# The geometry of human perception: RSA and multivariate models

Charles Zheng

Stanford University

November 16, 2015

(Joint work with Yuval Benjamini and Oluwasanmi Koyejo)

# Overview

*fMRI Background:*

- Nonparametric approaches: RSA.
- Parametric approach: Multivariate linear model.

*Questions:*

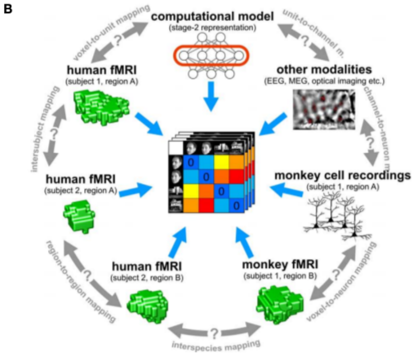
- Defining the RSA null and alternative hypotheses.
- Scientific interpretation of RSA results.
- Sensitivity to preprocessing choices.

*Proposed Projects:*

- Distribution-induced distance.
- Parametric RSA.

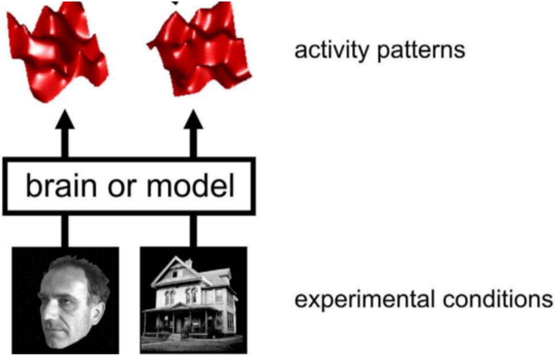
# Representation similarity analysis (RSA)

- Framework for studying how mental objects are represented in the brain, via brain activity (measured by fMRI, EEG) or behavior.
- Compare different brain regions or imaging modalities within a single subject, or compare multiple subjects.



# A typical RSA experiment

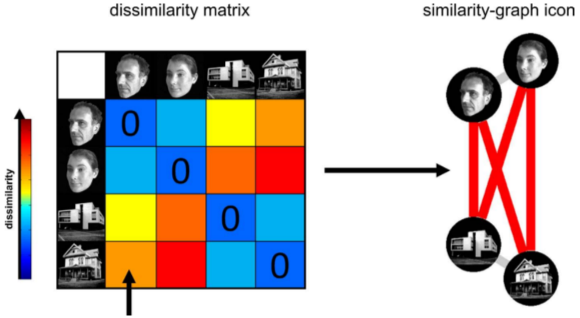
An experiment which demonstrates which regions of the brain differentiate between faces and objects.




Step 1: Present the subject with visual stimuli, pictures of faces and houses. Record the subject's brain activity in the fMRI scanner.

# A typical RSA experiment

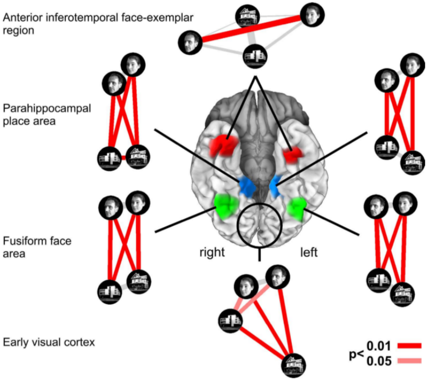
Step 2a: Process the data, and represent the brain activity of the subject for the $i$th stimulus as a real vector $y_i$. Form matrix of distances between $y_i$ and $y_j$, the *representation distance matrix* (RDM).




Step 2b: Assess statistical significance of distances to form similarity graph.

# A typical RSA experiment

Step 3: Compare similarity graphs between different brain regions.




Step 4: Draw scientific conclusions. (Step 5: Profit!!..?)

# Details of RSA

- Core methodology presented by Kriegeskort et al (2008) and extended by others.
- Suppose each of the stimuli have $r$ repeats, the responses for stimulus $i$ are $y_i^1, \ldots, y_i^r$, and the average over the repeats as $\bar{y}_i$.
- The representation distance matrix is computed as

$$D_{ij} = d(\bar{y}_i, \bar{y}_k)$$

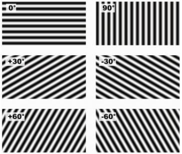
where $d$ may be Euclidean distance or correlation distance.

# Details of RSA

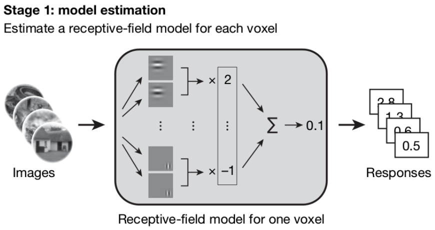
- Now let $D^A$, $D^B$ be *two* such distance matrices, e.g. two regions from same subject, or same region in two subjects.
- *Estimation.* One can define a distance metric $\mathfrak{d}$ *between* distance matrices, e.g. the Spearman correlation between the entries of $D^A$ and $D^B$, and estimate
$$\mathfrak{d}(D^A, D^B).$$
- *Testing.* One can test:
    - *Independence* of $D^A, D^B$.
    - *Equality* $D^A = D^B$.
    - Which of $D^A, D^B$ is closer to a reference distance matrix $D^0$.

# Comparison to parametric approach

- RSA is a "nonparametric" approach, because stimuli are treated as discrete classes.
- In contrast, one could consider presenting stimuli which are parameterized.
- Example: present the subject with gratings of varying orientation. Orientation $x$ is parameter.



# Example of parametric approach: natural images




- Kay et al (2008) parameterize natural images using *Gabor filters*.
- Let $x_i$ be the vector of 10000 Gabor filter coefficients for a natural image. Let $y_i$ be the vector of 20000 individual voxel responses.
- Kay fits a model of the form

$$y_i = B^T x_i + \epsilon_i$$

where $B$ is a 10000 x 20000 coefficient matrix, and $\epsilon_i$ is vector-valued noise with covariance $\Sigma$.

# Comparison of parametric and nonparametric approaches

Nonparametric: RSA

| Pros | Cons |
|---|---|
| Compare across subjects, regions, modalities | Can't generalize to new stimuli |

Parametric: Regression

| Pros | Cons |
|---|---|
| Predictive and descriptive power | Requires knowing featurization *a priori* |

See also Kriegeskorte and Bandettini (2007) "Analyzing for information..."

# Defining the RSA null

- Consider the problem of testing *equality* $D^A = D^B$.
- Obviously, the matrices $\hat{D}^A$ and $\hat{D}^B$ computed from data will *not* be equal!
- We need to define the *population* parameters $D^A$, $D^B$ in order to have a well-defined test.

# Possible definitions of population distance matrices

- Option 1: Let $\mu_i = \mathbf{E}[y_i]$ averaged over repetitions, then define the population parameter as

$$D_{ij} = d(\mu_i, \mu_j)$$

  This option *ignores noise*

- Option 2: Define

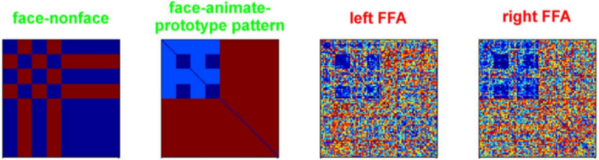
$$D_{ij} = \mathbf{E}[d(y_i^1, y_j^1)]$$

  where the average is over a single repetition. This option *includes the effect of noise* in the population parameter. Also, $\hat{D}_{ij}$ is an unbiased estimator of $D_{ij}$ if only one repeat per stimulus is used.

# Defining the RSA null

- The most commonly used test in RSA is a test of *independence* $D^A$ of $D^B$.
- This approach was suggested by Kriegeskorte (2008) and is well-established in ecological data analysis.
- Implemented via Mantel and partial Mantel tests (which use permutation).
- However, how can we define the null hypothesis of "independence"?? No matter how the *population* matrices $D^A$ and $D^B$ are defined, they are *deterministic* and hence it does not make sense to consider them *dependent*.
- Perhaps you could test if the entry-wise correlation is zero??

# Scientific interpretation of RSA results




A rejection of the independence null between $D^A$ and $D^B$ is taken to mean that outputs $A$ and $B$ are 'related.' For instance, $D^A$ might be the response from a subject's brain region, and $D^B$ is a 0-1 matrix reflecting *a priori* class membership. Rejection is taken to mean that the region $A$ have differential activation depending on the classes represented $D^B$.

# Scientific interpretation of RSA results

What does it mean to reject $D^A = D^B$? We can conclude that the means $\mu_i^A$ are *not* related to the means $\mu_i^B$ by a scaling factor and orthogonal rotation:

$$\mu_i^A \neq kH\mu_i^B$$

where $H^T H = I$.

But we have not ruled out the possibility that $\mu_i = \psi(\mu_j)$ for some bijection. Hence we can only draw a very weak conclusion about the difference between $D^A$ and $D^B$.

# Sensitivity to preprocessing choices

Even given the same raw data, there are a variety of choices for representing the response vectors $y_i$:

- Volumetric (voxels) vs surface-based coordinates;
- Voxel size and centering;
- Smoothing;
- Image registration;
- Representation of data via function basis.

The resulting distance matrix $D_{ij}$ depends on the above choices.

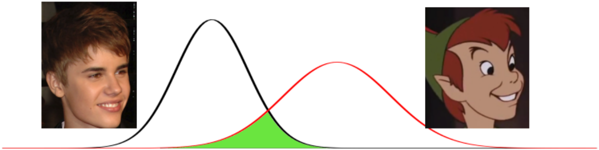
# Distribution-induced distance

A new definition for the *representation dissimilarity matrix* which:

- Allows more natural interpretation of the null $D^A = D^B$;
- Is nearly invariant to smoothing, registration, use of function bases;
- Downside: may be harder to estimate.

# Distribution-induced distance: motivation

Consider two stimuli $x_1$ and $x_2$ to be *distant* if the *response distributions* are statistically distant, or *close* if their response distributions overlap.



Note that the definition not only depends on the difference in *means* but also depends on the *noise distribution*.

# Distribution-induced distance: definition

- Let $\mathcal{F}_x$ denote the distribution of the reponse $y$ conditional on the stimulus $x$.
- Define the dissimilarity matrix

$$D_{ij} = \mathbb{D}(\mathcal{F}_{x_i}; \mathcal{F}_{x_j})$$

  where $\mathbb{D}$ is a measure of distance or divergence between probability measures.
- Example: if $y$ is conditionally multivariate normal, with covariance $\Sigma$ not depending on $x$, then

$$D_{ij} = \frac{1}{2}(\mathbf{E}[y|x_i] - \mathbf{E}[y|x_j])^T \Sigma^{-1} (\mathbf{E}[y|x_i] - \mathbf{E}[y|x_j])$$

  for either $\mathbb{D} =$ KL divergence or Hellinger distance.

# Distribution-induced distance: properties

- Invariance under bijections: Let $\tilde{y} = \psi(y)$ for some bijection $\psi$. Then if $\mathcal{F}_x^{\tilde{y}}$ denotes the conditional distribution of $\tilde{y}$ given $x$, we have

$$\mathbb{D}(\mathcal{F}_x^y, \mathcal{F}_{x'}^y) = \mathbb{D}(\mathcal{F}_x^{\tilde{y}}, \mathcal{F}_{x'}^{\tilde{y}})$$

  for any $f$-divergence $\mathbb{D}$.

- Near-invariance under sampling. Suppose the 'true' brain activity is represented by a random process $f$, and let $\mathcal{F}_x$ denote its conditional distribution given $x$. Define the vector $y$ as the values of linear functionals $\Lambda_1, \ldots, \Lambda_q$ evaluated on $f$. This corresponds to taking $y$ based on binning the signal $f$ or taking coefficients of $f$ from a function basis. Then, supposing $\Lambda_1, \ldots$ is 'dense'

$$\lim_{q \to \infty} \mathbb{D}(\mathcal{F}_x^y, \mathcal{F}_{x'}^y) = \mathbb{D}(\mathcal{F}_x^f, \mathcal{F}_{x'}^f).$$

# Distribution-induced distance: consequences

What does it mean to reject $D^A = D^B$?
Invariance under bijections means that $y^A$ and $y^B$ are different in a stronger sense: not only is it the case that $y^A$ and $y^B$ have different distributions, but we can also rule out the possibility that $y^A$ and $y^B$ are related by maps $\psi^{A\to B}$ and $\psi^{B\to A}$ where

$$y^A|x \overset{D}{=} \psi^{A\to B}(y^B)|x$$

$$y^B|x \overset{D}{=} \psi^{B\to A}(y^B)|x.$$

Hence the conditional distributions of $y^A$ and $y^B$ *carry different information.*

# Distribution-induced distance: consequences

Near-invariance under sampling means that the population distances $D_{ij}$ are robust to choices in coordinates, smoothing, etc, *supposing that sufficiently many dimensions are used*.

This is because any choice of extracting the vector $y$ from the 3-dimensional function-valued signal amounts to a choice of linear functionals $\Lambda_1, \ldots, \Lambda_q$. For example:

- Voxels + smoothing: each coordinate of $y_i$ takes the form

  $$y_i = \int \phi(z - c_i) f(z) fz$$

  where $\phi$ is a gaussian kernel, $c_i$ is the center of the $i$th voxel and $f(z)$ is the "true signal"

- Function basis:

  $$y_i = \int \phi_i(z) f(z) fz$$

  where $\{\phi_1, \ldots, \phi_q\}$ is the function basis.

# Parametric RSA

- Combine the parametric approach of multivariate regression with RSA.
- Model:
$$y \sim N(B^T x, \Sigma).$$
- The distribution-induced metric is therefore
$$D(x_i, x_j) = (x_i - x_j)^T B \Sigma^{-1} B^T (x_i - x_j).$$

# Parametric RSA

$$y \sim N(B^T x, \Sigma)$$
$$D(x_i, x_j) = (x_i - x_j)^T B\Sigma^{-1} B^T (x_i - x_j)$$

- Since all information about the distance is captured by the matrix $M = B\Sigma^{-1}B^T$, instead of testing

$$D^A = D^B$$

  we can test

$$M^A = M^B.$$

- We can compare two datasets with non-overlapping stimuli
- The approach is scalable in the number of distinct stimuli, since the size of $M^A$, $M^B$ only depend on the number of *features* rather than the number of stimuli

# References


- Kriegeskorte, N. (2008). Representational similarity analysis connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience
- Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. NeuroImage, 38(4), 649662. doi:10.1016/j.neuroimage.2007.02.022
- Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. Nature, 452(March), 352355. doi:10.1038/nature06713
- Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. Methods in Ecology and Evolution, 4(4), 336344. doi:10.1111/2041-210x.12018
- Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. Nature Neuroscience, 16(9), 13481355. doi:10.1038/nn.3470