Simitar: simplified searching of statistically significant similarity structure

Francisco Pereira
Imaging and Computer Vision
Siemens Corporation, Corporate Technology
755 College Road E, Princeton NJ 08540
francisco-pereira@siemens.com

Matthew Botvinick
Psychology Department and Princeton Neuroscience Institute
Princeton University
Princeton NJ 08540
matthewb@princeton.edu

***Abstract*—This paper describes Simitar, a toolbox for studying the similarity structure of patterns of brain activation in different experimental conditions. We focus on supporting two types of analysis, namely, the calculation of local similarity matrices for all locations in the brain and the identification of locations where similarity has a desired structure, via an intuitive interface.**

I. Introduction

Over the last five years there has been a growing interest in directly studying the representations of different stimuli in the brain [5]. This can be done purely to compare the similarity of representations [1] in different regions of interest (ROIs), or to compare them with those coming from animal studies [4], behavioural similarity judgements [9], models or other hypothesized structure of similarity between those stimuli [1], or even to perform decoding tasks across subjects [8].

The work above spans a range of spatial scales of similarity structure, and features different notions of what makes a particular structure interesting. We believe that those various possibilities can be encompassed by two general analysis procedures:

1) Producing a *similarity map* - This entails finding a similarity matrix in every possible location of interest; a location can be anything from a small $3 \times 3 \times 3$ searchlight to an entire ROI.
2) Producing a *similarity structure score map* - This requires specifying what characteristics are (un)desirable in a similarity matrix and producing a numeric score for how well a given matrix embodies those characteristics; at the simplest level, one should be able to express something like "condition A is similar to condition B but dissimilar to C, with other conditions being irrelevant". Given this, a score can them be computed for each location in the similarity map.

Simitar is a MATLAB/Octave toolbox that implements these procedures in an efficient manner; an entire similarity map containing correlation or euclidean distance matrices for every searchlight in the brain can be produced in seconds. The toolbox can be used in an exploratory fashion, where one might display all similarity matrices in a slice of interest or the similarity structure score map for the entire brain. It can also be used to test hypotheses about the presence of certain similarity characteristics, as it runs fast enough to make permutation tests feasible.

Finally, we note that it would be possible to run similar analyses using a general-purpose toolbox for analyzing brain imaging data, such as PyMVPA [3]. The main advantage of Simitar is that, in focusing solely on the key procedures, it has been optimized to perform them very efficiently. It also makes it easy for a user to try it as a black box, as there is little to learn besides how to prepare data for use and how to specify the similarity structure one is looking for. Finally, it is simple for a more advanced user to implement different methods for scoring desired similarity structures, and to write their own permutation tests around the core code.

[1]Sometimes what is considered is the dissimilarity between representations; we will use "similarity" throughout the paper to refer to both kinds.

II. Demonstration

A. Mock dataset

We will demonstrate the capabilities of Simitar using a mock dataset that simulates the results of an experiment in phonetic perception, using as stimuli video recordings of someone saying a particular phoneme. The experiment aims to study the McGurk Effect [6], a phenomenon whereby a subject hearing audio of someone saying a phoneme (BA) and video of them articulating a different phoneme (GA) perceives a third phoneme (DA). Experiment trials belong to one of these four conditions:

1) three where subjects see audio and video of someone saying BA, GA or DA, labelled with the phoneme name
2) one where they see mismatched audio (BA) and video (GA), which we will label MC (for McGurk)

The mock brain has a single slice, divided into 4 regions-of-interest (ROIs), associated with auditory, visual and perceptual representations, as well as "other things" that are common to all conditions. This is shown in Figure 1.

In each condition a subject will hear, see and perceive something, so each condition gives rise to a pattern of

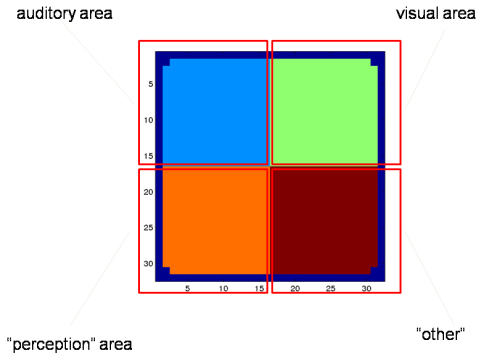

Figure 1. Mock brain, divided into ROIs

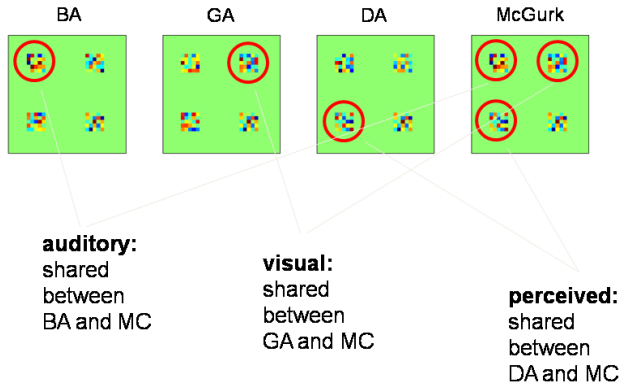

Figure 2. The patterns of activation across the brain, in the 4 conditions

activation in each of the four ROIs. We designed the dataset so that our experimental hypothesis was true:

- BA and MC share a pattern in the auditory ROI (subject hears the same)
- GA and MC share a pattern in the visual ROI (subject sees the same)
- DA and MC share a pattern in the "perception" ROI (subject perceives the same)
- all 4 conditions have the same pattern of activation over the remaining ROI

and this is illustrated in Figure 2.

The patterns of activation in each condition were used to generate a dataset by corrupting them with noise for each trial. The dataset produced had one such image per trial, and 10 trials of each condition in each of 4 "runs".

B. Producing similarity maps

Simitar can produce a local similarity matrix between the patterns of activity in all conditions, considered over all the voxels inside a given searchlight; the measure can be correlation, euclidean distance or a number of others. Figure 3 shows the correlation and euclidean distance matrices obtained in the $3 \times 3 \times 3$ searchlight around each voxel on the mock brain, laid out on their respective locations.

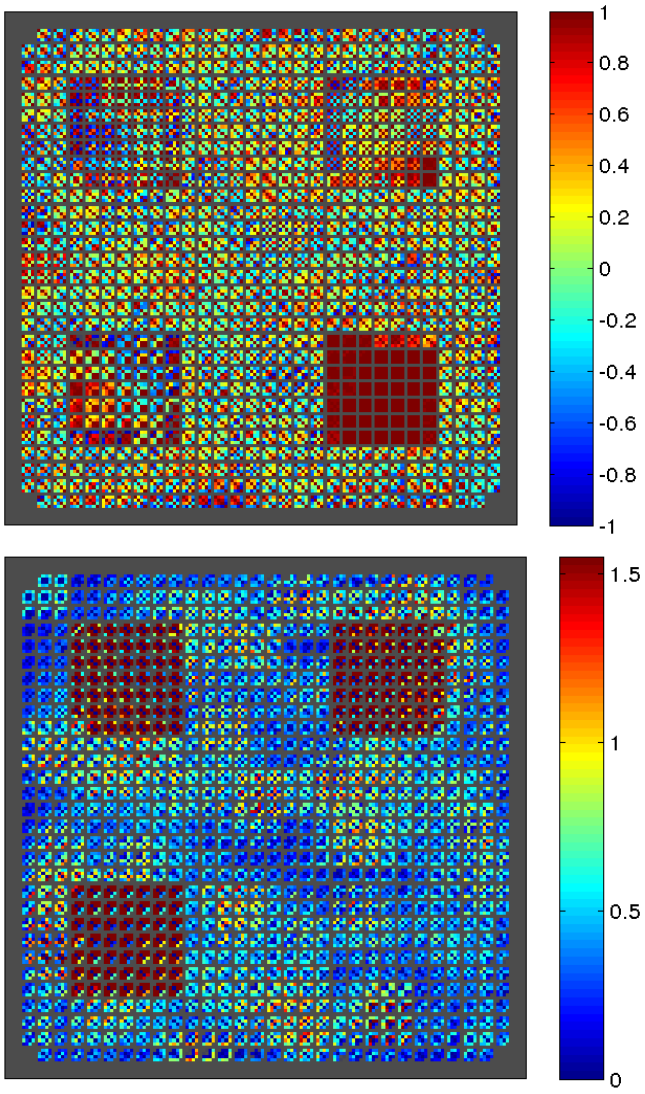

Figure 3. **Top:** correlation matrices for all searchlights in the mock brain **Bottom:** same as top, using euclidean distance

C. Producing similarity structure score maps

A similarity map can be used for exploratory data analysis, or if locations of interest are specified a priori. If, instead, the goal is to find locations where similarity structure displays certain characteristics, we first need to specify those. For instance, let us consider correlation matrices and suppose that we would like to find locations where BA is represented similarly to McGurk but differently from everything else. Intuitively, we want matrices where the correlation between BA and McGurk is high but that between BA and everything else (and McGurk and everything else) is low.

Simitar allows you to specify this through a *similarity structure scoring* scoring matrix

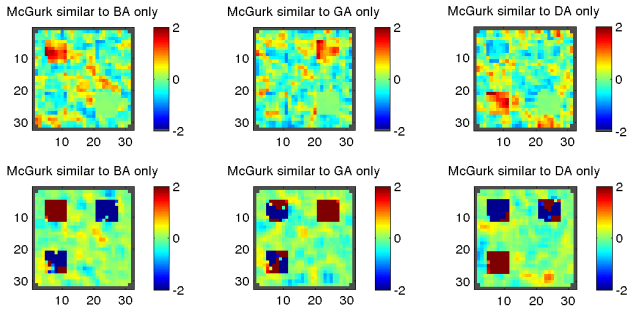

Figure 4. **Top:** similarity structure score maps for locations where each condition is similar to McGurk but dissimilar from others, using correlation **Bottom:** same as top, using euclidean distance

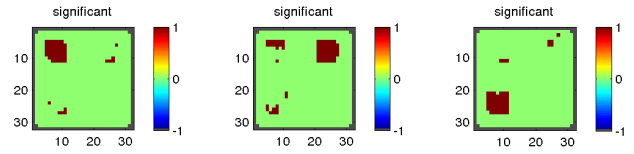

Figure 5. Locations where the result of the similarity score test was deemed significant.

```
    BA  GA  DA  MC
BA   0  -1  -1  +1
GA  -1   0   0  -1
DA  -1   0   0  -1
MC  +1  -1  -1   0
```

This matrix is multiplied elementwise by each searchlight similarity matrix and the elements of the resulting matrix are summed to produce a score. As desired, the higher the correlation between BA and McGurk the higher the score, but it will be penalized by correlation between other conditions and McGurk (0 entries are ignored). Simitar will, by default, automatically scale entries of this matrix so that the weight of rewards and penalties is balanced (i.e. penalty entries become $-\frac{1}{4}$, reward entries remain 1). We can thus produce a *similarity structure map* for the similarity structure scoring matrix above as well as analogous matrices for GA and DA, as shown at the top of Figure 4.

The same may be done using euclidean distance, except in this case the matrix should reward closeness between representations and penalize distance, i.e. again for BA similar to McGurk but different from everything else

```
    BA  GA  DA  MC
BA   0  +1  +1  -1
GA  +1   0   0  +1
DA  +1   0   0  +1
MC  -1  +1  +1   0
```

and the resulting map is shown at the bottom of Figure 4.

D. Statistical testing of similarity structure score maps

The similarity structure score maps above can be used for exploratory data analysis, but it is also possible to transform them into p-value maps by using permutation tests. Simitar supports two varieties:

1) Permute over example labels - If there are many examples of each condition, we can permute over their labels, within each run, and obtain a similarity structure map for that permutation. Repeated over many permutations, this yields a permutation distribution for the score at each voxel and a p-value for the score obtained using the original labels.
2) Permute over entries of the scoring matrix - If there are one or very few examples of each condition (e.g. one used deconvolution over many trials to get a single beta coefficient image as the example for one condition) there will likely not be enough examples to permute over example labels. One can, instead, permute over all distinct pairs of conditions in the score matrix (e.g. in our mock dataset there are only six, so hence 6!=720 permutations).

Note that the implicit null hypotheses are different in each case. In the first, we are assuming that there is no information in the condition labels, and this is the more typical test. In the second the implicit assumption is that the specific similarity structure scoring matrix selected does not matter (among all scoring matrices with the same numbers of -1/0+1). Figure 5 shows the locations where correlation similarity structure score was deemed significant, using 10000 permutations (of the first variety) and correcting for multiple comparisons using False Discovery Rate [2],($q = 0.01$), for similarity structure scoring matrixes picking BA similar to MC, GA similar to MC and DA similar to MC, respectively.

E. Real dataset

We used Simitar to analyze the dataset from the paper "Predicting Human Brain Activity Associated with the Meanings of Nouns" [7], which the authors have very kindly made public [2]. The examples used are from subject P1 and belong to one of 12 semantic categories [3]. For illustration we will look for locations where 'kitchen utensils' are represented similarly to 'tools' but differently from everything else. The score matrix would then be (columns 8 and 10 correspond to 'kitchen utensils' and 'tools'):

```
 0  0  0  0  0  0  0 -1  0 -1  0  0
 0  0  0  0  0  0  0 -1  0 -1  0  0
 0  0  0  0  0  0  0 -1  0 -1  0  0
 0  0  0  0  0  0  0 -1  0 -1  0  0
 0  0  0  0  0  0  0 -1  0 -1  0  0
 0  0  0  0  0  0  0 -1  0 -1  0  0
 0  0  0  0  0  0  0 -1  0 -1  0  0
-1 -1 -1 -1 -1 -1 -1  0 -1 +1 -1 -1
 0  0  0  0  0  0  0 -1  0 -1  0  0
-1 -1 -1 -1 -1 -1 -1 +1 -1  0 -1 -1
```

[2]http://www.cs.cmu.edu/~tom/science2008/

[3]The categories are 'animal', 'body parts', 'buildings, 'building parts, 'clothing', 'furniture', 'insect', 'kitchen utensils', 'man-made objects', 'tools', 'vegetable' and 'vehicle'.

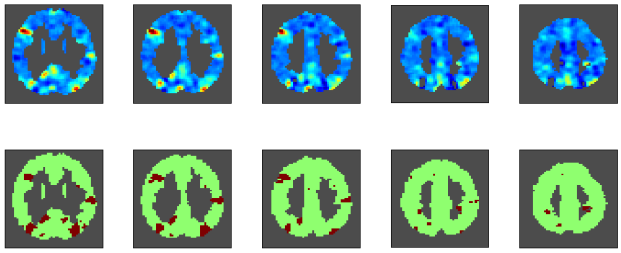

Figure 6. **Top:** Similarity structure score maps for 'kitchen utensils' similar to 'tools', using euclidean distance. **Bottom:** Locations where the structure score was deemed significant, using 10000 permutations and $FDR = 0.01$.

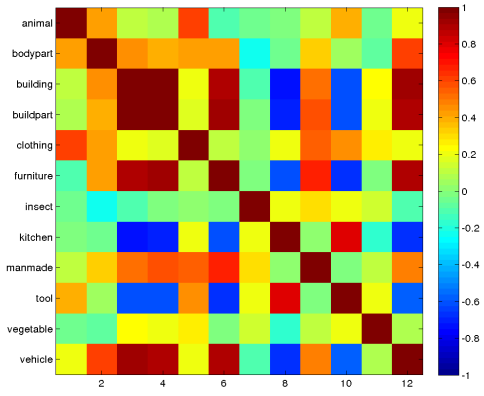

Figure 7. Correlation matrix in the location with the highest similarity structure score.

```
 0   0   0   0   0   0   0 -1   0 -1   0   0
 0   0   0   0   0   0   0 -1   0 -1   0   0
```

and, flipped for euclidean distance, would give rise to the structure score and significance maps shown in Figure 6. Several hundred voxels are deemed significant and those with the highest scores are in reasonable AAL ROIs in the occipital (`Occipital_Mid_L` and `Calcarine_R`), temporal (`Lingual_R` and `Fusiform_R`) and frontal (`Frontal_Inf_Tri_R`) lobes. Interpreting these is beyond the scope of this paper.

If we now look at the correlation matrix in the location with the highest score, show in Figure 7, it is clear that it matches the structure we specified, in that 'kitchen utensils' and 'tools' are represented similarly to each other, but dissimilarly from other things. Note that 'buildings', 'building parts' and 'furniture' are also represented similarly to each other in the same location; the structure matrix did not reward or penalize this. In order to understand what was being represented you could now look at the patterns of activation in the appropriate searchlight; it is possible that certain conditions have similar representations because there is little activation, for instance.

The dataset used had 360 examples and around 20000 voxels. It took Simitar ~ 0.3 seconds to compute all $3\times3\times3$ searchlight correlation matrices and $\sim$ 0.17 seconds to compute all euclidean distance matrices; the times using `pdist` and `squareform` in MATLAB were $\sim$ 7 seconds and ~ 5 seconds, respectively, sharing the rest of the toolbox code. Note that the speed in all cases is due to the fact that all searchlight neighbourhood relationships are pre-computed before running the code (a step which takes $\sim$ 1 second). All timings were obtained by computing the average of 30 runs on a 2.5GHz Intel Core i5 machine running MATLAB R2012a.

F. Distribution

Simitar is available online from http:\\minerva.csbmb.princeton.edu\simitarbeta. The site includes a demo, synthetic data for the mock brain and tutorials covering both data preparation and the use of the toolbox to produce all the plots shown in this paper.

References

[1] Andrew C Connolly, J Swaroop Guntupalli, Jason Gors, Michael Hanke, Yaroslav O Halchenko, Yu-Chien Wu, Hervé Abdi, and James V Haxby. The representation of biological classes in the human brain. *The Journal of Neuroscience*, 32(8):2608–2618, 2012.

[2] Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

[3] 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.

[4] Nikolaus Kriegeskorte and Marieke Mur. Representational similarity analysis of object population codes in humans, monkeys, and models. *Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging*, page 307, 2012.

[5] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008.

[6] John MacDonald and Harry McGurk. Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3):253–257, 1978.

[7] 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, 2008.

[8] Rajeev DS Raizada and Andrew C Connolly. What makes different people's representations alike: neural similarity space solves the problem of across-subject fmri decoding. *Journal of cognitive neuroscience*, 24(4):868–877, 2012.

[9] Matthew Weber, Sharon L Thompson-Schill, Daniel Osherson, James Haxby, and Lawrence Parsons. Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia*, 47(3):859–868, 2009.