Stability

Bin Yu
Statistics and EECS, University of California-Berkeley

SAMSI Opening Workshop on Massive Data, Sept, 2012

In honor of

John W. Tukey
June 16, 1915 – July 26, 2000

1962: “Future of Data Analysis” by Tukey

It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: intellectual adventure, demanding calls upon insight, and a need to find out "how things really are" by investigation and the confrontation of insights with experience.

2012: Information technology and data

"Although scientists have always comforted themselves with the thought that science is self-correcting, the immediacy and rapidity with which knowledge disseminates today means that incorrect information can have a profound impact before any corrective process can take place".

"Reforming Science: Methodological and Cultural Reforms," *Infection and Immunity* Editorial (2012) by Arturo Casadevall (Editor in Chief, mBio) and Ferric C. Fang, (Editor in Chief, Infection and Immunity)

2012: IT and data

"A recent study* analyzed the cause of retraction for 788 retracted papers and found that error and fraud were responsible for 545 (69%) and 197 (25%) cases, respectively, while the cause was unknown in 46 (5.8%) cases (31)." -- Casadevall and Fang (2012)

* R Grant Steen (2011), J. Med. Ethics

Casadevall and Fang called for

"Enhanced training in probability and statistics"

Scientific reproducibility

Statistical stability is a minimum requirement

for scientific reproducibility

Statistical stability: statistical conclusions should be stable to appropriate perturbations to data and/or models.

Roadmap for today

- Intellectual adventure: understanding of human visual pathway via fMRI
- Encoding and decoding models for natural-stimuli-fMRI data

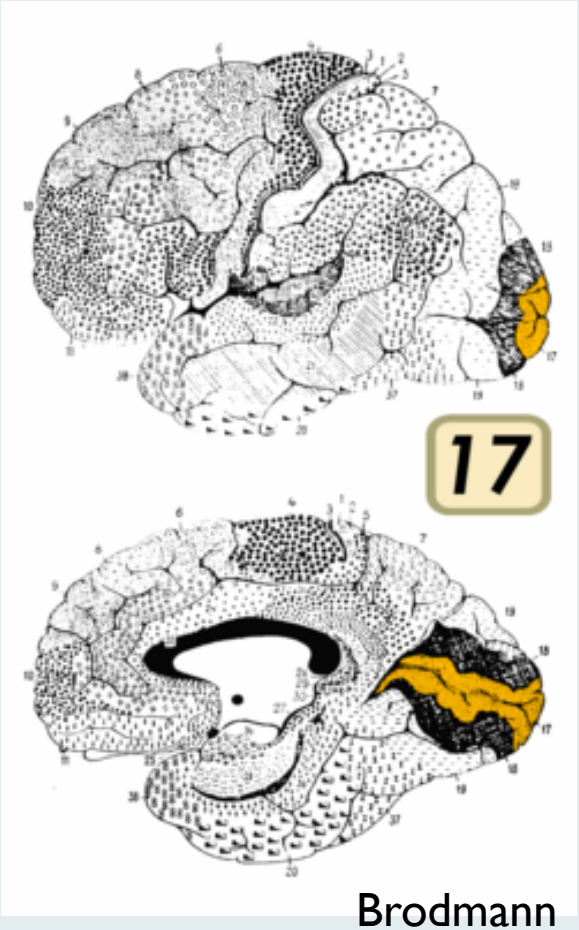
- Finding out "How things really are": interpretation needs stability
- Stability in the literature

- Proposed ES-CV for Lasso
- ES-CV applied to fMRI project

- Some theory: sample stability meets robust statistics in high-dim

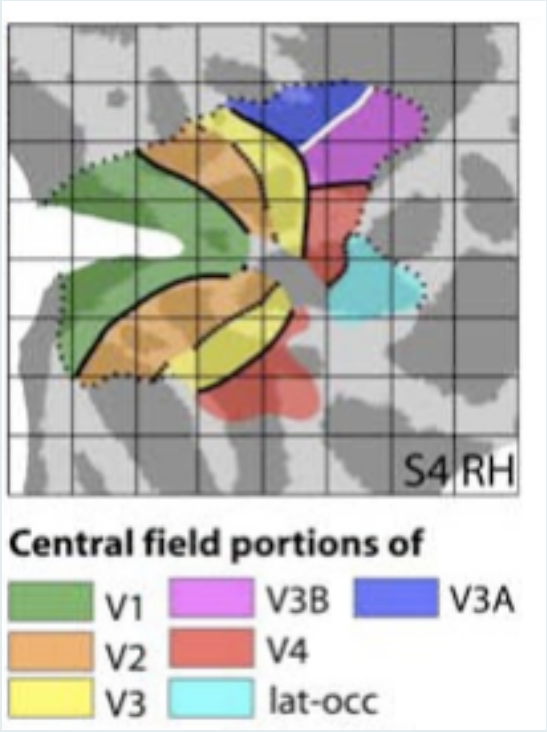
I. Intellectual adventure: Understanding visual system in Gallant Lab

- Retina (eye) ➔ LGN ➔ Early visual cortex (V1) ➔ V2 ➔ V3,V4…



Brodmann

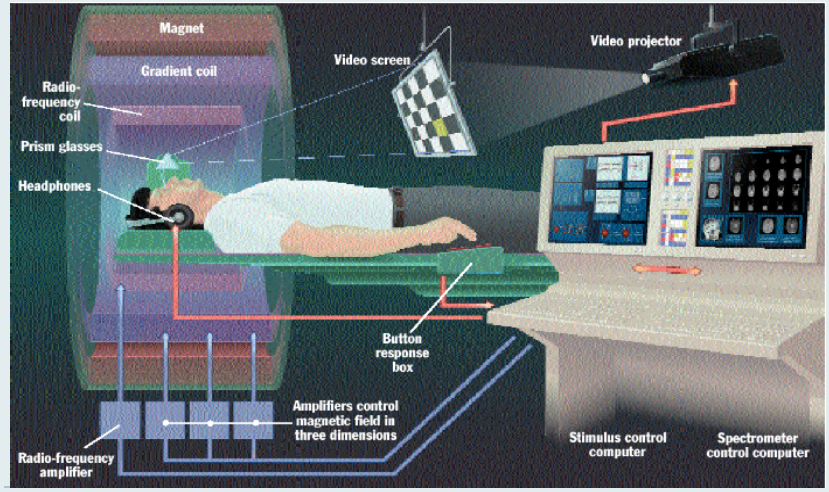
V1



Ringach, 2002

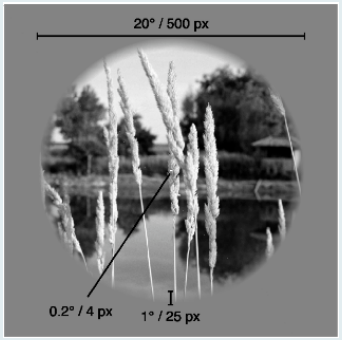
Functional MRI

- Non invasive and indirect recording technique
- Measures oxygenated blood flow – correlate of neural activity
- Low temporal resolution, a few seconds
- High spatial resolution (voxels of 1x1x1 mm cubes)
  - > 10,000 voxels in early visual areas (V1, V2 and V3)
  - each voxel covers > 100,000 neurons
- Can watch videos inside the machine



Goal: interpretable model

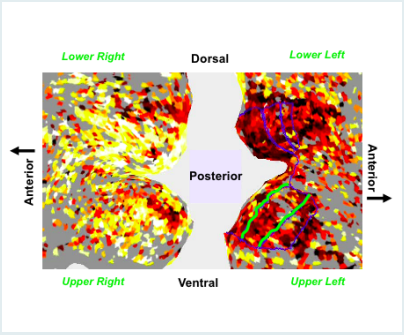
- Quantitative model - both stimulus and response high-dimensional



Natural Input
(Image or video/movie)

Encoding

Decoding

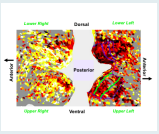


fMRI of Brain

  - Two related tasks:
  - Encoding predicts brain signal from visual stimuli
  - Decoding "validates" encoding and important in its own right

(cf. Dayan and Abbott, 2001)

Movie reconstruction:



Nishimoto, Vu, Naselaris, Benjamini, Yu and Gallant (2011)

Presented clip

Clip reconstructed from brain activity

“Mind-Reading Computers” in media

one of the 50 best inventions of 2011 by Time Magazine

Others: Economist, NPR, …

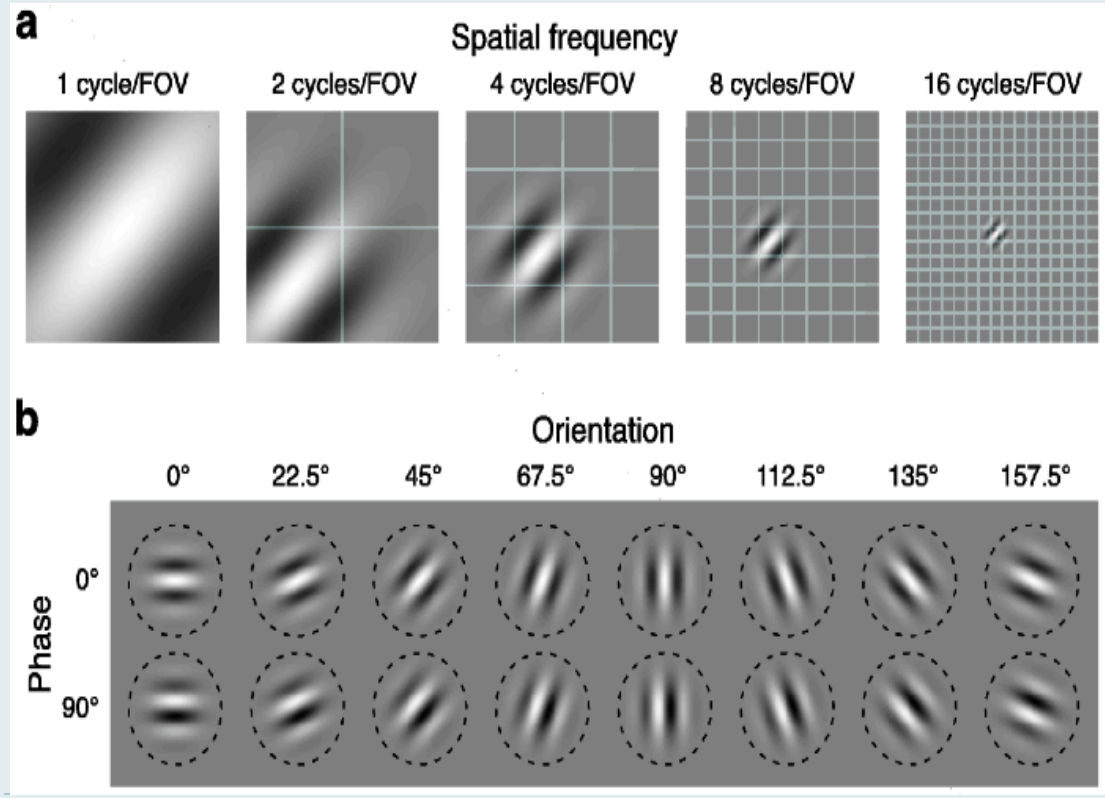
What model is behind the movie reconstruction algorithm?

Is the model interpretable?

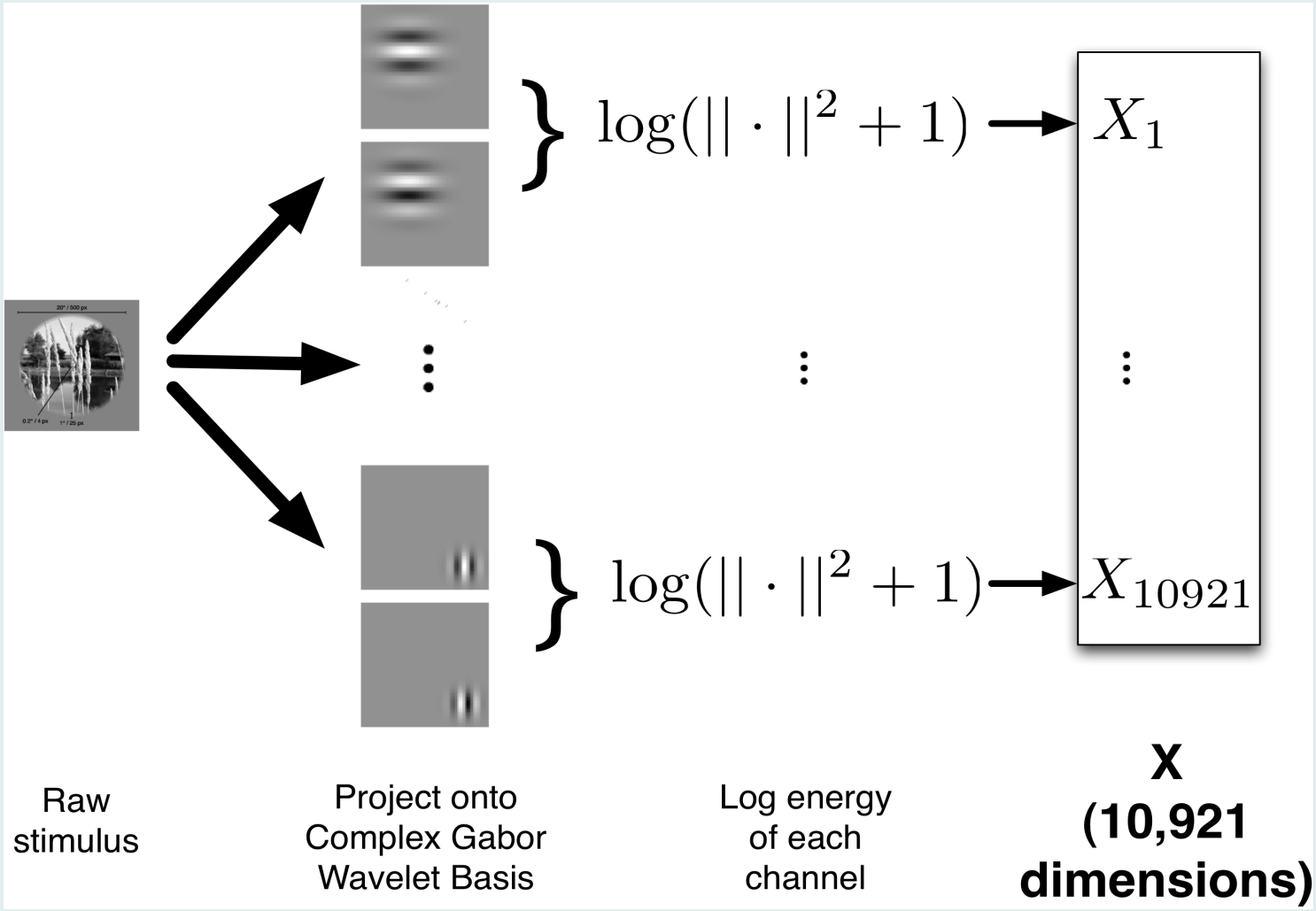
Early Visual Area V1

- Preprocessing an image:
  - Gabor filters corresponding to particular spatial frequencies, locations, orientations (Hubel and Wiesel, 1959,…)

Sparse representation after Gabor Filters, static or dynamic

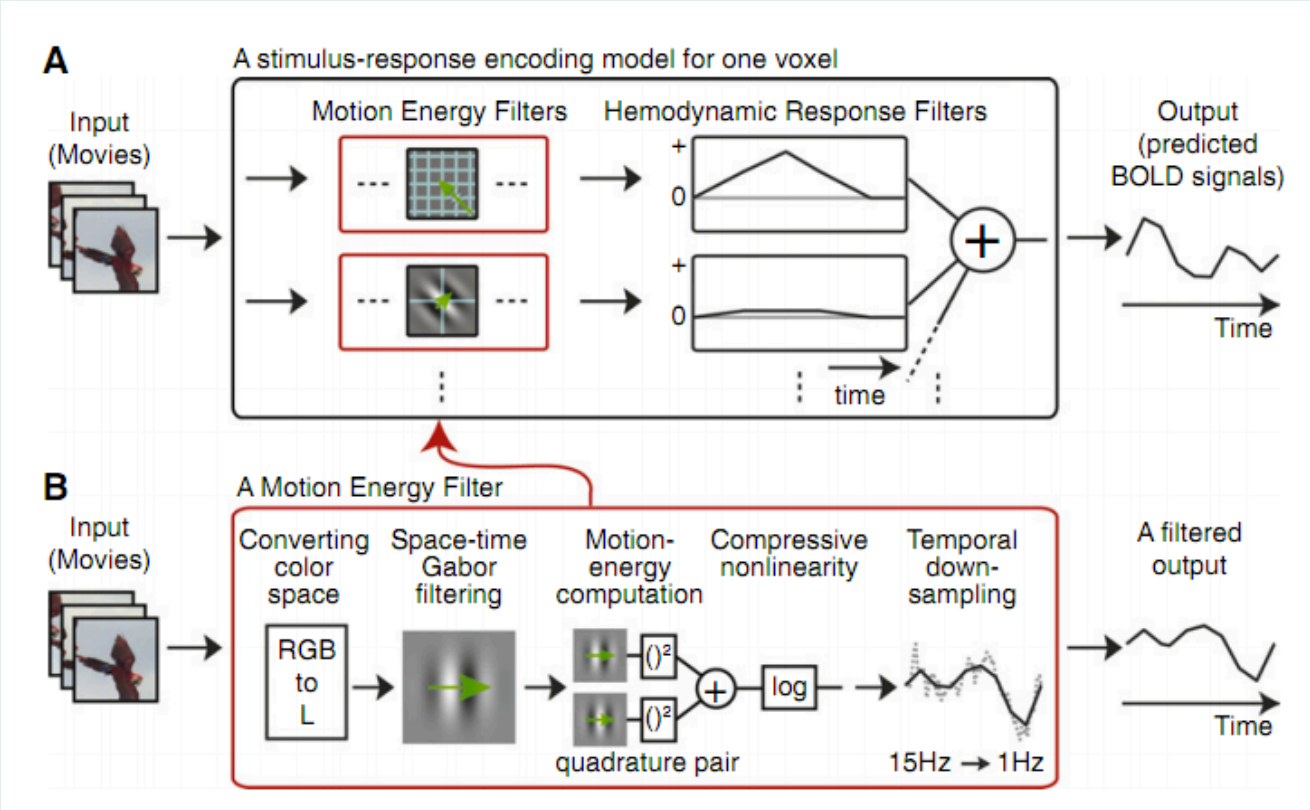


Features


$\log(|| \cdot ||^2 + 1)$
$\log(|| \cdot ||^2 + 1)$
X_1
X_{10921}
Raw stimulus
Project onto Complex Gabor Wavelet Basis
Log energy of each channel
X (10,921 dimensions)

From images to videos

- Features move (use 3D wavelets)
- Data split to 1 second chunks
- A lot more fitting going on



Encoding models in Gallant Lab

- Separate sparse linear model fitted to features for each voxel
  via e-L2Boost (Friedman, 2001) or Lasso (Tibshirani, 1996)

- Fitted model tested on 120 validation samples
  Performance measured by prediction **correlation**

Lasso: L_1-norm as a penalty to L2 loss

- The L_1 penalty is defined for coefficients β

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$$

- Used initially with L_2 loss:

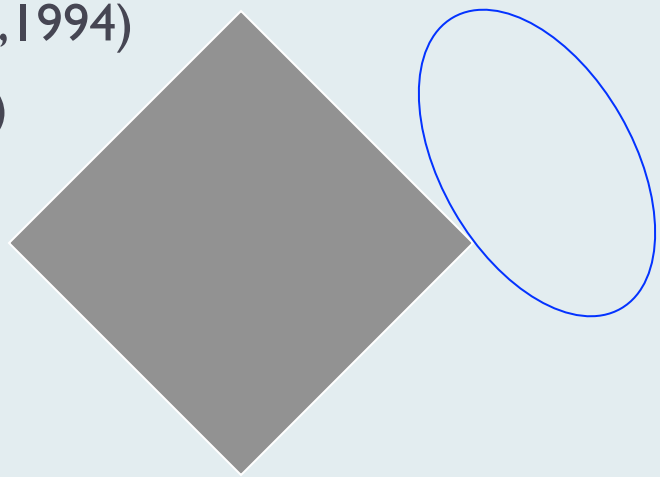
One particular regularization method (Tikhonov, 1969; …)

Signal processing: Basis Pursuit (Chen & Donoho,1994)

Statistics: Non-Negative Garrote (Breiman, 1995)

Statistics: LASSO (Tibshirani, 1996)

$$\hat{\beta}(\lambda) = argmin_{\beta}\{||Y - X\beta||_2^2 + \lambda||\beta||_1\}$$



Smoothing parameter often selected by Cross-Validation (CV)

Reconstruction of movies

(Nishimoto, Vu, Naselaris, Benjamini, Yu, and Gallant (2011)

1. Download N videos/images from the internet to construct an empirical prior (e.g. movie trailers, youtube videos, …)

2. Using fitted encoding models, we can get their posterior, and calculate posterior means.

3. Given an fMRI vector Y , for some image feature $g(Im)$, we can calculate on its expected mean using the posterior weights:

$$E[g(Im)|Y] =$$

$$\sum_{j \leq M} g(Im_j) \exp\{(Y - \hat{\beta}^T f(Im_j))^T \hat{\Sigma}_\omega^{-1} (Y - \hat{\beta}^T f(Im_j))\}$$

g's represent **red, green, blue** values at each pixel, frame.
Recover color based on the prior, not part of the model.

Reconstruction algorithm

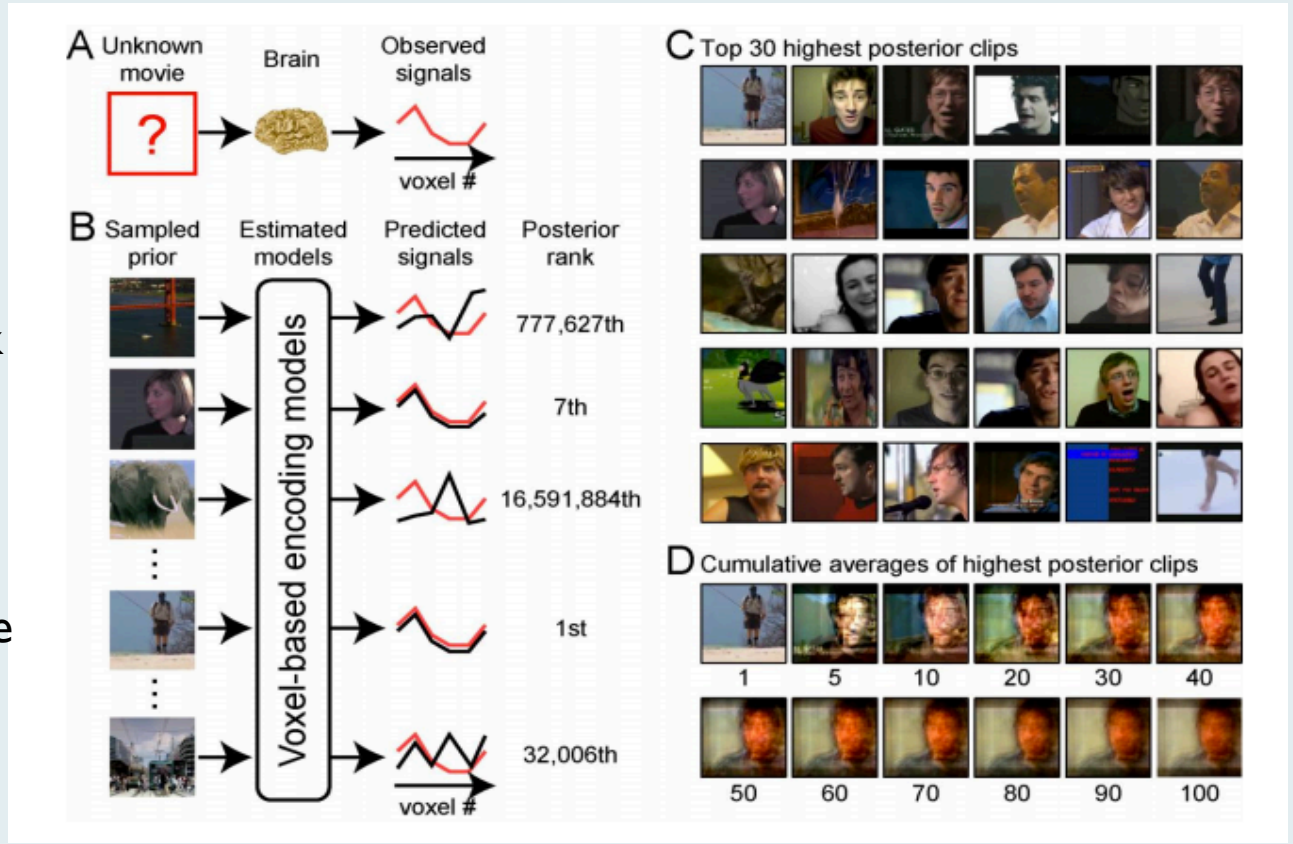
A. Record data

B. Fit encoding models

C. Sample prior and rank fit based on posterior

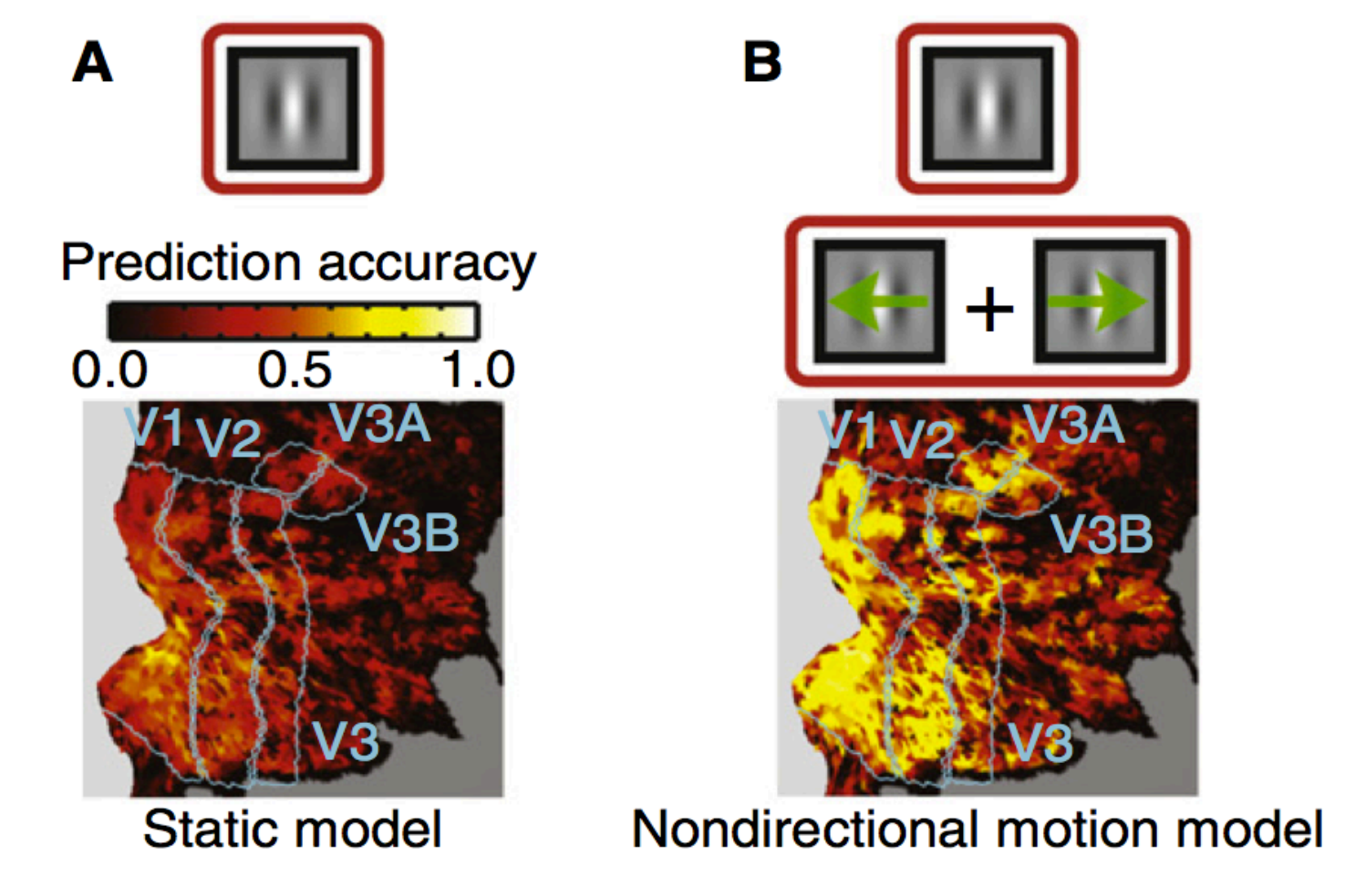
D. Pick top 100 ranks

E. Average in color space



Averaging in color-scale is quite rough

Encoding: energy-motion model necessary

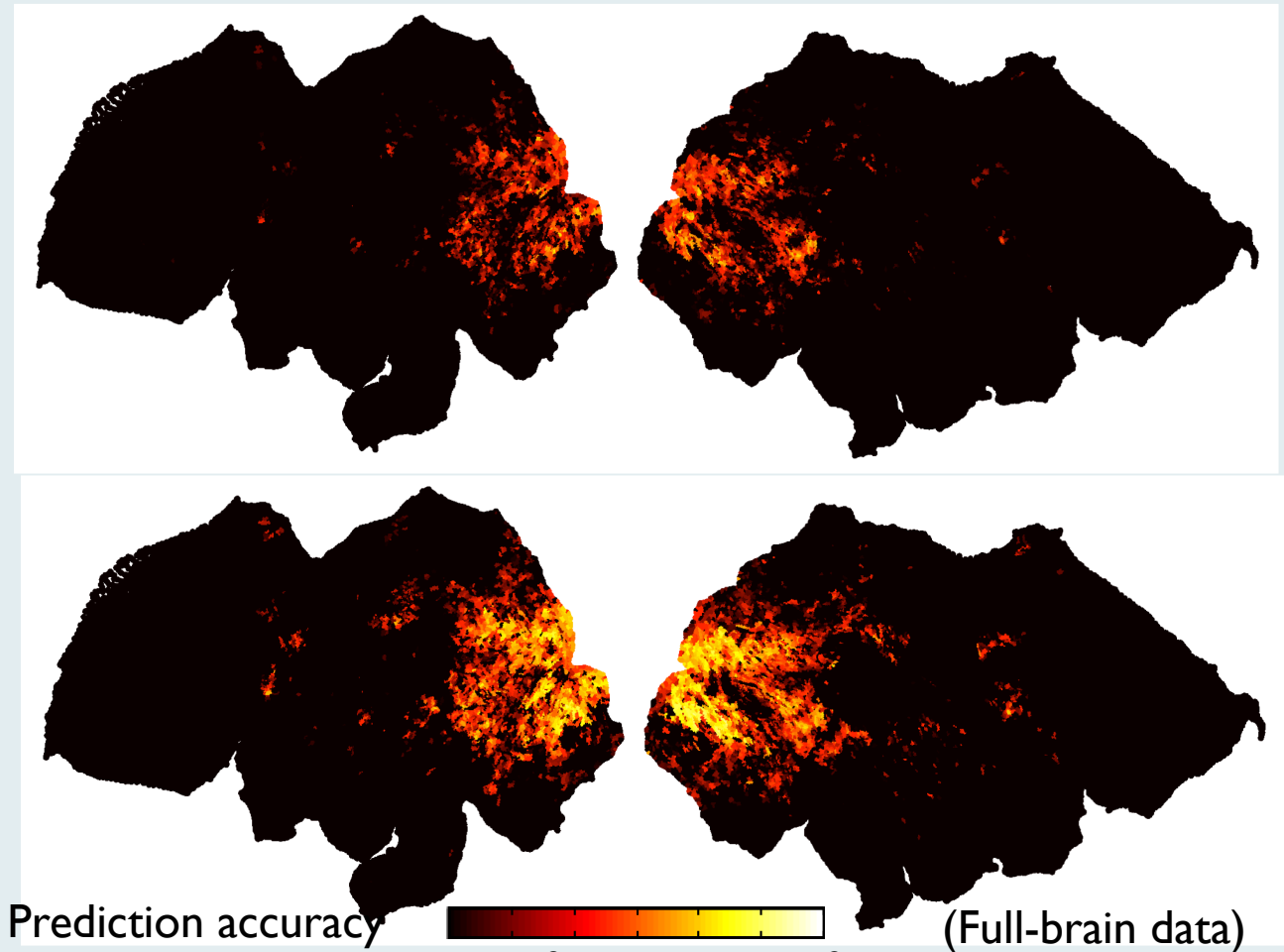


Encoding: sparsity necessary

Sparse regression improves prediction over OLS

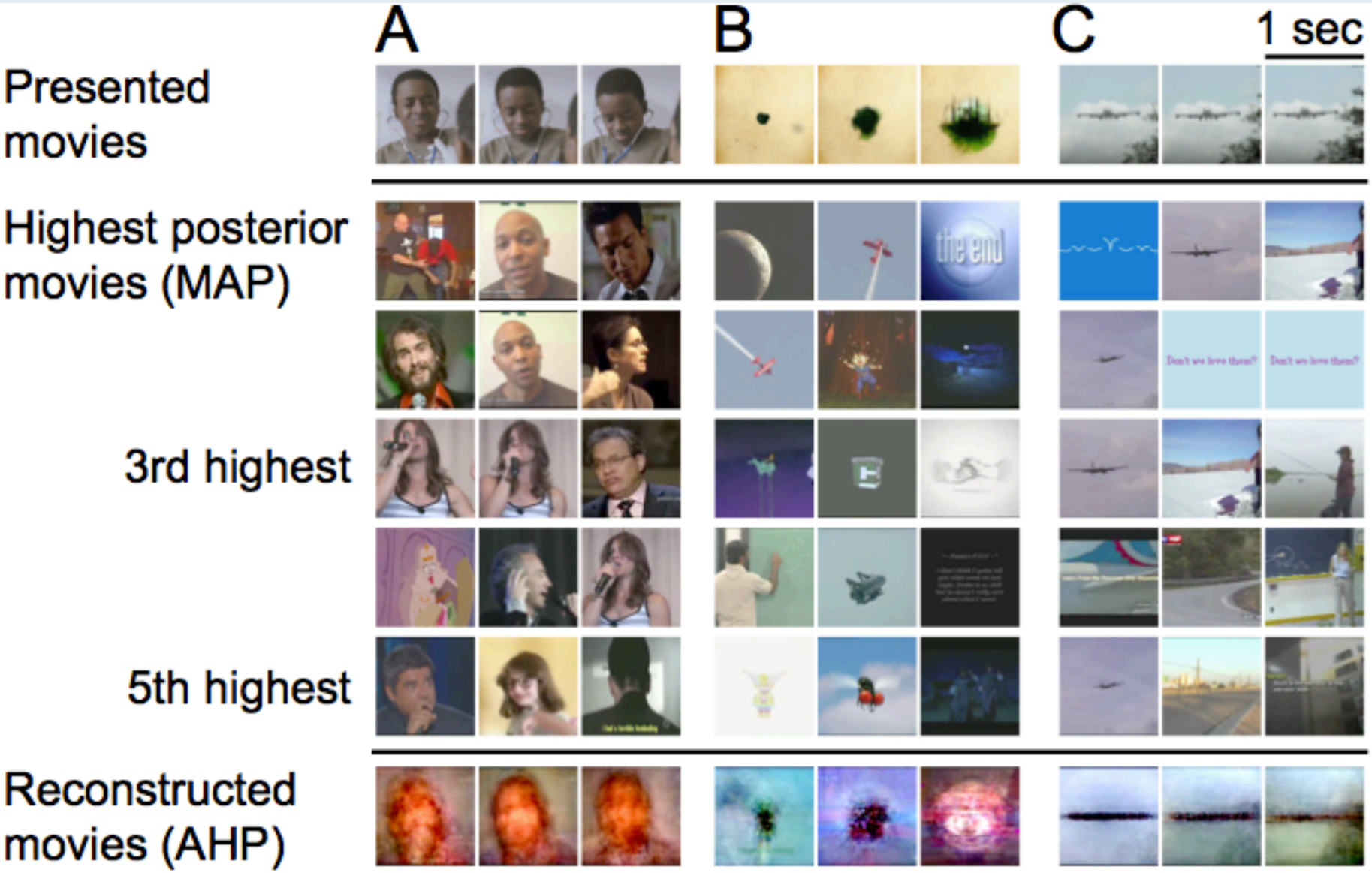
Linear Regression

Sparse Regression

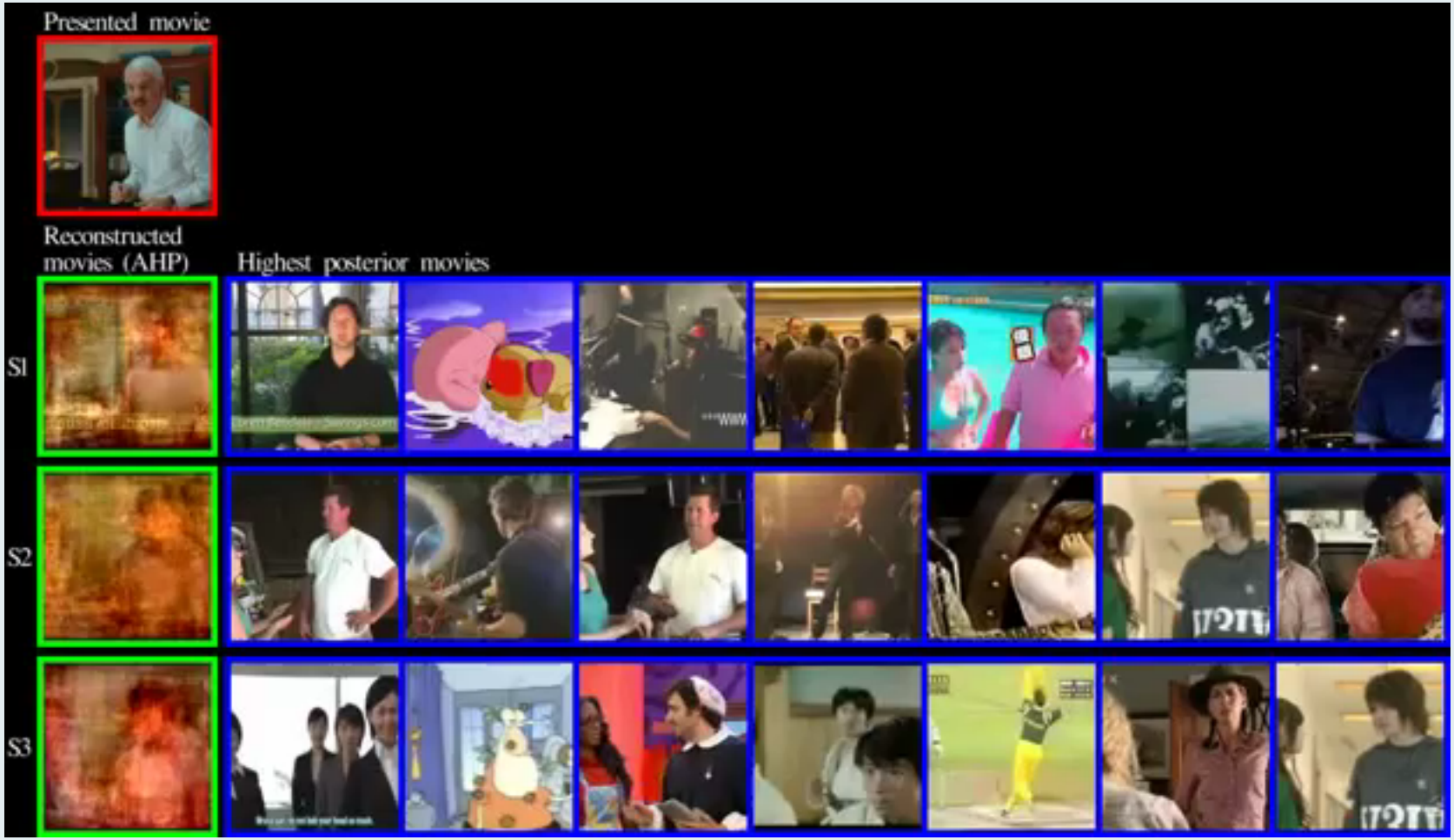


(Full-brain da

Reconstructing visual experience



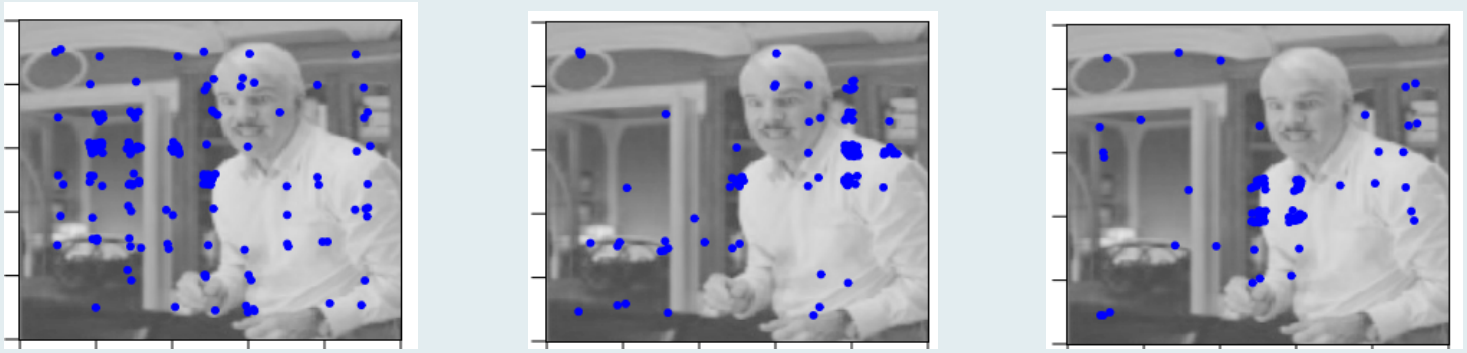
Movie reconstruction results for 3 subjects



Interpreting encoding models: spatial locations of selected features

Voxel A | Voxel B | Voxel C

Lasso +CV



Prediction scores on Voxels A-C are 0.72 (CV)

II. Stability for scientific reproducibilty

- Stability is a minimum requirement for reproducibility
- Statistical stability: statistical conclusions should be robust to appropriate perturbations to data.

Statistical stability is well defined relative to a particular aim and a particular perturbation to data (or model). Aims could be estimation, prediction or limiting law.

Data perturbation has a long history

Jacknife

… Quenouille (1949, 1956), Tukey (1958), Mosteller and Tukey (1977), Efron and Stein (1981), Wu (1986), Carlstein (1986), Künsch (1989), Wu (1990),..

Sub-sampling:

… Mahalanobis (1949), Hartigan (1969, 1970), Politis and Romano (1992 , 94), …

Cross-validation:

… Allen (1974), Stone (1974, 1977), Hall (1983), Li (1985), Zhang (1991), …

Bootstrap:

… Efron (1979), Bickel and Freedman (1981), Beran (1984), Künsch (1989), …

Perturbation argument is central to limiting law results

One proof of CLT (bedrock of classical stat theory):

1. Universality of a limiting law for a normalized sum of iid through a perturbation argument or Linderberg's swapping trick, i.e., a perturbation in a (normalized) sum by a random variable with matching first and second moments does not change the (normalized) sum distribution.
2. Finding the limit law via ODE.

cf. Lecture notes of Terence Tao at

http://terrytao.wordpress.com/2010/01/05/254a-notes-2-the-central-limit-theorem/

Limiting laws are stability results.

Perturbation argument is central (continue)

Recent generalizations to obtain other universal limiting distributions

e.g. Wigner law under non-Gaussian assumptions and last passage percolation...

Concentration results also assume stability-type conditions…

In learning theory, stability is closely related to good generalization performance…

Bringing stability considerations to Lasso

Precursors:

Shao (1995): perturbation via bootstrap for model selection criteria

Breiman (1996): perturbation by adding noise to responses

Stability and Lasso:

Bach (2008): perturbation via bootstrap for Lasso

Meinshausen and Bühlmann (08 or 10): perturbation via randomized Lasso (not selecting a particular smoothing parameter)

ES-CV: Statistical Stability (ES) (Lim & Yu, 2012)

Given a smoothing parameter λ, divide the data units into M blocks.

Get Lasso estimate $\hat{\beta}_m(\lambda)$ for each block m = 1, …, M, and form an estimate $X\hat{\beta}_m(\lambda)$ for the mean regression function.

$$\bar{\hat{\beta}}(\lambda) = \frac{1}{M}\sum_m \hat{\beta}_m(\lambda)$$

Define the estimation stability (ES) measure as

$$ES(\lambda) = \frac{\frac{1}{M}\sum_m ||X\hat{\beta}_m(\lambda) - X\bar{\hat{\beta}}(\lambda)||^2}{||X\bar{\hat{\beta}}(\lambda)||^2}$$

ES-CV: Statistical Stability (SS)+CV (continue)

ES aims at estimation stability, while CV aims at prediction stability.

ES is the reciprocal of a test statistic for testing

$$H_0 : X\beta = 0$$

with the variance of the mean function estimator estimated by delete-d Jacknife.

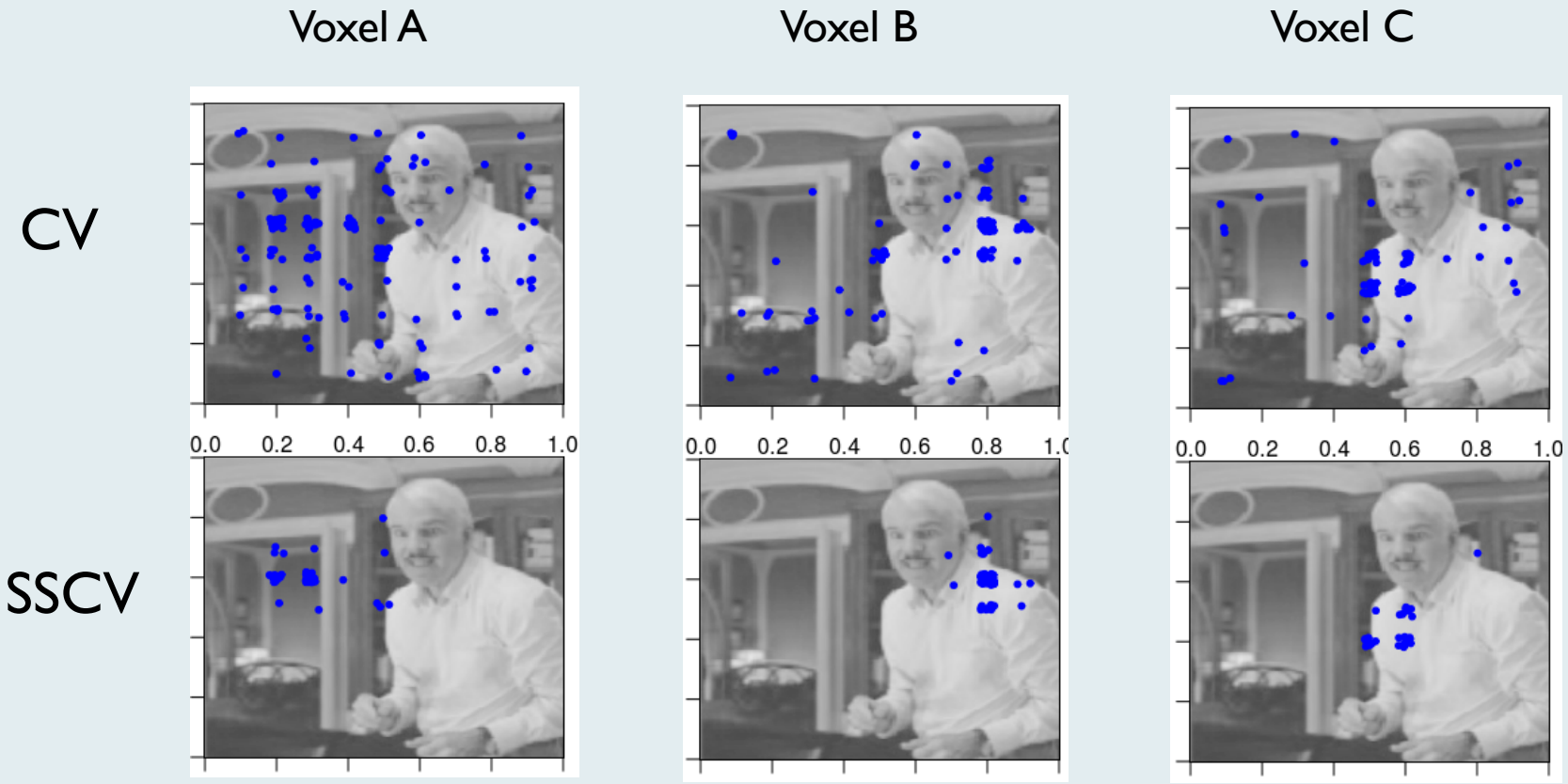
ES-CV: Statistical Stability (SS)+CV (continue)

ES-CV selection criterion for smoothing parameter λ :

Choose the λ that minimizes $ES(\lambda)$ and is not smaller that the CV selection.

Applicable to smoothing parameter selection in Lasso and Tikhonov regularization or Ridge and possibly more.

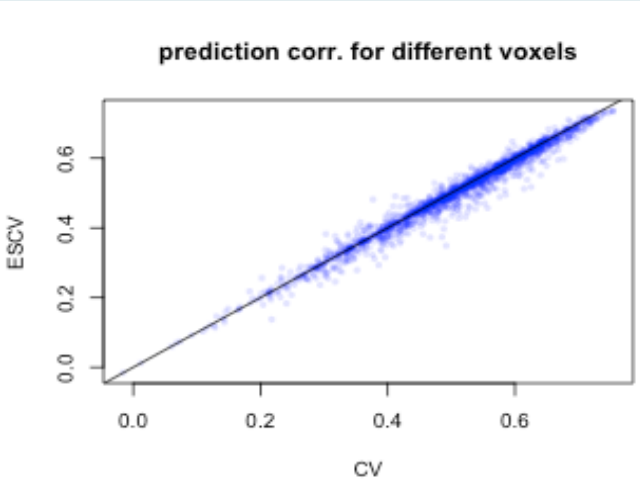
Back to fMRI prblem: spatial locations of selected features

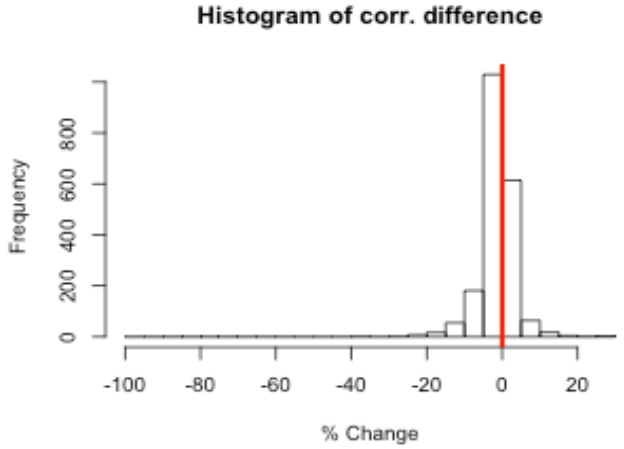


Prediction on Voxels A-C: CV 0.72, SSCV 0.7

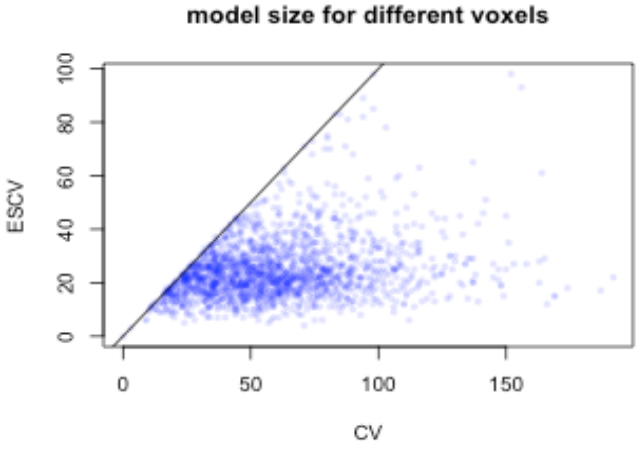
ESCV: sparsity gain (60% reduction) with minimal prediction loss (1.3%)

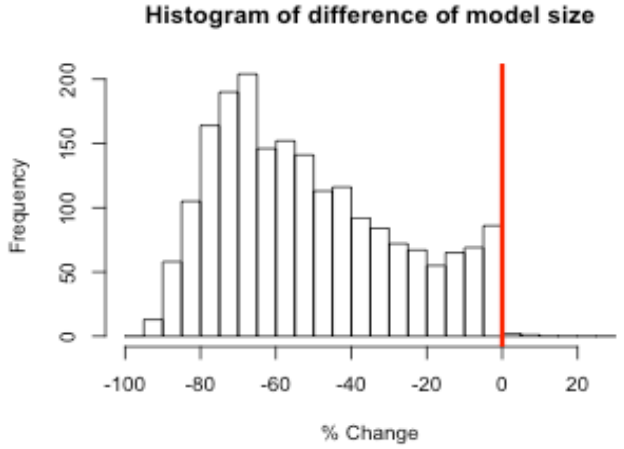
Prediction (correlation)





Model size





Based on validation data for 2088 voxels

III. Robust statistics also aims at stability

Mean functions are fitted with L_2 loss.

What if the "errors" have heavier tails than Gaussian tails?

L_1 loss is commonly used in robust statistics to deal with heavier tail errors in regression.

Will L_1 loss add more stability?

Model perturbation is used in Robust Statistics

Tukey (1958), Huber (1964, …), Hampel (1968, …),
Bickel (1976, …), Rousseeuw (1979, …), Portnoy (1979, ….),
….

"Overall, and in analogy with, for example, the stability aspects of differential equations or of numerical computations, robustness theories can be viewed as stability theories of statistical inference."

- p. 8, Hampel, Rousseeuw, Ronchetti and Stahel (1986)

Seeking insights through analytical work

For high-dim data such as ours, removing some data units could change significantly the outcomes of our model because of feature dependence.

This phenomenon is also seen in simulated data from Gaussian linear models in high-dim.

How does sample to sample variability

interact with heavy tail errors?

Sample stability meets robust statistics in high-dim (El Karoui, Bean, Bickel, Lim and Yu, 2012)

Set-up: Linear regression model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

For i=1,…,n:

$$X_i \sim N(0, \Sigma_p), \quad \epsilon_i \; iid, \;\; E\epsilon_i = 0$$

We consider the random-matrix regime: $p/n \to \kappa \in (0, 1)$

$$\hat{\beta} = argmin_{\beta \in R^p} \sum_i \rho(Y_i - X_i'\beta)$$

Due to invariance, WLOG, assume $\Sigma_p = I_p, \beta = 0$

Sample stability meets robust statistics in high-dim

RESULT (in an important special case):

1. Let $r_\rho(p,n) = ||\hat{\beta}||$, then $\hat{\beta}$ is distributed as $r_\rho(p,n)U$

   where $U \sim uniform(S^{p-1})(1)$

2. $r_\rho(p,n) \to r_\rho(\kappa)$, let $\hat{z}_\epsilon := \epsilon + r_\rho(\kappa)Z$, Z indep of ϵ

$$prox_c(\rho)(x) = argmin_{y \in R}[\rho(y) + \frac{(x-y)^2}{2c}]$$

Then $r_\rho(\kappa)$ satisfies

$$E\{[prox_c(\rho)]'\} = 1 - \kappa$$

$$E\{[\hat{z}_\epsilon - prox_c(\hat{z}_\epsilon)]^2\} = \kappa r_\rho^2(\kappa)$$

Sample stability meets robust statistics in high-dim (continue)

In our limiting result, a normalization constant stabalizes.

Sketch of proof:

"Leave-one-out" trick both ways (perturbation argument)
(reminiscent of "swapping trick" for CLT)
Analytical derivations (prox functions)
(reminiscent of proving normality in the limit for CLT)

Sample stability meets robust statistics in high-dim (continue)

Corollary: Under our regression model with iid exponential errors,

when $\kappa = p/n > 0.3$,

L_2 loss fitting (OLS) is better than

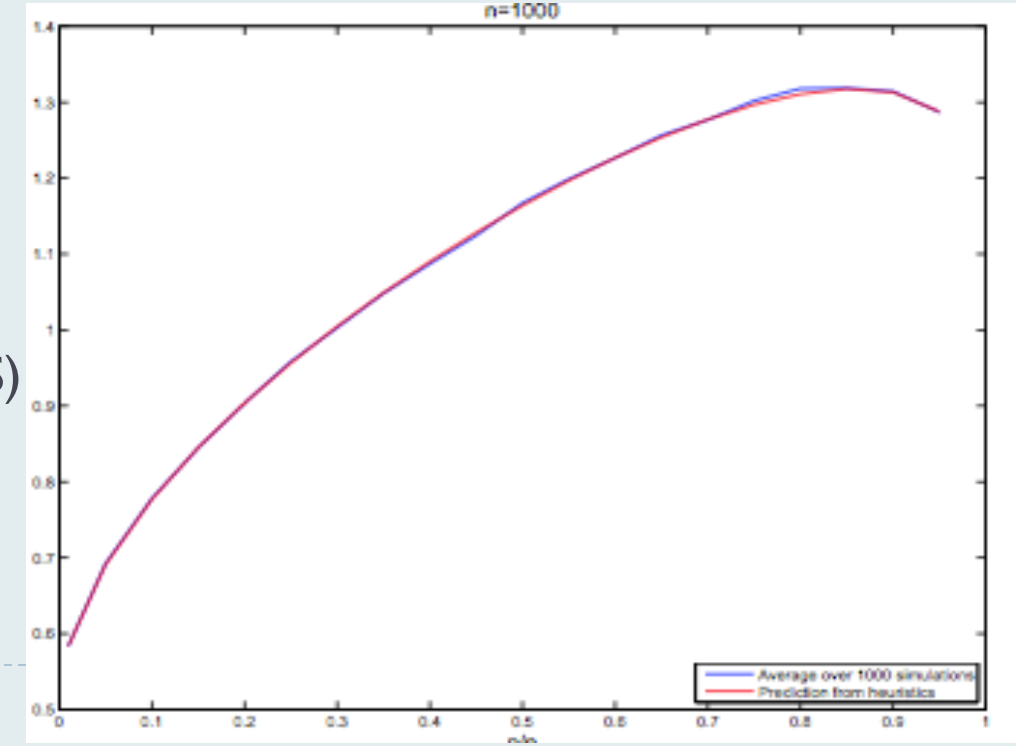
L_1 loss fitting (LAD) in terms of MSE or Var.

Plot on the right for n=1000

x-axis: κ = p/n in (0,1)

y-axis:

ratio of MSE (LAD) to MSE (OLS)



Blue curve – simulated →

Magenta curve – analytic →

Sample stability meets robust statistics in high-dim

Remarks:

MLE doesn't work here for a different reason than in cases where penalized MLE works better than MLE. We have unbiased estimators, a non-sparse situation and the question is about variance.

Optimal loss function can be calculated (Bean et al, 2012)

Simulated model with design matrix from fMRI data and double-exp. error shows the same phenomenon: p/n > 0.3, OLS is better than LAD. Some insurance for using L2 loss function in fMRI project.

Results hold for more general settings.

Current directions: adventure continues

- Challenge in the past 5 years: encoding and decoding of V4 brain signals

  V4: known difficult region where memory and attention matter

- Our encoding model for V4 single-neuron data gives the best prediction performance known

  Mairal, Benjamini, Willmore, Gallant & Yu (2012) (in prep)

  fMRI data to be looked at…

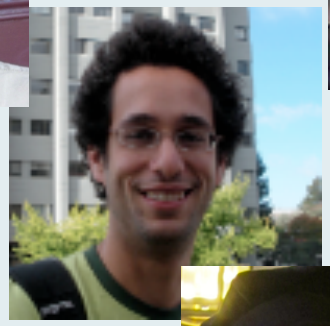
  Resulted decoding performance?

Collaborators

fMRI

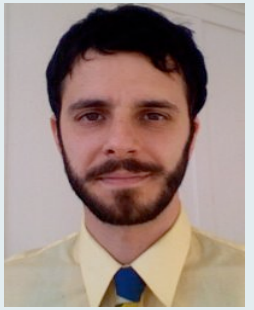
S. Nishimoto

J. Gallant



Y. Benjamini

A. Vu



T. Naselaris

ES-CV

C. Lim

Robust

N. El Karoui

D. Bean

P. Bickel

Funding acknowledgements

Guggenheim Foundation

NSF (DMS and CISE)

Parting message

What of the future? The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it? That remains to us, to our willingness to take up the rocky road of real problems in preferences to the smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments. Who is for the challenge?

--- John W. Tukey (1962) "Future of Data Analysis"