2002), designed to make use of this similarity structure algorithm in localizing the 'context vector' hypothesized in the 'temporal context model' (Howard and Kahana, 2002). Finally, we describe progress so far on attempts to predict the degree of behavioral forgetting as a function of context change

3 Introduction

Functional MRI's three-dimensional, millimeter-resolution images provide tens of thousands of windows to the soul. This is too many, and they are too grimy for us to meaningfully look through with the human eye. Only by systematically, algorithmically winnowing the data down to a manageable form can we make sense of them. A plethora of voxel-by-voxel tests, dimensionality reduction techniques and clustering tools exist to help us do just that. Building on these methods, this paper emphasizes the value of searching for *isomorphisms* between patterns of activity in the BOLD response and the fine-grained predictions of rich psychological models.

4 The similarity structure algorithm

There is good reason to think that the brain processes information by representing and re-representing it. The multiple retinotopic maps (Engel et al., 1997), multiple tonotopic maps (Engelien et al., 2002), and multiple somatosensory and pain maps (Mazzola et al., 2005) all exemplify chains of processing, each step further emphasizing and de-emphasizing aspects of the world. For instance, we might talk about the the similarity structure of representations in area MT (Tootell et al., 1995) collapsing color distinctions and emphasizing depth and motion distinctions, or vice versa in V4 (Wandell, 2000). MT 'cares about' motion, and V4 'cares about' color - what counts as similar in MT may differ from what counts as similar in V4.

How can we formalize what counts as similar for a group of voxels in the brain? Or, to use the terminology we will adopt: for a given *voxelset*[1] of n

[1] Importantly, the voxels in a voxelset might have been drawn from a contiguous cluster, but they might equally have been drawn from locations dispersed throughout the brain. In the fMRI literature, the set of voxels to which an analysis is confined is usually referred to as a 'region of interest' or 'volume of interest'. Both these terms at least imply that the regions so defined are contiguous. Instead, we will adopt the term 'voxelset' to refer to a single voxel

voxels drawn individually from the functional volume, how can we quantify how similar the neural activity pattern at time t_i is to the neural activity pattern at time t_j?

Geometrically, a neural activity pattern can be thought of as a point in a high-dimensional space. Each voxel is a dimension in this space. A sequence of brain states changing through time traces a trajectory through this space. Similar patterns of activity can be visualized as points close together, and distinct patterns of activity are distant from one another. Activity patterns might be close together on some dimensions and differ on others.

A natural next step would be to formalize the notion of similarity between the pair of activity patterns at timepoints (i.e. 'scans' or 'TRs') t_i and t_j in terms of the Euclidean distance between two points in an n-dimensional space:

$$d_{i,j} = \sqrt{\sum_{n=1} (v_{n,i} - v_{n,j})^2} \tag{1}$$

where:

$d_{i,j}$ is the Euclidean distance between the activity patterns at timepoints t_i and t_2

$v_{n,i}$ and $v_{n,j}$ are the activity values for voxel n at timepoints i and j respectively

In the special case where the voxelset only includes a single voxel, the space is one-dimensional, and the Euclidean distance between a pair of timepoints boils down to a simple subtraction:

$$d_{i,j} = |v_{n,i} - v_{n,j}| \tag{2}$$

Of course, distance and similarity have an inverse relationship, and so similar points will have a small Euclidean distance and distinct points will have a high Euclidean distance.

We can look at the activity patterns in a voxelset elicited by a series of events or cognitive states, and compute the Euclidean distances between them. This complete set of pairwise comparisons, every timepoint compared with every other timepoint, will form a symmetrical *(timepoints* x *timepoints)* matrix, where the diagonal comparisons of timepoints with themselves will always have a distance of zero. We will refer to this distance matrix, computed from the voxelset data, as the 'data distance matrix'. It describes which timepoints' activity patterns are similar to each other, and which are distinct - in other words, it formalizes what that voxelset 'cares about'.

or group of voxels that will be included in an analysis. These voxels might have been chosen based on anatomy, some previous functional statistic or with tarot cards, although this paper will mostly discuss functionally-defined voxelsets.

If we have a *(timepoints* x *conditions)* regressors design matrix, as would be needed to run a general linear model (Worsley and Friston, 1995) , then we can construct a 'model distance matrix' in the very same way, substituting 'conditions' for 'voxels':

$$d_{i,j} = \sqrt{\sum_{c=1} (x_{c,i} - x_{c,j})^2} \tag{3}$$

where:

$d_{i,j}$ is the distance between the design matrix timepoints at t_i and t_2

$x_{c,i}$ and $x_{c,j}$ are the regressor values for condition c at timepoints i and j respectively

Considering this geometrically, we are treating our design matrix as tracing a trajectory through some c-dimensional condition space, where each dimension is a condition. In calculating our model distance matrix, we are comparing every point in the condition space with every other point in the condition space. This model distance matrix formalizes the similarity structure that we would like to see reflected in a voxelset.

We will now have two pairwise distance matrices - one for the data, and one for the model. They will be the same size *(timepoints* x *timepoints)*, since they will each contain the pairwise distances between every timepoint. If we flatten these matrices into a single long line, then we can simply correlate the values in the two lines. This is akin to asking whether the similarity structure of the model is mirrored by the similarity structure of the data. Ideally, timepoints that are very different (high Euclidean distance) in our model should be very different in the data, and similar (low Euclidean distance) timepoints in the model should be similar in the data. This correlation score between the flattened data distance matrix and the flattened model distance matrix is our index of 'goodness' in a voxelset. Voxelsets with high correlations care about the same things as our model - that is, they reflect its similarity structure.

Specifying a distance matrix between every timepoint and every other timepoint scales quadratically with the number of timepoints, but only linearly with the number of voxels in the voxelset. For computational convenience then, it makes sense to average timepoints from the same condition together, cancelling out noise, and drastically reducing the number of computations.

The similarity structure algorithm describes how to quantify the isomorphism between what a voxelset cares about and the predictions of a theory or model. In essence, this defines a criterion, or 'objective function' by which we may determine whether a voxelset is 'good' or not. In this way, we can search through the brain, looking for 'good' voxelsets, that exhibit the similarity structure in our model. Or, we could compare multiple models from different theories,

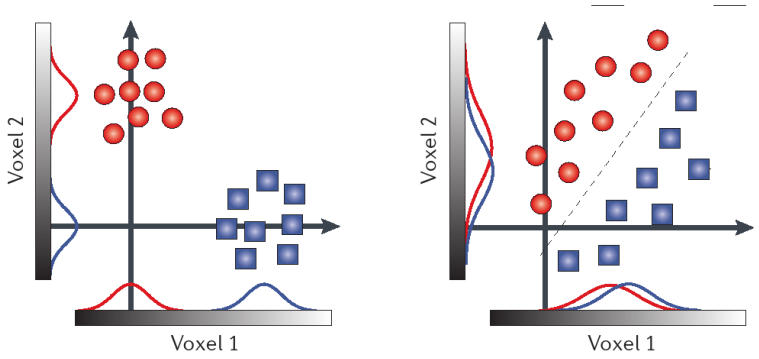

Figure 1: *Two-voxel toy examples to illustrate single- vs multi-voxel feature selection scenarios.* In the first case, the conditions are just as discriminable with a single voxel's values as when using information from both voxels at the same time. In the second case, the two conditions are not discriminable at all when viewing the values from voxel 1 in isolation, or when viewing the values from voxel 2 in isolation. Only when both voxels are considered at the same time can a decision boundary that splits the conditions be formed. From Haynes & Rees (2006).

and see which of them is best matched by the voxelset, as a means of evaluating theories about what that voxelset represents.

We can then conduct such a search through the brain treating each individual voxel as its own voxelset, applying the similarity structure algorithm to each in turn. This is the single-voxel, i.e. mass univariate, version of the algorithm, since it evaluates each voxel in isolation at a time. To create a binary mask indicating which voxels to include in further analyses, we can set a p-value or r threshold, or just pick the best n voxels. For simplicity, we will adopt this latter approach for all single-voxel feature selection from now on.

4.1 Multi-voxel similarity structure

4.2 Why might multivariate feature selection help? Considering the two-voxel case

Haynes and Rees (2006) discuss how discriminating between conditions in the most simple two-voxel case might be impossible if those voxels are considered in isolation. The most illustrative example can be seen in figure 1 - in this case, the boundary dividing categories A and B runs such that the values from v_1 are completely uninformative about category, as are the values from v_2 alone. However, in concert, the (v_1, v_2) vector is highly discriminative.

This toy example serves to illustrate how a multi-voxel voxelset might be treated differently from multiple single-voxel voxelsets. The relevance of this will become more apparent later when discussing classification.

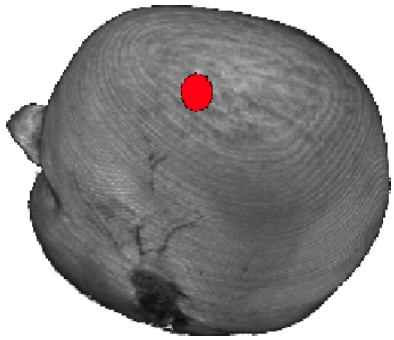

Figure 2: *Glass brain showing an example sphere of radius 4.* Some small amount of stretching may be visible, since the sphere-creating algorithm does not take inter-slice gaps into account when judging distances between voxels.

4.2.1 Searchlight spheres

Unfortunately, the combinatorics of testing multiple voxels at a time can quickly become intractable. After all, if we were to decide on a whim to evaluate every single 10-voxel combination in a 50,000-voxel brain, this would yield $\binom{10}{50000}$ combinations (about 10^{40}) to evaluate.

As a simple and reasonable alternative, Kriegeskorte et al. (2006) proposed evaluating a sphere's worth of voxels at a time, with each sphere centered on a different voxel. The radius of the sphere could be large or small (see discussion below). For a fixed radius of, say, 2, then each voxel would lie at the center of its own sphere containing roughly 30 voxels, with voxels at the edge of the brain having truncated spheres. To visualize this, see figure 2 for a randomly-chosen example sphere of radius 4).

Obviously, this approach is making assumptions about locality, since only geographically-proximal voxels are being treated together - there is no scope for evaluating voxels scattered around the brain at the same time. However, this can be seen as a boon too, since it dramatically restricts the number of possible subsets of voxels to consider.

The final implementational detail to consider relates to how the searchlight spheres should be scored in order to create a binary voxelset mask. Running the multi-voxel similarity structure algorithm produces a performance score (the correlation between data and model distance matrices) for each searchlight sphere. Three options seem worthy of consideration:

> a) *Center-voxels only*: each voxel is assigned the score for the sphere of which it is the center. To define a voxelset mask, choose the best n voxels.
>
> b) *Entire spheres*: to define a voxelset mask, choose the best m spheres, and include all of their constituent voxels. Though spheres may overlap, each voxel is included in the voxelset only once.

c) *Average participation*: each voxel participates in multiple spheres. The score for a voxel is the average of all the spheres' scores in which it participates. To define a voxelset mask, choose the best n voxels.

In this paper, we consider only the first 'center-voxels only' option, though the others would be interesting to try in the future.

4.3 Making use of the voxelsets

Evaluating voxelsets can be an end in and of itself. Localizing the areas that reflect the similarity structure of a model on a brain map can be very informative. Indeed, sometimes just demonstrating that there is a region in the brain that reflects some similarity structure, irrespective of where it is, can provide evidence for a theory that posited the existence of such a process or representation.

Alternatively it might be possible to use similarity structure as a 'feature selection' algorithm to choose the best voxelsets on which to run some kind of further classification or regression analysis. Following Haxby et al. (2001), a growing number of neuroimaging researchers have argued that multi-voxel methods such as classification may provide a more sensitive measure of the informational content of a voxelset than the mass univariate approach (Mitchell et al., 2004; Kamitani and Tong, 2005; Haynes and Rees, 2005; Norman et al., 2006). Voxels that might not be significant when tested in isolation with a mass univariate contrast might still provide information that improves classification performance when part of a larger aggregate. For instance, Kamitani and Tong (2005) demonstrated that the line orientation of a grating is classifiable with a linear support vector machine using only voxels from V1. This is particularly surprising because the microscopic organization of orientation-selective neurons in V1 cortical columns is not visible macroscopically with a mass univariate GLM. They argue that this is possible because tiny orientation-selective biases in individual voxels become highly informative when aggregated over many such voxels.

Many of these rely on a leave-one-out cross-validation framework, where the feature selection and classifier algorithms are trained on all but one run of the data, and then tested on the remaining run. Each run gets a turn at being the withheld testing run on one of the iterations in this cross-validation loop. Though time-consuming, this approach makes good use of the limited data.

Seen in this way, we could use classification performance as a means of comparing feature selection algorithms. This will be our aim in the next section, where we benchmark multiple feature selection algorithms on a delayed saccade visual attention and working memory dataset.

Better still, we could train a classifier or regression algorithm to make behavioral predictions, based on subject's behavior in one phase of the experiment, as per Polyn et al. (2005). This will be the aim of the final section, where we define our voxelsets using the similarity structure algorithm, and then attempt to make predictions about behavior using these voxelsets.

4.4 The delayed saccade dataset

4.4.1 Methods and analysis path

The primary dataset I will use for benchmarking the similarity structure algorithm was collected by DeSimone et al. (submitted) to examine topographic maps in frontal and parietal areas using a memory-guided delayed saccade design (Sereno et al., 1995). The task required subjects to saccade to one of twelve peripheral locations arranged clockwise around a central fixation point. Each trial, lasting a total of 5s, consisted of a brief period fixating on the central cross while a target appeared in one of the twelve clockface positions at approximately 10° eccentricity for 500ms. Subjects had to remember this location while multiple sets of distractors were presented for 3s. The disappearance of the fixation point cued subjects to saccade to the remembered target location and back to fixation, in time for the next trial's fixation cross to appear in 1500ms.

Each run began with the 3 o'clock position, and trials proceeded sequentially anti-clockwise around the clock. A complete cycle around the clock took 60s, with 8 cycles in a scanning run. The subject was allowed to rest after each of the 6 runs, but there were no rest trials during scanning runs. The position of each target was randomly jittered by up to 2.5° in each direction.

Three subjects participated in the study. Data were acquired with a 3T Siemens Allegra head-dedicated MRI scanner using a standard birdcage coil, using a gradient echo, echo planar sequence with a 128 square matrix, in-plane resolution of 2x2mm^2, 20 axial slices each 2mm thick with a 1mm gap between slices, and a repetition time of 2s. The acquisition volume was positioned to cover frontal, parietal and dorsal occipital cortex. A high-resolution anatomical scan was taken at the end of the session with an MPRAGE sequence of $1mm^3$ resolution and a 256x256 matrix. An in-plane magnetic field map image was acquired to perform echo planar imaging undistortion.

The functional images were motion-corrected (Cox and Jesmanowicz, 1999) to the image acquired closest in time to the anatomical scan and undistorted using the field map scan. Linear and quadratic trends within runs were subtracted, and each voxel was z-scored within runs to give it a mean of 0 and standard deviation of 1, following Polyn et al. (2005).

The simplest way to have characterized the design of this experiment would have been to assign each timepoint a discrete label from 1–12, referring to the position on the clockface being saccaded to in that trial. Unfortunately, because each timepoint's repetition time (TR) was 2s and trials lasted 5s, a trial consisted of 2.5 timepoints. Sophisticated ways to take these 'halfway' timepoints into account were considered, but for reasons of expediency, they were instead simply ignored.

In order to account for the haemodynamic lag, the labels in each run were shifted forward by 3 timepoints, such that each data timepoint now corresponded to the label from 3 timepoints (6s) previously, corresponding roughly with the peak of most standard haemodynamic response functions, following Polyn et al. (2005). For reasons that will be discussed further below, it was

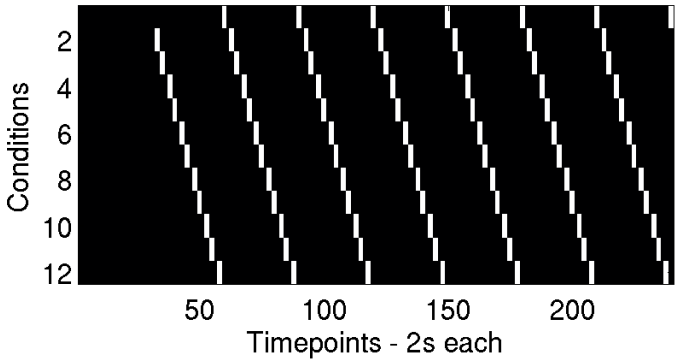

Figure 3: *Condition labels for every timepoint in a run*, after shifting 3 timepoints along, removal of halfway timepoints, and removal the first cycle.

felt that more sophisticated haemodynamic response function convolution techniques would complicate things without any real benefit.

By shunting the labels along like this, the first few timepoints in each run no longer had a label, and the last few 2 o'clock timepoints in each run were truncated. This meant that there were fewer timepoints from the 2 o'clock condition than from the others. Following DeSimone et al. (submitted) , we re-balanced the conditions by excluding almost all of the first cycle of each run, such that the first label in a run was a 2, and the last label in the run was a 1. Figure 3 makes the end result clear for a single run.

Finally, all of the timepoints from each given condition from a given run were averaged together. This is not strictly necessary, but cancelled out a great deal of the noise, making the results much cleaner. This left 1 averaged-timepoint per condition per run, making 72 in total.

This labels matrix can be re-expressed in terms of the Cartesian coordinates of the clockface position being saccaded to, with the eccentricities normalized to lie on the unit circle. Thus 12 o'clock would have coordinates (0,1), and 9 o'clock would have coordinates (-1,0). This is the key step that will allow us to treat some conditions as more similar to each other than others.

Using tools provided by DeSimone et al. (submitted) and following Sereno et al. (1995), the data were first analyzed using a Fourier decomposition, applied to each voxel in isolation. A 'good' voxel should show a roughly sinusoidal oscillation at the frequency of the clock cycles (1/60Hz). The greater the power of that frequency, the better the voxel.

The Fourier decomposition rests on the well-founded assumption that each voxel has a Gaussian-like tuning curve. That is, a '6 o'clock' voxel with a Gaussian tuning curve would respond maximally when the subject was saccading to the 6 o'clock position, respond somewhat when saccading to 5 or 7 o'clock, only a little when saccading to 4 or 8 o'clock, and so on (see Figure 4). This is, in effect, the 'model' underlying the Fourier analysis - it looks for voxels with

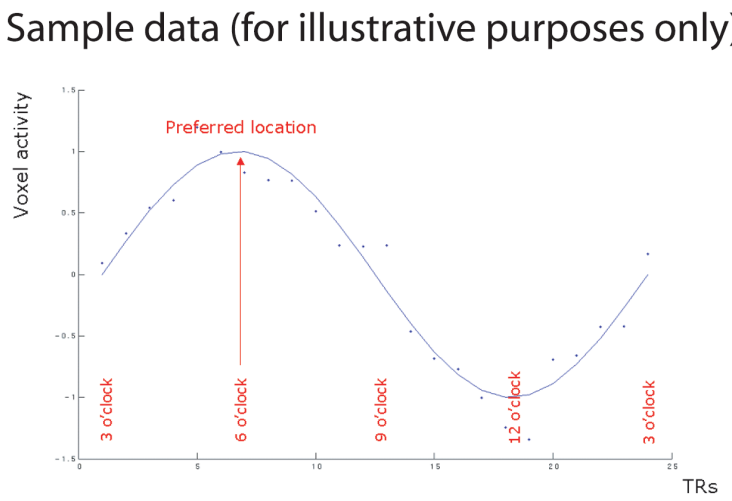

Figure 4: *Depiction of what a perfect voxel with a Gaussian tuning curve centered on the 6 o'clock clockface position might look like.*

high power at the frequency of the clock cycles in the task, since voxels that show Gaussian tuning curves for saccade angle will look roughly sinusoidal after the cyclic sequence of Gaussians have been convolved with a haemodynamic response function (Cox, unpublished).

In order to establish that the similarity structure algorithm is indeed producing sensible results, brain maps produced by the Fourier and single-voxel similarity structure algorithms were compared. In order to do this, it was necessary to construct a model distance matrix from the Cartesian coordinates design matrix already described. Thus, for the model, the distance between timepoint t_i and timepoint t_j is just the Euclidean distance between the clock face position being saccaded to at t_i and the clock face position being saccaded to at t_j.

We now need a corresponding distance matrix for our actual data, which we will compare with our model distance matrix. To calculate each distance, the voxelset's activity patterns at timepoint t_1 are compared with the set of values at t_j.

Another way of putting this, most usefully for our purposes, is to consider that a good voxelset's response will be similar for nearby positions on the clockface, and different for distant/opposite positions on the clockface. Its activity patterns will be similar for 12 and 1 o'clock, but very different for 3 and 9 o'clock, as can be seen in Figure 5 .

As a means of visualizing the similarity structure of the voxelsets in a more intuitable way than the numbers in the data distance matrices, multidimensional scaling (Shepard, 1980) was used to create two-dimensional plots depicting which conditions are similar to which (Edelman et al., 1998; O'Toole

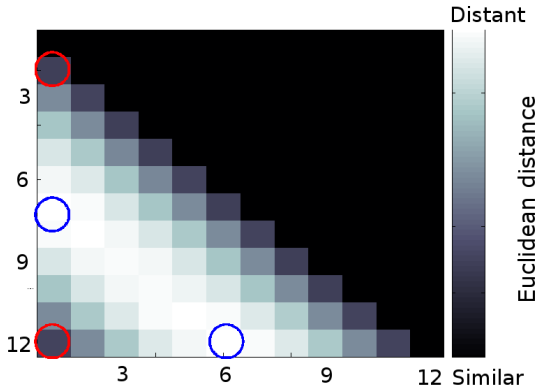

Figure 5: *Model distance matrix created from the Cartesian coordinates design matrix.* The red circles highlight examples of very similar pairs of conditions - 1 o'clock & 12 o'clock, and 1 o'clock & 2 o'clock. The blue circles highlight examples of very distant conditions - 6 o'clock & 12 o'clock, and 1 o'clock & 7 o'clock.

et al., 2005). Although these previous efforts have visualized the reduced dimensionality space, we are not aware of any efforts to use the shape of this space as an objective function for evaluating voxelsets, as described here.

To corroborate the findings that each half of the visual field is represented contralaterally, two further model distance matrices were constructed. One only computed pairwise distances between the conditions from the left half of the visual field (7, 8, 9, 10 and 11 o'clock), while the other only computed pairwise distances between the conditions on the right half of the visual field (1, 2, 3, 4 and 5 o'clock). It was hoped that these two models would pick out contralateral areas as best reflecting their respective similarity structures.

Finally, in order to quantify the relative efficacy of different feature selection methods, a full-scale cross-validation feature selection and classification analysis was run. The classifier was trained on 5 of the 6 runs to discriminate between the activation patterns from the twelve different clockface position conditions. A classifier's guess was considered to be correct if the (x,y) coordinate of its output was closer to the target clockface position's location than any of the other 11 positions. The Matlab (Natick, MA) Neural Networks Toolbox 'trainscg' backpropagation algorithm was used without a hidden layer to predict real-valued x and y coordinates.

Since Haxby et al. (2001) simply used a mass univariate omnibus ANOVA to test whether each voxel's activity varied significantly between conditions, we included this algorithm as a baseline performance estimate. We expected classification performance using voxelsets selected with this algorithm to be worst since it did not take into account any of the information about which conditions were more or less similar to each other, but simply treated all the conditions

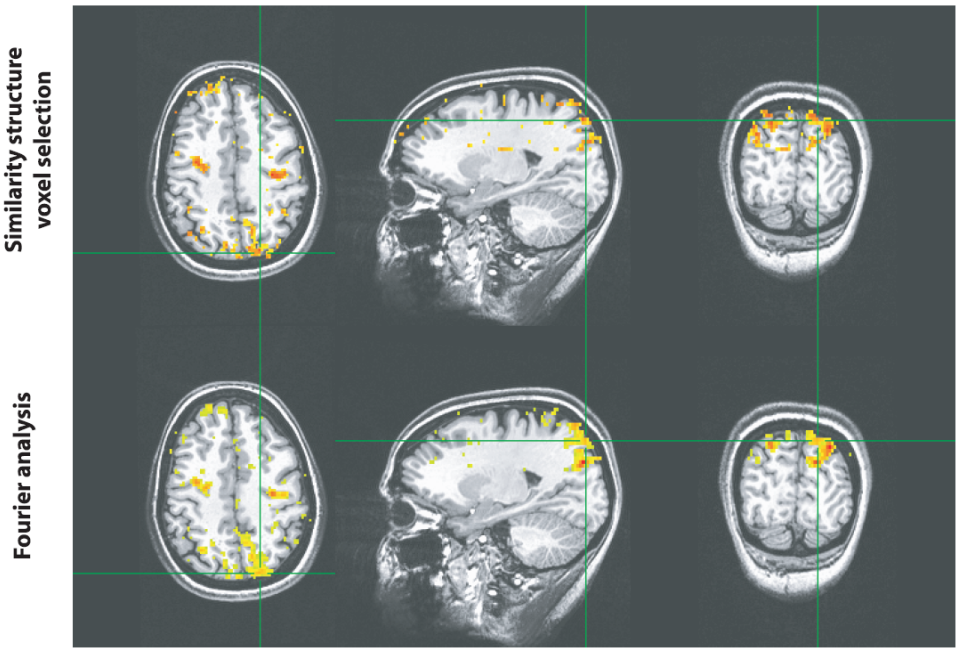

Figure 6: *Comparison of similarity structure and Fourier feature selection algorithms on subject 1.* Thresholding on the two maps has been arbitrarily chosen to facilitate comparison. Intraparietal sulcus, frontal eye fields and dorsolateral prefrontal cortex regions are picked out by both algorithms.

as distinct groups. For this to work, each timepoint has to be considered as 'belonging' to one condition, and so the 12 binary boxcar condition labels were used (see Figure 3).

4.4.2 Results and discussion

Previous results, (DeSimone et al., submitted) have shown a clear lateralization of representation, with each half of the visual field being represented by the contralateral hemisphere in three main areas we can broadly term 'intraparietal sulcus' (IPS), 'frontal eye fields' (FEF), and 'dorsolateral prefrontal cortex' (DLPFC).

In order to determine whether the similarity structure algorithm was returning results in line with conventional analyses, we produced thresholded brain maps using the the similarity structure (see figure 6 top pane) and standard Fourier analysis (figure 6 bottom pane) algorithms. Little effort was made to threshold the two maps in exactly the same way, but it the two appear to be picking out similar results.

The IPS and FEF regions are clearly visible in Figure 6 (bottom pane), though the head was tilted during scanning so the bilateral symmetry is not

visible in a single slice. The dorsolateral prefrontal regions that DeSimone et al. (submitted) found were also delineated to some degree using the similarity structure algorithm (also not visible in this slice).

We also ran a single subject in a modified design where the condition order was completely randomized, in order to facilitate certain GLM contrast analyses that could not be run on the delayed saccade dataset because of the collinearity of the regressors. Although we do not report the results here, the single-voxel GLM and similarity structure algorithms produced similar brain maps.

The low-dimensionality representation of the neural activation patterns in figure 7 is clearly isomorphic to the original clockface condition-structure used to select the voxels. To be interesting, such plots should be generated from the withheld runs in a cross-validation framework, though this one was not. This is a proof of concept but the approach could be useful in the future, perhaps as a means of visualizing the similarity structure for more complex distance matrices.

This mass univariate omnibus ANOVA provided a baseline against which the similarity structure algorithm could be evaluated, since the ANOVA did not not take into account any prior knowledge about the similarity between conditions. It was, in effect, just a simple GLM with 12 boxcar regressors. Figure 9 shows that the ANOVA drastically under-performs the similarity structure algorithm when picking voxels. There are improvements to the way this ANOVA analysis was run that might perhaps bring up its performance, such as adding in the scanning runs as a random effect to remove a possible source of variance. However, when all other things are kept equal, adding in continuous-valued information describing the similarity structure between conditions helps, relative to not using it at all.

Just as different regressor matrices in a GLM will give rise to different beta weights, creating different model distance matrices will provide different data/model correlations. The simplest way to illustrate how this might be used is to test alternative similarity structure models. For instance, DeSimone et al. (submitted) showed hemispheric lateralization of visual field representations, especially in lower extrastriate areas. In other words, the right hemisphere cares about the left half of the visual field, while the left hemisphere cares about the right half of the visual field. That is, the similarity structure of the left clockface positions should be preserved more carefully in voxels from the right hemisphere, and vice versa.

Indeed, as expected, the model distance matrix of the left visual field alone activated primarily the contralateral hemisphere (see figure 8 left pane), and vice versa for the right visual field. Since this is a radiological view of the brain, the right hemisphere is actually pictured on the left hand side, and clearly shows greater correlations in the intraparietal sulcus than the right hemisphere. The opposite holds more or less true for the right half of the visual field. This lateralization is less clearly visible in the more frontal FEF and DLPFC areas, however. We will return to this idea of model-testing later, when analyzing the context dataset.

The single- and multi-voxel versions of the similarity structure algorithm

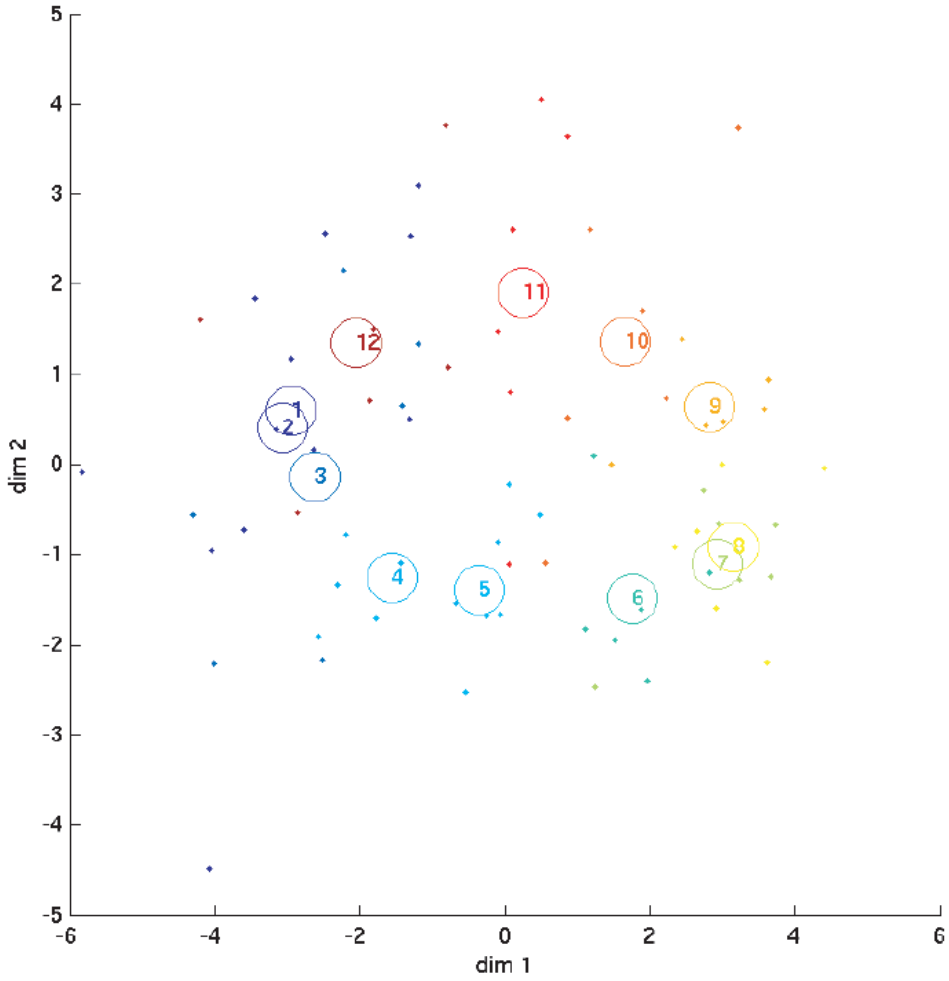

Figure 7: *The first two dimensions of a multidimensional scaling analysis run on the averaged neural activity patterns from the best 200 voxels chosen by the similarity structure algorithm.*

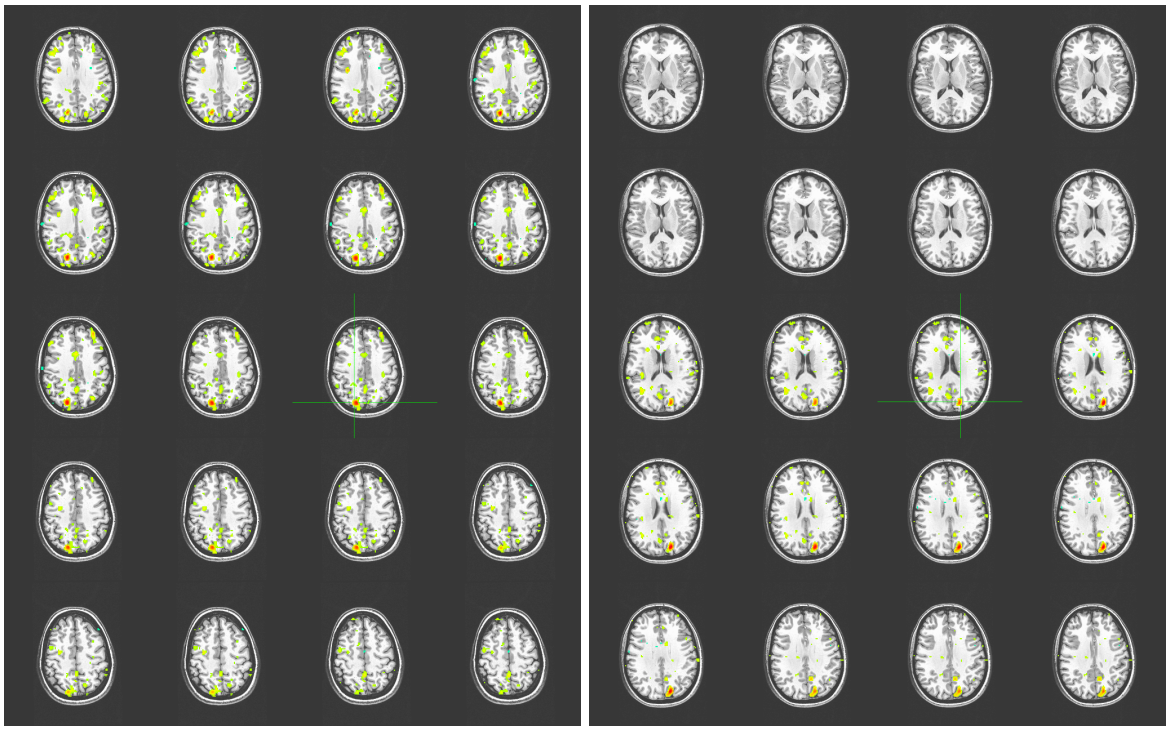

Figure 8: *Thresholded brain maps produced using the similarity structure for the left visual field conditions only (left side) and right visual field conditions only (right side).* Following radiological conventions, left=right. As predicted, the left visual field similarity structure picks out primarily the right hemisphere (confusingly shown on the left), and and the right visual field similarity structure picks out primarily the left hemisphere (on the right). Note: the two montages are plotted using slightly different coordinates, because the subject's head is tilted, making it difficult to see the effect clearly with the same set of coordinates.

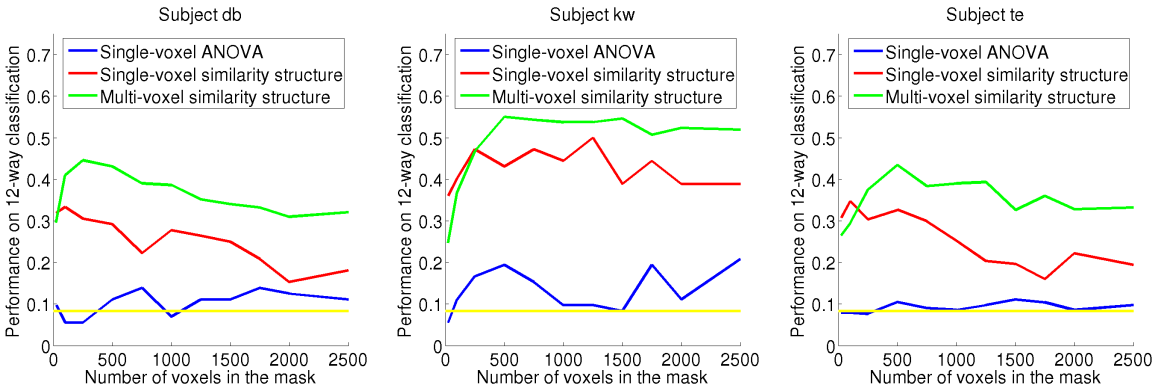

Figure 9: *Comparison of: single-voxel ANOVA (blue); single-voxel similarity structure (red); multi-voxel similarity structure (green)* - three subjects. Chance performance (yellow) is 8%. Unsmoothed data, averaged timepoints (one per condition per run).

were compared directly on the delayed saccade dataset (see Figure 9). 'Performance' here refers to 12-way classification performance, using the cross-validation generalization framework described above. Figure 9 shows a healthy improvement when the multi-voxel version of the similarity structure algorithm is used over the single-voxel version, for all three subjects. All three algorithms perform above chance (the flat yellow line around 8%), though the ANOVA is clearly dramatically less effective than the similarity structure algorithm.

4.5 Comparison with spatial averaging or smoothing

There are many possible explanations of the finding that the multi-voxel algorithm performed better than the single-voxel algorithm. It could be that the multi-voxel algorithm is working better simply because it is averaging over multiple voxels, cancelling out uncorrelated noise. This would be an unsophisticated win for multivariate methods, and could be more easily achieved simply by averaging all the values within a sphere together, or smoothing with a Gaussian kernel. Kriegeskorte et al. (2006) set out to test exactly this, and showed benefits to using multi-voxel feature selection on unsmoothed data over single-voxel feature selection with smoothed data, although their results were not unequivocal.

In order to understand this better, we also experimented with synthetic data (Kriegeskorte et al., 2006). Our aim was to show that there are cases where simply smoothing or spatially averaging does not choose voxels as well as taking multiple voxels into account would. Of course, one can also create synthetic data where there is no benefit to taking multiple voxels into account beyond averaging, and we compared these two cases. Since this work with synthetic data is only tangentially relevant to the use of richer psychological

models in fMRI, we do not discuss it further here.

The fairest way to test whether the multi-voxel algorithm is benefiting solely from averaging/smoothing would be to create masks using:

1. the single-voxel feature selection algorithm on smoothed or spatially-averaged data

2. the multi-voxel feature selection algorithm on unsmoothed data

and then compare their cross-validation classification performance on unsmoothed data. We are in the process of running this analysis.

However, even these results will need careful consideration. The delayed saccade task elicits activation in large, topographically-organized areas, where most of the signal appears to exist at a very low spatial frequency. However, there may be other datasets or other brain areas where the information is represented at a higher spatial frequency. It may be that averaging will help with the delayed saccade data, but may impair performance on other datasets. As a result, we are also working on benchmarking other datasets, such as from the Experience-Based Cognition (EBC) competition (Schneider et al., 2006).

4.6 Comparison with standard GLM

We have described how a model distance matrix for the similarity structure algorithm could be produced by coding each condition as an (x,y) coordinate on the unit circle, and computing their pairwise distances to each other.

More or less the same analysis could be conducted using a standard mass univariate general linear model (Worsley and Friston, 1995), using a design matrix consisting of just two regressors, the x and y coordinates. The GLM predicts each voxel's values as a weighted sum of the regressors in the design matrix (ignoring for now regressors of no interest such as head motion, linear or quadratic trends and other artifacts). The amount of variance accounted for in predicting each voxel's timecourse from the x and y coordinates would be the goodness value used in selecting voxels.

$$y' = \mathbf{X}\beta + \epsilon \tag{4}$$

where:

y' is the single-voxel timecourse being predicted

$\mathbf{X}$ is the *(timepoints* x *conditions)* design matrix

β is the vector of beta weights

ϵ is the residuals vector of variance unaccounted for by the GLM

Though the results are not reported here, this has been found to yield a very similar brain map to that of the *single*-voxel similarity structure algorithm, since they are based on the same psychological model, and both deal with single voxels and multiple conditions at a time.

The primary advantage of the similarity structure algorithm is that it can be naturally extended to take multiple voxels into account, rather than just a single voxel at a time. In other words, the similarity structure algorithm takes into account multiple voxels *and* multiple conditions.

4.7 Future directions

The results shown in Figure 9 are a promising start, but it may be possible to bolster these results in the future. We make two suggestions, both of which allow for the possibility of evaluating non-local and non-spherical voxelsets. Both suggestions are agnostic about the particular multi-voxel objective function used.

A simple but promising alternative to the searchlight sphere approach has been proposed by Bryan and Haxby (2006). They start with the best voxel, as evaluated by some single-voxel algorithm. They then run a multi-voxel feature selection algorithm on every pairwise combination of that voxel with every other voxel. In so doing, they determine the best voxel to add to create a two-voxel voxelset. They then exhaustively determine the best voxel to add to create a three-voxel voxelset, and so on, until adding a voxel causes performance to drop. This method allows the voxels to be distributed in completely arbitrary fashion. This should at least match or exceed the performance of a single-voxel approach, and indeed, Bryan and Haxby (2006) have demonstrated tangible benefits over single-voxel feature selection.

A second future direction would be to consider feature selection as a search problem. The stepwise approach just described is, in effect, a simple search algorithm, though more sophisticated techniques exist, such as genetic algorithms (Mitchell, 1996) or beam searches (Russell and Norvig, 2002). Indeed, one could start with a mask defined already, either using a single-voxel approach, the searchlight spheres, or the stepwise approach, and then try adding and removing single voxels, or combining multiple masks together in a systematic fashion to dramatically expand the range of voxel subsets being considered. Of course, this comes at a combinatorial cost, and it lacks the simplifying locality constraint of the searchlight spheres.

Much of the preceding discussion about feature selection algorithms was motivated by a desire for new methods that incorporate rich psychological models when evaluating voxelsets. We can now apply the similarity structure now to a new domain, where a rich theory exists that has yet to be addressed with neuroimaging.

5 Temporal context model analyses

The final set of analyses center around a memory-related dataset. This paradigm builds on the rich episodic memory literature, concerned with the storage and retrieval of memories of events occurring at a specific place and time. We create an objective function based on the predictions of a prominent model of episodic memory, and attempt to use the voxelsets so defined to track changes in subjects' internal context as a predictor of forgetting.

5.1 Background

5.1.1 The temporal context model

Howard and Kahana (2002)'s 'temporal context model' (TCM) of episodic memory provides the underlying framework for the experiment. Their theory builds on an old idea, that the context we are in affects the way we remember things (Estes, 1955; Godden and Baddeley, 1975; Mensink and Raaijmakers, 1989). 'Context' is used here in a broad technical sense as an amalgam of the subject's external physical environment as well as a plethora of internal mental and physiological variables, such as one's mood, current activity, or background thoughts.

TCM considers there to be two components to an episodic memory:

a) The primary episodic *content* of the memory, which might comprise any kind of occurrence specific in place and time

b) A snapshot of the *context* state at the moment of encoding

Following a standard practice in the memory literature, we can treat any neural representation as a vector - so the episodic content of a memory is a vector, and the context snapshot is a vector too. Simplifying considerably, we can think of the encoding of an episodic memory as storing the combined content and context-snapshot vectors. Correspondingly, we can think of retrieval as the business of returning the stored vectors that somehow match a partial or noisy cue vector.

TCM considers the context vector to change from moment to moment. Moving to a new environment produces a large and abrupt change. A late-morning hunger that burgeons over the course of an experiment drives the context to change more gradually. The many, many internal and external variables that make up our mental state could each be conceptualized as dimensions of the context vector. Context, viewed in this way, has a somewhat Heraclitan feel to it.

TCM makes one further, vital supposition - the retrieval of an episodic memory actually activates, or 'reinstates', its context vector payload, making that past context current again. By rewinding our brains a little to the way they were when we first encoded the memory, we make it easier to retrieve other memories that contain similar context vectors. Concretely, this means that retrieving a

memory makes it easier to recall other memories encoded at around the same time, in the same place, or under the same circumstances. This 'lag recency' effect has been clearly demonstrated in free recall (Kahana, 1996). Having recalled the item from timepoint t_i, subjects are much more likely to next recall the item from timepoint t_{i+1} or possibly t_{i-1}.

5.1.2 Context and forgetting

Sahakyan and Kelley (2002) ran an elegant between-subjects behavioral study using a simple manipulation. Subjects were presented with a list of words, then performed a task, then saw more words, and then freely recalled all of the words they'd seen. The task was manipulated in one of two ways:

1. during the *invisibility task*, subjects had to imagine what they would do if they were invisible for a day, without considering the implications of their actions

2. during the *waiting task*, subjects simply had to wait for a period of the same duration

The invisibility condition was deliberately designed to disrupt the subjects' mental context. Intuitively, the idea is that the imagining of salacious, murderous, voyeuristic, mischievous or other vividly evocative acts would banish all thoughts of the prior list of words. The rate of change of mental context would be rapid during this invisibility task, especially relative to the more pedestrian foot-tapping and thumb-twiddling going on during the 'waiting' control task.

Sahakyan and Kelley (2002) showed clearly that subjects who performed the invisibility task in between lists remembered the second list better and the first list worse than subjects who had simply waited between lists. This finding fit neatly within the Howard and Kahana (2002) temporal context model framework. After all, if every new episodic memory is tagged with the context vector present at the moment of encoding, then two episodic memories laid down a minute or so apart will share similar context vectors, and the retrieval of one should facilitate retrieval of the other ... *unless* something had occurred in that short interval that produced an abrupt change in context. In this case, the two memories' context vectors would bear considerably less similarity to each other, and retrieval of one would hardly facilitate retrieval of the other.

In the case of the Sahakyan and Kelley (2002) experiment, the context when free recall began would be most similar to the context present at the end of the second list, moments before. Thus, retrieval of the second list should be good. Retrieval of the first list would be substantially affected by the interval between the two lists - if we imagine that the invisibility task caused a considerable change in context, then retrieval of the second list context will barely facilitate retrieval of the first list. This would explain why performing the invisibility task causes impaired performance on the first list.

Sahakyan and Kelley (2002) went one step further, in a second experiment. They showed that simply by asking subjects to think back to the goings-on at

the beginning of the experiment, they were able to neutralize the forgetting effect resulting from performance of the invisibility task. That is, deliberate reinstatement of the context from the first list facilitated recall of that list, allowing subjects to jump across the cross-list context chasm created by the invisibility task. This second experiment lent considerable support to the TCM model, though it is the first experiment that we sought to replicate.

5.2 The temporal context dataset

In the following analyses, we will aim to define a voxelset whose rate of change between lists is greater when subjects perform tasks that should greatly disrupt their context than when they perform tasks that should barely disrupt their context. Moreover, we hope that this rate of change will be directly predictive of the degree of behavioral forgetting of the first list.

5.2.1 Methods

Following a series of behavioral pilot studies, we alighted on a modified version of Sahakyan & Kelley's design that could be run within-subject, in part because of the greater difficulty and expense involved in scanning subjects with fMRI, but also because good cross-subject MVPA methods have yet to be developed.

In order to fit as many runs into a single scanning session as possible, items were presented faster, the task durations and free recall periods were shorter and the entire design was quickened wherever possible, providing 8 high-disruption and 8 low-disruption runs per subject. Only three subjects' data have been collected so far using this pilot design. Each run consists of a 15-second initial blank period while the scanner settles, 8 seconds of simple arithmetic tasks, 8 concrete, imageable words, each presented for 2 seconds, 40 seconds of task, 8 more words, 8 seconds of arithmetic, and then a 40-second free recall period. At the end of 16 runs, a final 3-minute free recall run for the entire experiment was included, though it won't be analyzed here.

For the low-disruption task, subjects fixated on a central fixation cross, and were asked to count the number of times it changed luminance. Subjects were not asked to report this number. The high disruption task involved short imagination tasks, such as imagining being invisible for a day.

A second set of modifications related to a possible confound in the original Sahakyan & Kelley design. In their design, each item was presented for 5 seconds, with no encoding task being performed, so it seems at least possible that subjects were rehearsing items they had seen so far as they went along. Moreover, it seems likely that subjects in the waiting condition spent some of that time rehearsing the first list. This alone could explain their improved performance on that list. It is, of course, also possible that subjects in the invisibility condition were rehearsing the first list too, since they were not required to talk out loud about their imaginings. Our concern was that greater rehearsal in the waiting condition could account for improved performance on list 1 in the waiting condition, rather than changes in temporal context. This is why

our low disruption task still required subjects to pay attention to the changing luminance of the fixation cross, a simple but engaging task that should make it harder to rehearse. We also reduced the item presentation time to just 2 seconds, and asked subjects to picture each item in isolation to minimize elaborative encoding of multiple items at a time, to minimize rehearsal during the lists.

Our aim was to see whether we could look at the timepoints from just before the task period, and from just after, and measure the change in brain state between them. We hoped that the magnitude of this neural change, measured as the Euclidean distance between the 'before' and 'after' activity patterns, would predict the degree of disruption from the task, and therefore, the degree of forgetting shown behaviorally.

The original block design matrix was convolved with a haemodynamic response function (Cox, unpublished) (which effectively shifts each of the condition labels three timepoints to the right), then renormalized so that the values ranged between 0 and 1, and finally thresholded at the 0.8 level to binarize the condition values again.

None of the timepoints used to calculate the magnitude of neural change incorporated any of the timepoints in which the disruption task was being performed. Further, we were careful to ensure that none of the timepoints from directly after the task were included either, lest the sluggish haemodynamic response to the task leak into the timepoints used for the post-task snapshot. To visualize which timepoints were included for the pre-task and post-task snapshots, see Figure 10c. The unconvolved task block (Figure 10a, middle row) finishes on timepoint 40, and the first timepoint to be used as part of the post-task snapshot (Figure 10c, bottom row) is timepoint 46. The three pre-task timepoints in Figure 10c were averaged together to create the pre-task snapshot, as were the three post-task snapshot timepoints. The scalar magnitude of neural change was then calculated as the Euclidean distance between the average of the three pre-task activity patterns and the average of the three post-task activity patterns.

The behavioral forgetting effect for each run was computed as per the following equation:

$$f_r = (n_{r,2} - n_{r,1}) \tag{5}$$

where:

f_r is the behavioral forgetting effect measure for run r, for a single subject. A higher f means that list 2 is being remembered better than list 1 in that run, as we hope to see for the high disruption runs.

r is the index of the run (numbered from 1–16)

$n_{r,1}$ is the number of items remembered in the r^{th} run from the first list

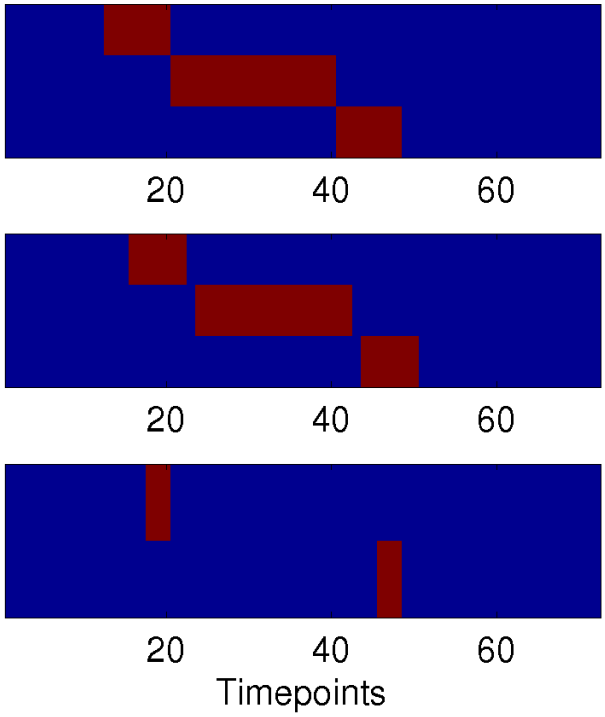

Figure 10: *Timepoints from list and task for a single run.* Plot (a) shows the original, unconvolved design matrix, with the list1, task and list2 timepoints in red. Plot (b) shows the design matrix after being convolved with a haemodynamic response function, renormalized to a 0-1 range, and thresholded at 0.8. Plot (c) shows just the timepoints for list1 and list2 that were averaged together to create the pre-task and post-task snapshots. Note: the other conditions that are not shown are blank periods, instructions, recall, arithmetic etc.

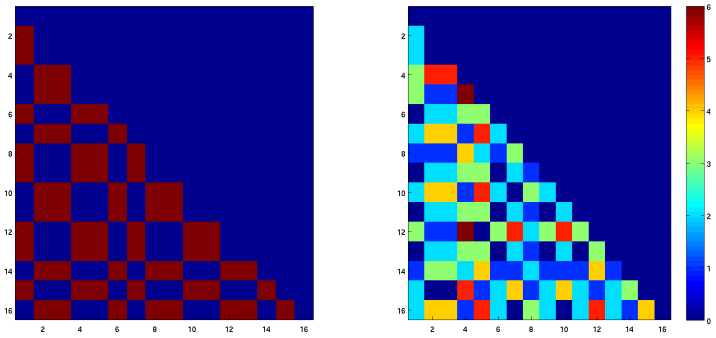

Figure 11: *a) Model distance matrix for low vs high disruption task b) Model distance matrix for behavioral forgetting effect.* Both figures are drawn from subject 1 and differ according to the counterbalancing of the task conditions across subjects. As always, higher values mean greater distances.

$n_{r,2}$ is the number of items remembered in the r^{th} run from the second list

Other ways of quantifying the forgetting effect could be considered in the future.

Following Sahakyan & Kelley, we had expected that there would be a greater behavioral forgetting effect in the runs where subjects performed the more disruptive task. We hypothesized that this forgetting might be caused by greater changes in context when performing the imaginative task, thus making the first-list items less readily retrieved using a second-list context cue. A proximal and less ambitious goal would be to find a voxelset whose magnitude change over the course of the task interval is predictive of which task is being performed. Our ultimate aim is to find a voxelset whose magnitude of change during the task interval is predictive of the behavioral forgetting effect.

We can redescribe our hypothesis in terms of a model distance matrix, which we can then use to evaluate voxelsets. In this case, we have two model distance matrices we can use - one for high vs low disruption task, and one based on the degree of behavioral forgetting, where each run has been collapsed down to a single high/low disruption or f_r behavioral forgetting value. The timecourses for the two models can be seen in figure 12, from which the corresponding model distance matrices will be produced (e.g. figure 11 for subject 1). Unfortunately, complicated distance matrices such as these are difficult to interpret visually.

5.2.2 Results

Though it is hard to see from Figure 12, subjects remembered more words from the second list than the first list when the task performed was more disruptive (see Figure 13). This is the behavioral effect that was sought, where the

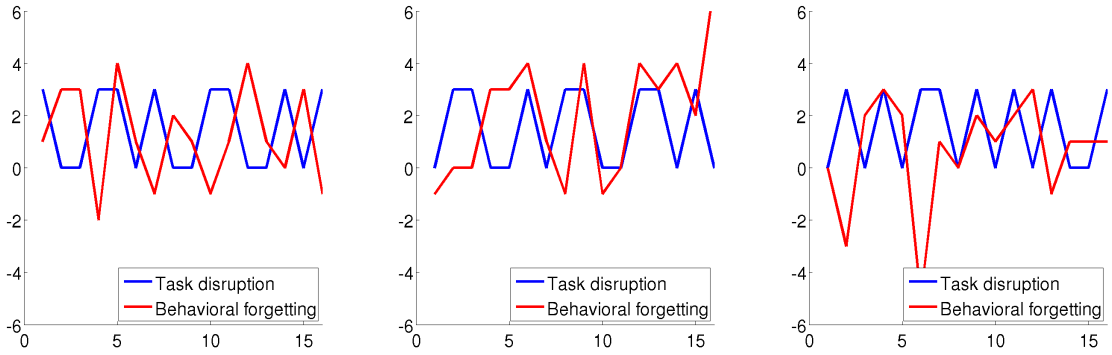

Figure 12: *Low vs high disruptiveness of task and degree of behavioral forgetting for each run, for all three subjects.*

more disruptive (imagination) task caused improved recall of the second list and worsened recall of the first list than the less disruptive (luminance cross) task. However, the within-subjects effect is not very strong.

Figure 13 shows performance at predicting whether a given run contained a high or low disruption task, and the behavioral forgetting effect for that run. It appears then from these preliminary analyses that we can predict which task was being performed quite well on two of the three subjects (see figure 13a), but we cannot yet predict the degree of behavioral forgetting (see figure 13b) reliably.

We have not yet thoroughly investigated where in the brain voxels are being drawn from.

5.2.3 Discussion

There are many possible explanations for the failure to predict behavior. Of course, it could be that the basic TCM conception of context, at least as operationalized here, needs re-working. At this stage, that would be a prematurely strong conclusion. A more likely showstopper is simply that fMRI technology is poorly-suited or inadequate for the analyses attempted. This could be because higher spatial resolution and an improved signal-to-noise ratio is necessary to distinguish tiny variations in the context vector over the course of a 40-second task. It could even be that the blood-based BOLD response is a poor proxy measure for stable but subtly-varying patterns of neural activity like the context vector.

Before sinking into this kind of fatalism, it is worth noting that there are a variety of more addressable issues that might be affecting the results.

The well-known issue of low-frequency artifacts in fMRI signal poses a particularly awkward dilemma. This 'drift' is often dealt with by looking for and removing linear and quadratic trends within runs, as an early preprocessing

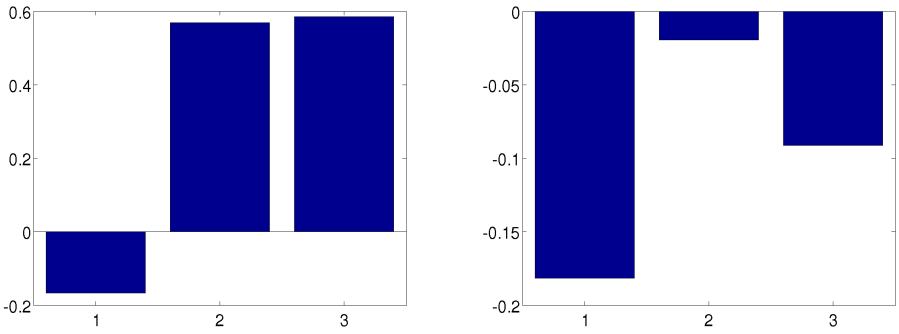

Figure 13: *Left (a): correlation (r) when predicting high vs low disruption tasks. Right (b): correlation (r) when predicting the degree of behavioral forgetting* . In both cases, the values shown are the r values for a given subject, averaged over all 16 runs. The performances are for cross-validation generalization for all 3 subjects, using 200 voxels chosen using the multi-voxel similarity structure algorithm.

step. We have deliberately avoided removing such trends, to avoid accidentally subtracting away the very effect being sought. In order to visualize the timecourses and see how much of an issue this is, Figure 14 shows the mean brain-wide activity for both the low- and high-disruption runs for all three subjects. As a first pass check, these figures provide reassurance that there are no obvious global trends affecting the runs. More work is needed to confirm this in a more careful way, and especially to consider the possibility of artifactual trends on an individual voxel level.

It is noticeable that each run starts with an abnormally high level of activity as the scanner restarts, and there is another period of increased activity just before free recall. These periods coincide with the arithmetic distractors, which are both engaging and involve multiple choice key-presses. These figures use the entire brain as a voxelset, and only provide information about mean *activity*, and tell us very little about changes in the fine-grained patterns so we will not consider them any further.

In its current form, the experimental design for the context experiment had multiple constraints:

- to maximize the amount of context disruption in the high disruption task, minimize it in the low disruption task, and in general, maximize the discrepancy in forgetting between the two conditions
- minimize the opportunities for rehearsing previous items
- make the duration of the lists long enough to allow the haemodynamic lag from prior timepoints to subside

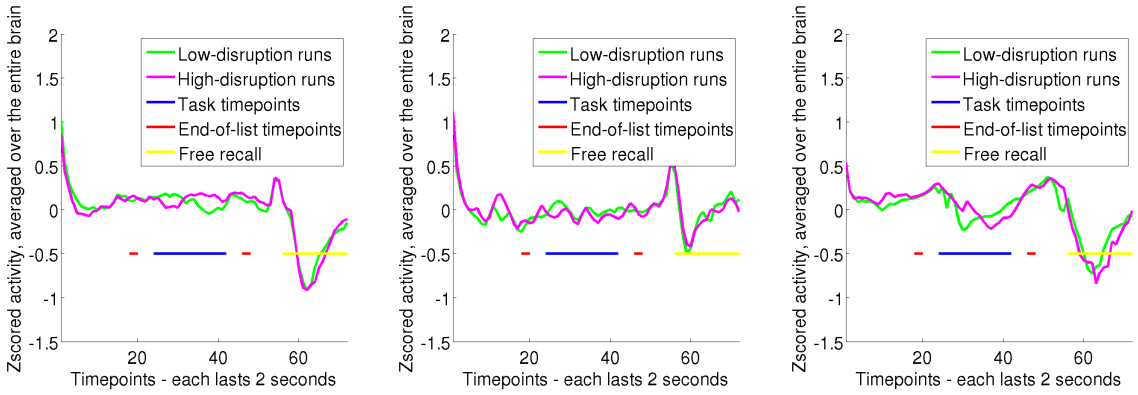

Figure 14: *Timecourses showing the mean brain-wide activity over the course of a run.* Each graph depicts a single subject. The green line shows this brain-wide activity averaged over the low-disruption runs, and the magenta line shows this brain-wide activity averaged over the high-disruption runs. The red lines pick out the timepoints towards the end of the two lists used across which the context change is being measured. The blue line picks out the task timepoints, and the yellow line picks out the free recall timepoints.

- provide as many observations of between-list context change in both conditions as possible

It seems likely that these constraints might be further optimized in a revised design. For instance, it is possible that other tasks might be found that cause either a much smaller or much greater degree of context change. Two other tasks, involving arithmetic problems and stimulating visual images were included in behavioral pilots, but did not merit inclusion. Adding further runs would lengthen the experiment (currently lasting 42 minutes without rest periods), at the risk of tiring subjects, but would provide more observations of between-list context changes. Unfortunately, concatenating multiple scan sessions from the same subject would introduce considerable complication to the model.

One potential flaw in the existing data relates to the stimuli. The words presented were drawn from the Toronto Noun Pool (Thorndike and Lorge, 1944) chosen for their imageability, concreteness and simple phonology. However, no effort was made to exclude semantically-related or similar-sounding words from the same list. As a result, many of the lists contained words with pre-existing associations. These sets of words tended to be recalled better, adding a large source of variance to the behavioral data that is unexplained by the simplified context model being used. The easiest solution would be to exclude related items when constructing future lists to maximize the proportion of variance in the behavioral data accounted for by temporal context model.

There is a more sophisticated alternative. In our characterization of the

TCM, we have only considered the *rate* at which the context vector changes from moment to moment. Drawing on geometric intuitions, the context vector charts a trajectory over a period of time through high-dimensional 'context space', where each point represents the context state at a given moment. The rate of context change is then just the distance travelled through this context space, divided by the time taken. In the full TCM account, the direction taken by the context vector is a fully described as a function of the stimuli being encoded. In a richer model, we might incorporate semantic information about the individual words being presented (Howard et al., in press), which would make extremely precise predictions about the exact trajectory that the context vector should trace. These extra constraints might help define the voxelset more precisely.

There is of course room to question the entire premise of the experiment. In effect, we have described a design for localizing (or defining a voxelset that contains) the context vector. Prefrontal and entorhinal cortical areas seemed likely substrates for the context vector. Both receive projections from all over the brain, allowing them to incorporate a diverse range of information about external and internal state. There is neurophysiological evidence (Howard et al., 2005; Rougier et al., 2005) indicating that both areas may contain specialized cytoarchitectural and neuronal machinery for supporting persistent maintenance of activation patterns, such as for working memory. Such mechanisms might be useful for maintaining a stable context vector, parts of which can be selectively updated in the face of external or internal changes of state. In other words, the context vector has to be able to 'hover' more or less stationary in context space, but also drift in a specified direction and rate as a function of change in mental state. Norman et al. (in press) speculate that the short term memory buffer might actually serve quite well as the context vector, since it changes over time in exactly the way required by TCM. This parsimonious proposal would fit with the constraints described, and help a great deal in concretizing an otherwise purely theoretical construct.

A quite alternative conception of the context vector would banish any such notion of a localized, constantly-updated executive summary of the current mental state. Instead, one could view the entire brain itself as the context vector. After all, the entire brain is stable, but selectively incorporates and is driven by the diverse range of internal and external variables relevant to the context vector.

Despite the failure to predict the degree of behavioral forgetting from context change, prediction of the binary high/low disruption condition appears to be robust in 2 of the 3 subjects. This raises questions of its own, and deserves a note of caution. We describe how the timepoints at the end of the lists were chosen in order to minimize the possibility that the slow ramping down of the haemodynamic response function has smeared information from the task period into the timepoints at the end of the second list. This remains a possibility, since the haemodynamic response function model is just a model, and has been shown to differ from region to region. Aside from that, there are other extraneous factors that might be helping with this prediction. For instance, it may be

that subjects find the imagination task more arousing, and that these global physiological changes appear as a change in context, though an argument could be made that such physiological changes *are* context changes. Less interestingly, it could be that one of the tasks elicits greater head motion than the other, which could look like a brain-wide drift, though the head motion parameters estimated by the volume registration do not support this hypothesis,

In the future, we hope to enrich the model used even further. It might be possible to define a voxelset objective function that takes the free recall data into account in the following way. TCM proposes that the reinstatement of prior context during recall occurs automatically as a product of retrieving memories, and in so doing, alters the cues used in future retrieval. Sahakyan & Kelley demonstrated this behaviorally with their 'think back to the beginning of the experiment' reinstatement manipulation. Although our design does not incorporate any such instructions, we might expect to see such reinstatement evidence by the changes in context state during the final free recall. Indeed, this would be a very natural extension of the Polyn et al. (2005) design - where they tracked the dynamics of free recall using reinstatement of semantic category, we could track the dynamics of free recall using encoding periods. Where they showed that the brain resembles its 'celebrity face' encoding state just before recalling celebrity faces, we might hope that the brain will resemble its 'first list of run 5' encoding state just before recalling a slew of items from the first list of run 5.

6 Conclusion

We have provided evidence that by taking multiple voxels and multiple conditions into account at once, the similarity structure algorithm may provide a more sensitive measure of a model than the mass univariate GLM. We have shown how to frame existing models in terms of their similarity structure, and attempted to introduce a novel paradigm that attempts to forge closer ties between theory, neural activity patterns and behavior.

References

Bryan R, Haxby JV (2006) Stepwise k-nearest neighbor classification In *Meeting of the Organization for Human Brain Mapping*, Florence, Italy.

Cox RW (unpublished) The AFNI 'waver' haemodynamic response function. *http://afni.nimh.nih.gov/afni/doc/faq/17/* .

Cox RW, Jesmanowicz A (1999) Real-time 3D image registration for functional MRI. *Magnetic resonance in medicine* .

DeSimone K, Schneider KA, Weiner KS, Kastner S (submitted) Topographic maps in human frontal cortex revealed in delayed saccade and spatial working memory tasks .

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

Engel SA, Glover GH, Wandell BA (1997) Retinotopic organization in human visual cortex and the spatial precision of fMRI. *Cerebral Cortex* .

Engelien A, Yang Y, Engelien W, Zonana J, Stern E, Silbersweig D (2002) Physiological mapping of human auditory cortices with a silent event-related fMRI technique. *Neuroimage* .

Estes WK (1955) Statistical theory of spontaneous recovery and regression. *Psychological Review* .

Godden DR, Baddeley AD (1975) Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology* .

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–2429.

Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience* 8:686–691.

Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* .

Howard MW, Addis KM, Jing B, Kahana MJ (in press) Semantic structure and episodic memory In Landauer T, McNamara D, Dennis S, Kintsch W, editors, *LSA: A Road Towards Meaning*. Lawrence Erlbaum, Mahwah, NJ.

Howard MW, Fotedar MS, Datey AV, Hasselmo ME (2005) The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review* 112:75–116.

Howard MW, Kahana MJ (2002) A distributed representation of temporal context. *Journal of Mathematical Psychology* 46:269–299.

Kahana MJ (1996) Associative retrieval processes in free recall. *Memory and Cognition* 24:103–109.

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* 8:679–85.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proceedings of the National Academy of Sciences* 103:3863–3868.

Mazzola L, Isnard J, Mauguiere F (2005) Somatosensory and pain responses to stimulation of the second somatosensory area (SII) in humans. a comparison with SI and insular responses. *Cerebral Cortex* .

Mensink GJM, Raaijmakers JGW (1989) A model for contextual fluctuation. *Journal of Mathematical Psychology* .

Mitchell M (1996) *An Introduction to Genetic Algorithms* MIT Press, Cambridge, MA.

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S (2004) Learning to decode cognitive states from brain images. *Machine Learning* 5:145–175.

Norman KA, Detre GJ, Polyn SM (in press) Computational models of episodic memory In Sun R, editor, *The Cambridge Handbook on Computational Cognitive Modeling*. Cambridge University Press.

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive science* .

O'Toole AJ, Jiang F, Abdi H, Haxby JV (2005) Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience* 17:580–590.

Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes recall during memory search. *Science* 310:1963–1966.

Rougier NP, Noelle D, Braver TS, Cohen JD, O'Reilly RC (2005) Prefrontal cortex and the flexibility of cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences* .

Russell S, Norvig P (2002) *Artificial Intelligence: A Modern Approach* Prentice Hall.

Sahakyan L, Kelley CM (2002) A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology* .

Schneider W, Bartels A, Formisano E, Haxby JV, Goebel R, Mitchell T, Nichols T, Siegle G (2006) Competition: Inferring experience based cognition from fMRI. *Proceedings of the Organization of Human Brain Mapping Florence Italy June 15 - http://www.ebc.pitt.edu* .

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brandy TJ (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* .

Shepard RN (1980) Multidimensional scaling, tree-fitting and clustering. *Science* .

Thorndike E, Lorge I (1944) *The Teacher's Word Book of 30,000 Words* New York: Teachers College, Columbia University.

Tootell RB, Reppas JB, Kwong KK, Malach R, Born RT, Brady TJ, Rosen BR, Belliveau JW (1995) Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *Journal of Neuroscience* .

Wandell BA (2000) The new cognitive neurosciences. *Computational neuroimaging: color representations and processing* .

Worsley K, Friston K (1995) Analysis of fMRI time-series revisited - again. *Neuroimage* .