Reverse-engineering the brain

Charles Zheng

Stanford University

January 29, 2016

HUMAN ENGINEERING
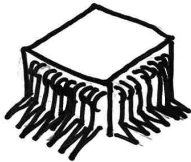
NATURAL SYSTEM

coarse grained

$\mathbb{E}[\cdot]$
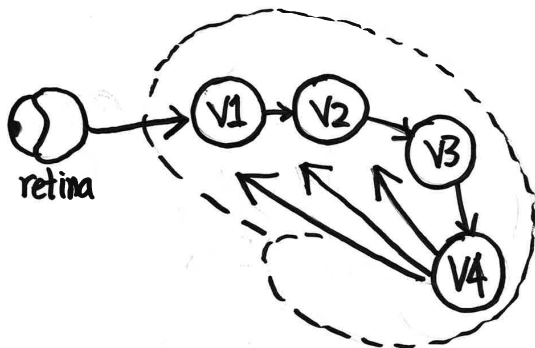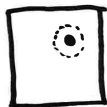$Var[\cdot]$

fine grained

$\frac{dx}{dt} = f(x)$

Level	Neuroscience	Statistics
Reductionist	Segmentation	Clustering
Topological	Connectomics	Graphical Lasso
Information Flow	RSA?	**this thesis!**
Explicit Mechanisms	Encoding models	Regression

STIMULUS   retina   V1 → V2 → V3 → V4

Retina   V1   V2?   · · ·   V4??

Voxel 1

3.2   2.2   4.3

Voxels (~10,000)

$$Y =$$

3.2
2.2
4.3
:
:

· · ·

Stimuli (~1000)
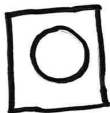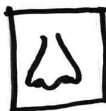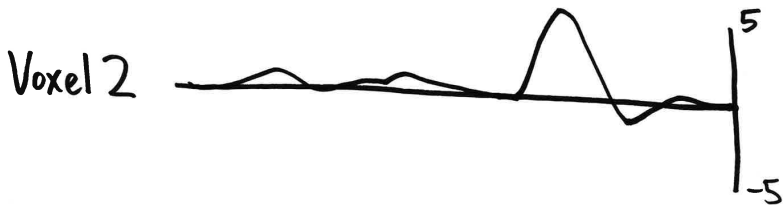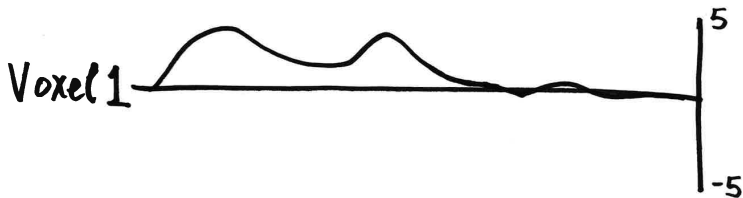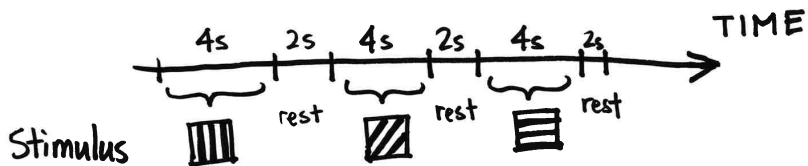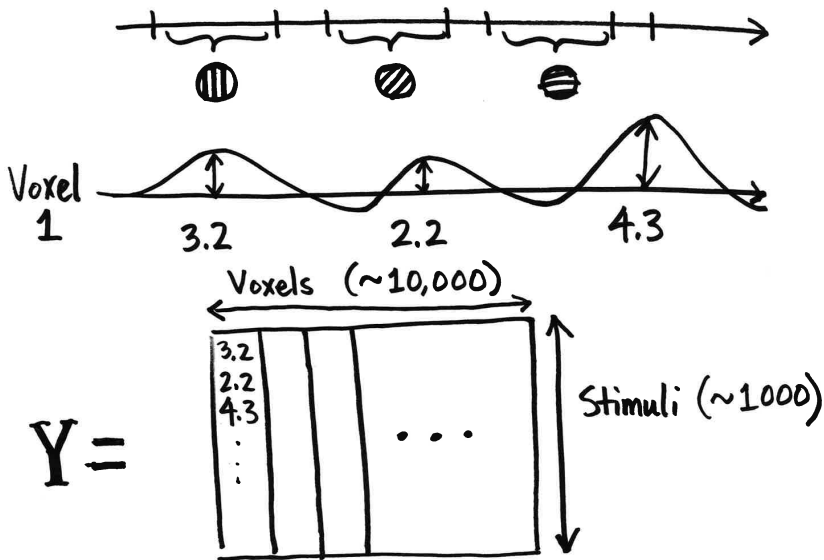
Stimuli x	Response y
$\begin{pmatrix} 1.0 \\ 0 \\ 3.0 \\ 0 \\ -1.2 \end{pmatrix}$	$\begin{pmatrix} 1.2 \\ 0 \\ -1.8 \\ -1.2 \end{pmatrix}$
$\begin{pmatrix} 0 \\ -2.2 \\ -3.1 \\ 4.5 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1.2 \\ -1.9 \\ 0.5 \\ 0.6 \end{pmatrix}$
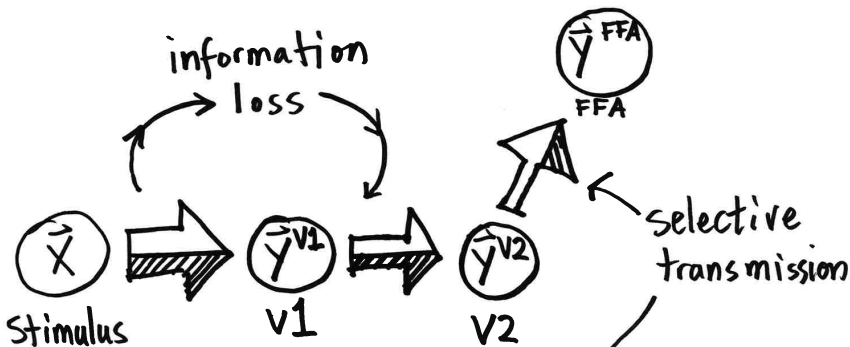
classification

Part I
of talk

information

distance
metric

Part II of
talk
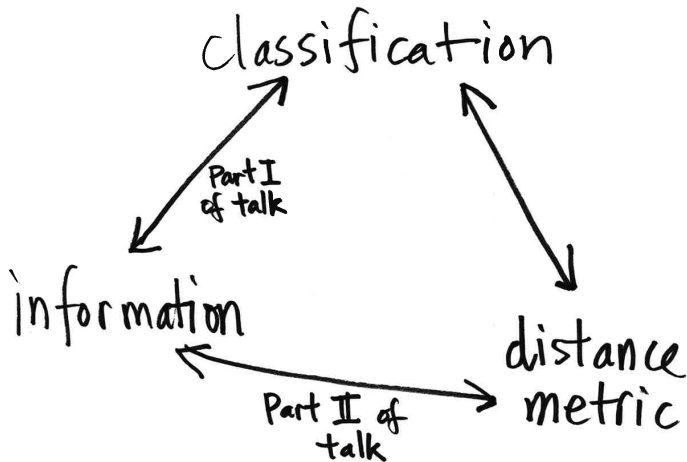
Training Data
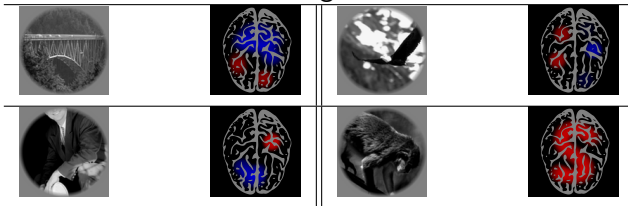
Test Data
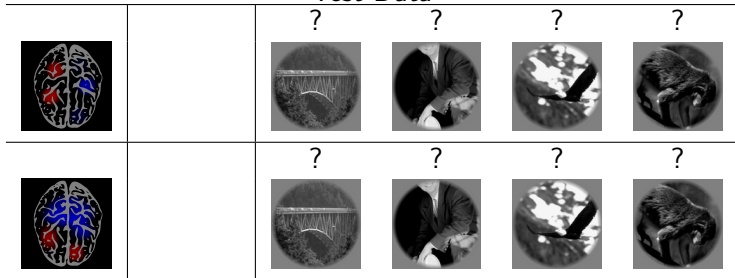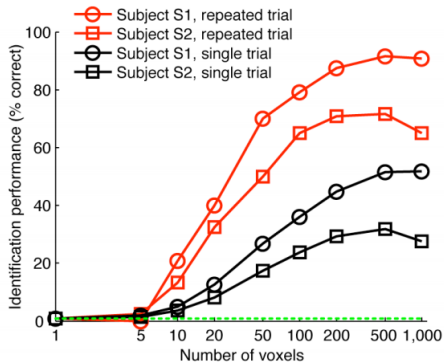
Redundancy of neural coding
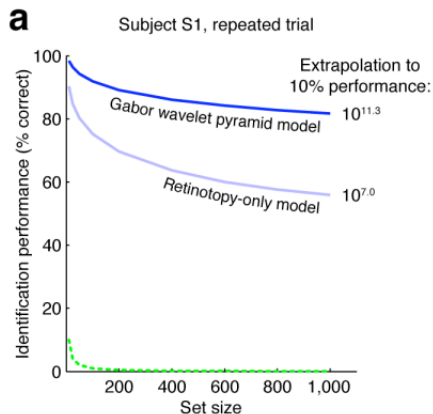
Can we extrapolate the classification curve?

How many neurons does it take to classify a lightbulb?

Charles Zheng

Stanford University

January 6, 2016

(Joint work with Yuval Benjamini.)

Experimental design

Sample stimuli iid from $p(\mathbf{x})$. Repeated measures.
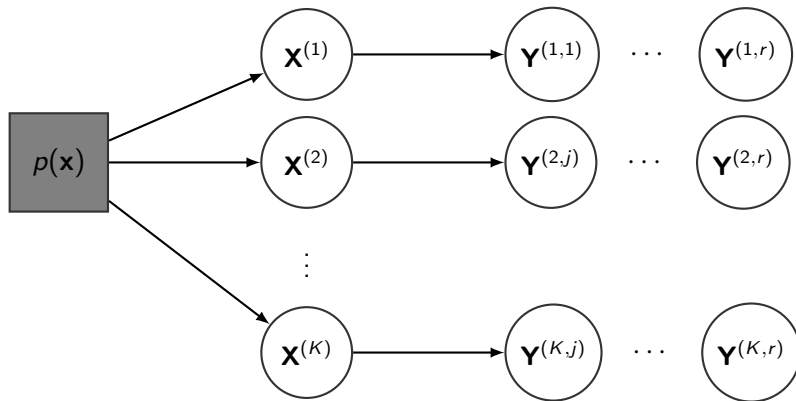
- Main question: How to estimate $I(\mathbf{X}; \mathbf{Y})$ from such data?
- Side question (experimental design): How to choose K?

Entropy and mutual information

X and Y have joint density $p(x, y)$ with respect to μ.

Quantity	Definition	Linear analogue
Entropy	$H(X) = -\int (\log p(x)) p(x) \mu_X(dx)$	$\text{Var}(X)$
Conditional entropy	$H(X\|Y) = \mathbf{E}[H(X\|Y)]$	$\mathbf{E}[\text{Var}(X\|Y)]$
Mutual information	$I(X; Y) = H(X) - H(X\|Y)$	$\text{Cor}^2(X, Y)$

The above definition includes both *differential* entropy and *discrete* entropy.
Information theorists tend to use log base 2, we will use natural logs in this talk.

Can we learn $I(\mathbf{X}; \mathbf{Y})$ from such data?

Answer: yes.

- We have $I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$.
- We can estimate $H(\mathbf{Y})$ from the data
- We can estimate $H(\mathbf{Y}|\mathbf{x}^{(i)})$ from the data, and define

$$\hat{H}(\mathbf{Y}|\mathbf{X}) = \frac{1}{K} \sum_{i=1}^{K} \hat{H}(\mathbf{Y}|\mathbf{X}^{(i)})$$

- As K and r both tend to infinity,

$$\hat{I}(\mathbf{X}; \mathbf{Y}) = \hat{H}(\mathbf{Y}) - \hat{H}(\mathbf{Y}|\mathbf{X})$$

  is consistent for $I(\mathbf{X}; \mathbf{Y})$.

Limitations with the 'naïve' approach

Naïve estimator:

$$\hat{I}(\mathbf{X}; \mathbf{Y}) = \hat{H}(\mathbf{Y}) - \frac{1}{K} \sum_{i=1}^{K} \hat{H}(\mathbf{Y}|\mathbf{X}^{(i)})$$

- If K is small, the naïve estimator may be quite biased, even for low-dimensional problems. Gastpar et al. (2010) introduced an *antropic correction* to deal with the small-K bias.
- Difficult to estimate differential entropies $H(\mathbf{Y})$, $H(\mathbf{Y}|\mathbf{x}^{(i)})$ in high dimensions. Best rates are $O(1/\sqrt{n})$ for $d \leq 3$ dimensions. Convergence rates for $d > 3$ unknown!

- Supervised learning becomes an extremely common approach for dealing with high-dimensional data, for numerous reasons!
- Perhaps we can use supervised learning to estimate $I(\mathbf{X}; \mathbf{Y})$ as well.
- Existing approach uses confusion matrix (Treves et al, 1997)

- Successful supervised learning exploits structure in the data, which *nonparametric methods ignore.*
- Using supervised learning to estimate mutual information can be viewed as *using prior information* to improve the estimate of $I(\mathbf{X}; \mathbf{Y})$.
- So while the general problem of information estimation is nearly impossible in high dimensions, the problem might become tractable if we can exploit known structure in the problem!

Relationship between mutual information and classification

- Suppose X and Y are discrete random variables, and X is uniformly distributed over its support.
- Classify X given Y. The optimal rule is to guess

$$\hat{X} = \text{argmax}_x \, p(Y|X = x).$$

- Bayes error:

$$p_e = \Pr[X \neq \hat{X}].$$

- Fano's inequality:

$$I(X; Y) \geq (1 - p_e) \ln K - \text{const}.$$

where K is the size of the support of X.

Define the *K*-class *average Bayes error* as

$$ABE_K = \mathbf{E}[\Pr[\hat{X}_{Bayes}(Y) \neq X]]$$

where $\hat{X}_{Bayes}$ is the optimal classification rule, and the expectation averages over randomness in $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$.

Then, under a certain asymptotic regime, there is an *exact* relationship between ABE_K and $I(\mathbf{X}; \mathbf{Y})$ (in contrast to the lower bound given by Fano's inequality or confusion matrix method.)

Theorem. *Take a sequence of joint distributions* $(\mathbf{X}, \mathbf{Y})$ *where the dimensionality of* $\mathbf{X}$ *and* $\mathbf{Y}$ *grow to infinity, but where*

$$\lim I(\mathbf{X}; \mathbf{Y}) = \iota < \infty.$$

Under certain regularity conditions, the limiting K-class average Bayes error is given by

$$\lim_{d \to \infty} ABE_K = \pi_K(\sqrt{2\iota}).$$

where

$$\pi_K(\mu) = 1 - \int_{-\infty}^{\infty} \phi(z - \mu)(1 - \Phi(z))^{K-1} dz.$$

Sidenote: interpretation of π_K

The function $\pi_K(\mu)$ gives the probability that a $N(\mu, 1)$ variable is smaller than the minimum of $K - 1$ other $N(0, 1)$ variables (all independent.) Hence $\pi_K(0) = \frac{K-1}{K}$ due to symmetry. (This is also the misclassification rate from pure guessing.)
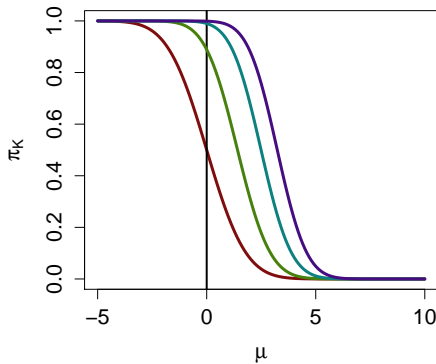
Legend: $K = \{$ 2 , 9 , 99 , 999 $\}$

Regularity Conditions

- We require the conditional log-likelihoods

$$\log p(\mathbf{Y}^{(i)}|\mathbf{X}^{(j)})$$

to have a nondegenerate jointly multivariate normal limiting distribution (when scaled appropriately).

- Define

$$u(x, y) = \frac{p(x, y)}{p(x)p(y)} - 1.$$

Draw $\mathbf{X} \sim p(x)$ and $\mathbf{Y}^* \sim p(y)$ independently. We require $u(\mathbf{X}, \mathbf{Y}^*)$ have a nondegenerate univariate normal limiting distribution (when scaled appropriately). Also, $\mathrm{Cov}[u(\mathbf{X}^{(i)}, \mathbf{Y}^{(j)}), u(\mathbf{X}^{(k)}, \mathbf{Y}^{(j)})] \to 0$.

These conditions are satisfied quite generally, e.g. if $(\mathbf{X}, \mathbf{Y})$ multivariate normal with limiting spectrum.

The low-SNR estimator of $I(\mathbf{X}; \mathbf{Y})$

Our proposed estimator for mutual information is

$$\hat{I}_{ls}(\mathbf{X}; \mathbf{Y}) = \frac{1}{2}\pi_K^{-1}(\widehat{ABE})^2$$

where $\widehat{ABE}$ is the test error of the classifier. (The subscript ls stands for low-SNR.)

Note: there is still a lot of improvement for this estimator, mainly correcting for bias due to variability in $\widehat{ABE}$. This is the first version we tried.

Simulation study

Models.

- Multiple-response logistic regression model

$$X \sim N(0, I_p)$$

$$Y \in \{0, 1\}^q$$

$$Y_i | X = x \sim \text{Bernoulli}(x^T B_i)$$

where B is a $p \times q$ matrix.

Methods.

- Nonparametric: $\hat{I}_0$ naive estimator, $\hat{I}_\alpha$ anthropic correction.
- ML-based: $\hat{I}_{CM}$ confusion matrix, $\hat{I}_F$ Fano, $\hat{I}_{LS}$ low-SNR method.

Fig 1. Low-dimensional results ($q = 3$)

Sampling distribution of $\hat{I}$ for $\{p = 3,\ B = \frac{4}{\sqrt{3}}I_3,\ K = 20,\ r = 40\}$.
True parameter $I(X; Y) = 0.800$ *(dotted line.)*
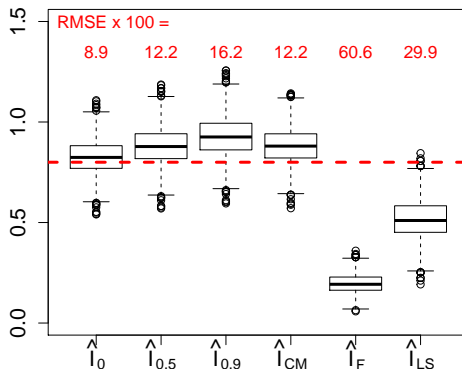
Naïve estimator performs best! $\hat{I}_{LS}$ not effective.

Fig 2. High-dimensional results ($q = 50$)

Sampling distribution of $\hat{I}$ for $\{p = 50, B = \frac{4}{\sqrt{50}} I_{50}, K = 20, r = 8000\}$.
True parameter $I(X; Y) = 1.794$ (dashed line.)
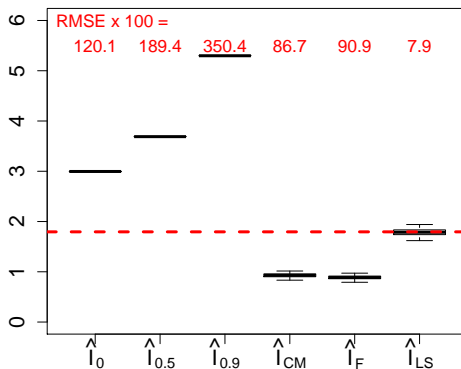
Non-parametric methods extremely biased.

Fig 5. Dependence on K given fixed N ($q = 10$)

Sampling distribution of $\hat{I}_{LS}$ for $\{p = 10,\ B = \frac{4}{\sqrt{10}} I_{10},\ N = 80000\}$, and $K = \{5, 10, 15, 20, \ldots, 80\}$, $r = N/k$.
True parameter $I(X; Y) = 1.322$ (dashed line.)
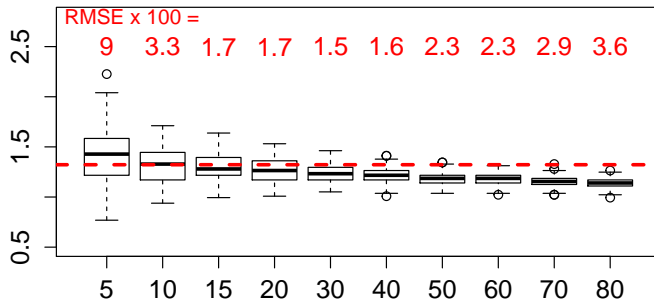
Decreasing variance as K increases. Bias at large and small K.

Conclusions

- We derive a relationship between average Bayes error (ABE) and mutual information (MI), motivating a novel estimator $\hat{I}_{LS}$.
- Theory based on high dimensional, low SNR limit, where

$$\text{ABE} \leftrightarrow \text{MI}.$$

- In ideal settings for supervised learning, ABE can be estimated effectively and $\hat{I}_{LS}$ can recover MI at much lower sample sizes than nonparametric methods.
- In simulations, $\hat{I}_{LS}$ works better than Fano's inequality or the confusion matrix approach.

The geometry of human perception: RSA and multivariate models

Charles Zheng

Stanford University

November 16, 2015

(Joint work with Yuval Benjamini and Oluwasanmi Koyejo)

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.
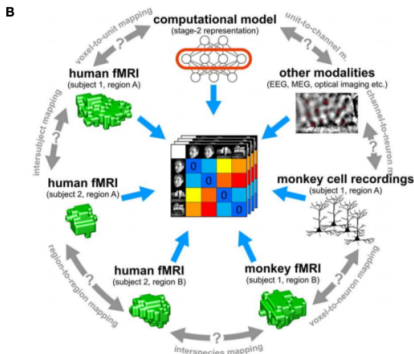
A typical RSA experiment

An experiment which demonstrates which regions of the brain differentiate between faces and objects.
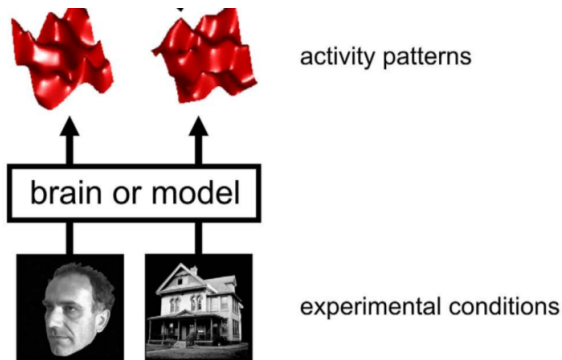
Step 1: Present the subject with visual stimuli, pictures of faces and houses. Record the subject's brain activity in the fMRI scanner.

A typical RSA experiment

Step 2a: Process the data, and represent the brain activity of the subject for the ith stimulus as a real vector y_i. Form matrix of distances between y_i and y_j, the *representation distance matrix* (RDM).
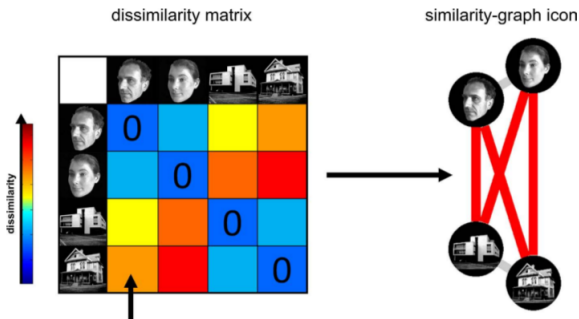
Step 2b: Assess statistical significance of distances to form similarity graph.

A typical RSA experiment

Step 3: Compare similarity graphs between different brain regions.
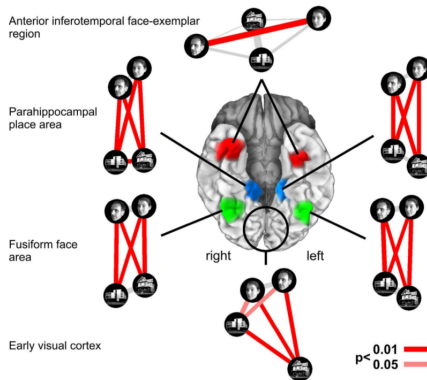
Step 4: Draw scientific conclusions. (Step 5: Profit!!..?)

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.

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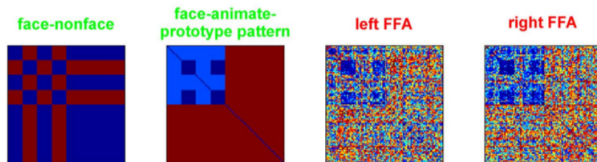
A rejection of the independence null between D^A and D^B is taken to mean that outputs A and B are 'related.' For instance, D^A might be the response from a subject's brain region, and D^B is a 0-1 matrix reflecting *a priori* class membership. Rejection is taken to mean that the region A have differential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

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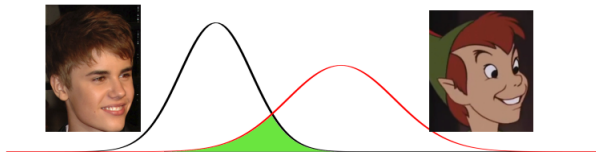
Note that the definition not only depends on the difference in *means* but also depends on the *noise distribution*.

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.
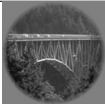
- Example: if y is conditionally multivariate normal, with covariance Σ not depending on x, then

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

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

Functional MRI allows us to infer the metric

- In fMRI, one records the subject's response y_i to a stimulus x_i, for $i = 1, \ldots, n$ (e.g. $n = 3000$)
- The recorded y_i is actually a 'filtered' version of the brain activity (some information is lost)
- By fitting a model to the data, one can estimate the conditional distribution of $Y|x$ and hence estimate the perception-induced metric d_ι.

Stimuli	Response

Parametric RSA

- Combine the parametric approach of multivariate regression with RSA.
- Model:

$$y \sim N(B^T x, \Omega^{-1}).$$

- The distribution-induced metric is therefore

$$D(x_i, x_j) = (x_i - x_j)^T B \Omega B^T (x_i - x_j).$$

Parametric RSA

$$y \sim N(B^T x, \Omega^{-1})$$

$$D(x_i, x_j) = (x_i - x_j)^T B \Omega B^T (x_i - x_j)$$

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

$$D^A = D^B$$

we can test

$$\Sigma^A = \Sigma^B.$$

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

FFA

F2

F1

$\Sigma_{FFA} = \begin{pmatrix} 3 & 1 \\ & 2 \end{pmatrix}$
$\phantom{\Sigma_{FFA} =} H \; F1 \; F2$

information loss

H

F2

F1

V4

$\Sigma_{V4} = \begin{pmatrix} 3 & 3 \\ & 3 \end{pmatrix}$
$\phantom{\Sigma_{V4} =} H \; F1 \; F2$

H

F2

F1

Stimulus

$\Sigma_{X} = \begin{pmatrix} 5 & 5 \\ & 5 \end{pmatrix}$
$\phantom{\Sigma_{X} =} H \; F1 \; F2$

selective transmission

$\Sigma_{AFER} = \begin{pmatrix} 1 & 3 \\ & 3 \end{pmatrix}$
$\phantom{\Sigma_{AFER} =} H \; F1 \; F2$

AFER
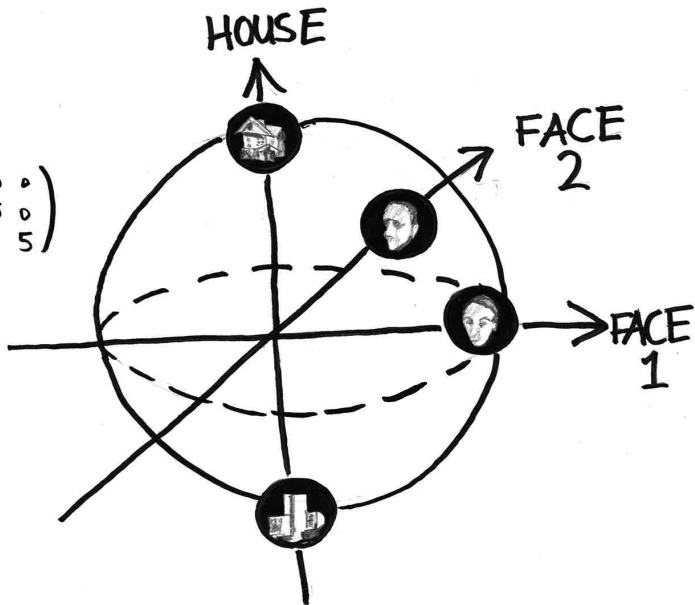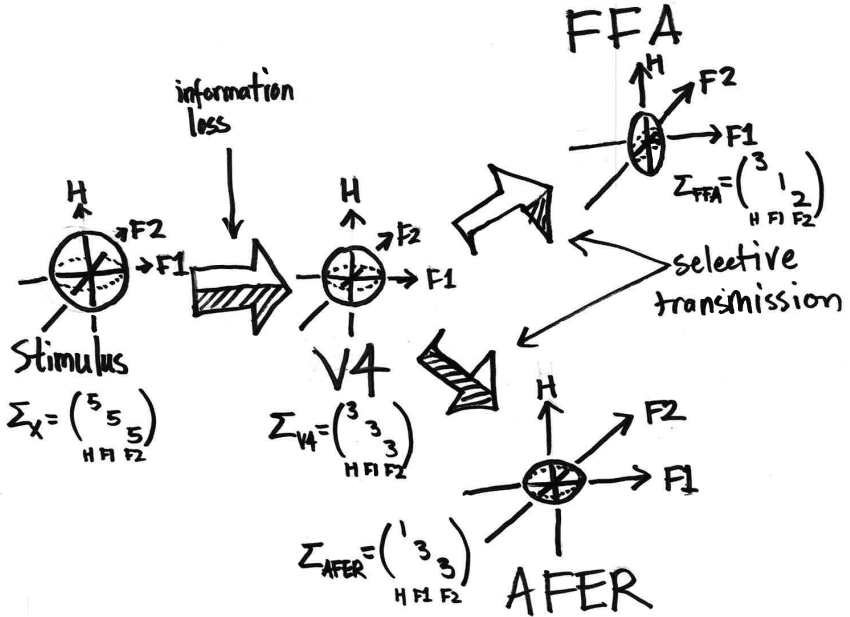
Comparison with information approach

- Parametric RSA makes strong assumptions on the model (e.g. linearity.)
- We can track information loss: $I(X; Y^{V1}) > I(X; Y^{V2})$
- But in cases of *selective transmission*, we can infer *which dimensions of information* are being lost (face features, house features, etc.)
- We get a finer-grained picture of information flow than the Shannon information approach.

Section 4

To do

Ongoing

- Apply low-SNR information estimator and parametric RSA to real data.
- Develop bootstrap-based hypothesis tests for parametric RSA: e.g. $H_0 : \Sigma^A = \Sigma^B$.

Possible extensions of information theory work

- Our proposed estimator performs well in small sample sizes, yet is *not consistent*... can it be fixed?
- Develop complimentary theory for estimating Bayes error. (This would allow us to state risk bounds for estimating $I(X;Y)$.)
- Even if you don't care about Shannon information, the theory allows you to potentially extrapolate classification curves:

$$ABE_{10} \to I(X;Y) \to ABE_{10000}$$

  (But more work is needed.)
- Use our theory to address the question of optimal experimental design.

Possible extensions of RSA work

Beyond testing $\Sigma^A = \Sigma^B$

- Inference on which components of Σ^A differ from Σ^B.
- Multiple comparisons: comparing Σ^A to Σ^B to Σ^C
- Inferring latent features X from data! *Multiple MDS*

$$\min_{x, \Sigma^k} \sum_{i,j,k} w_{ijk}(D_{ij}^k - (x_i - x_j)^T \Sigma^k (x_i - x_j)).$$

FFA

information loss

H ↗F2 →F1

$\Sigma_{FFA} = \begin{pmatrix} 3 & 1 \\ & 2 \end{pmatrix}$
$\quad$ H F1 F2

H ↗F2 →F1

Stimulus

$\Sigma_X = \begin{pmatrix} 5 & 5 \\ & 5 \end{pmatrix}$
$\quad$ H F1 F2

H ↗F2 →F1

V4

$\Sigma_{V4} = \begin{pmatrix} 3 & 3 \\ & 3 \end{pmatrix}$
$\quad$ H F1 F2

selective transmission

H ↗F2 →F1

$\Sigma_{AFER} = \begin{pmatrix} 1 & 3 \\ & 3 \end{pmatrix}$
$\quad$ H F1 F2

AFER

References
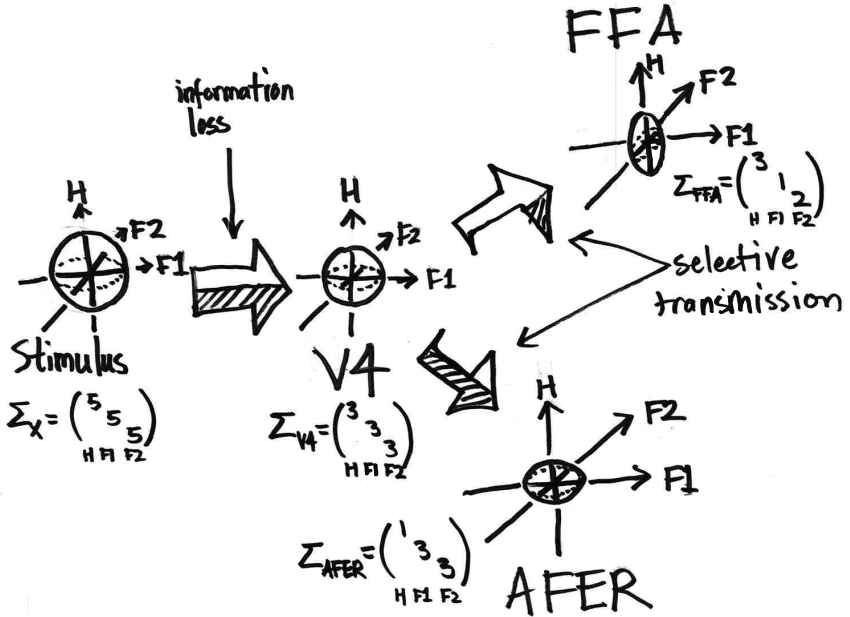
- Kay, KN., Naselaris, T., Prenger, R. J., and Gallant, J. L. "Identifying natural images from human brain activity". *Nature* (2008)
- Naselaris, et al. "Bayesian reconstruction of natural images from human brain activity". *Neuron* (2009)
- Vu, V. Q., Ravikumar, P., Naselaris, T., Kay, K. N., and Yu, B. "Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models", *The Annals of Applied Statistics*. (2011)
- Chen, M., Han, J,. Hu, X., Jiang, Xi., Guo, L. and Liu, T. "Survey of encoding and decoding of visual stimulus via fMRI: an image analysis perspective." *Brain Imaging and Behavior*. (2014)

References

- Cover and Thomas. Elements of information theory.

- Gastpar., Gill., Huth., Theunnisen. "Anthropic Correction of Information Estimates and Its Application to Neural Coding." IEEE Trans I.T.

- Quiroga and Panzeri. "Extracting information from neuronal populations: information theory and decoding approaches." Nature Rev. Neuro.

- Muirhead. Aspects of multivariate statistical theory.

- van der Vaart. Asymptotic statistics.

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