Case Study 3: fMRI Prediction

~~LASSO Regression~~

LARS, Fused LASSO

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington

Emily Fox

February 21th, 2013

LASSO Regression

- **LASSO:** least absolute shrinkage and selection operator
- New objective:

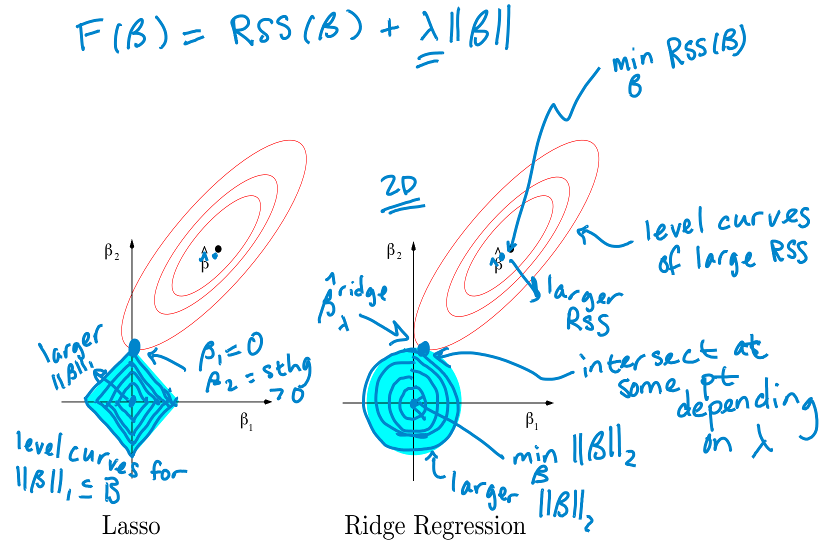
$$\min_{\beta} \underbrace{\sum_{i=1}^{N} (y^i - (\beta_0 + \beta^T x^i))^2}_{RSS(\beta)} + \lambda \|\beta\|_1$$

$$\Updownarrow$$

$$\min_{\beta} RSS(\beta) \quad \text{s.t.} \quad \|\beta\|_1 \leq B$$

Geometric Intuition for Sparsity

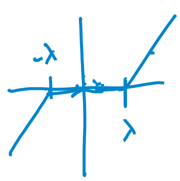
$F(\beta) = RSS(\beta) + \lambda \|\beta\|$

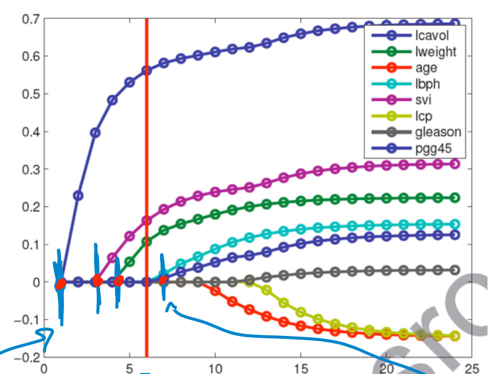


©Emily Fox 2013 3

Now: *LASSO Coefficient Path*

Again, each λ indexes a diff. soln





From
Kevin Murphy
textbook

$\|\beta\|_1 \le B$

there are only a few critical values of λ where new cov. enter model

CV soln

B

inc. λ

solns are sparse for any given λ

©Emily Fox 2013 4

LASSO Algorithms

- Standard convex optimizer
- Least angle regression (LAR)
  - Efron et al. 2004
  - Computes entire path of solutions
  - State-of-the-art until 2008
- Pathwise coordinate descent – new
- More on these "shooting" algorithms next time…

LARS – Efron et al. 2004

- LAR is an efficient stepwise variable selection algorithm
  - "useful and less greedy version of traditional forward selection methods"

Efron

- Can be modified to compute regularization path of LASSO
  - → LARS (Least angle regression and *shrinkage*)
- Increasing upper bound *B*, coefficients gradually "turn on"
  - Few critical values of *B* where support changes
  - Non-zero coefficients increase or decrease linearly between critical points
  - Can solve for critical values analytically

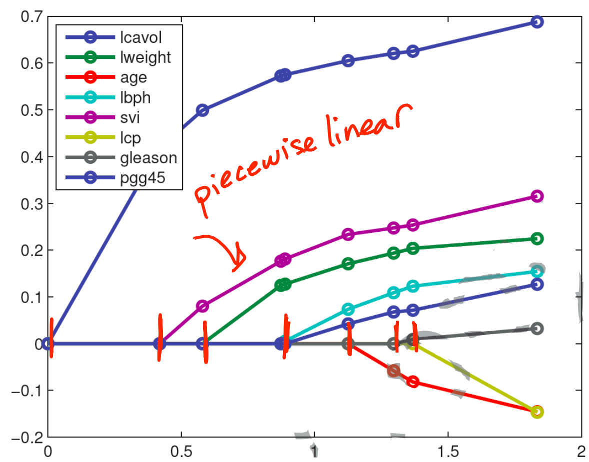
key to providing full reg path

- Complexity:

$O(\min(Np^2, pN^2))$

\# of obs. \# of covariates = cost of a single LS soln

LASSO Coefficient Path



From
Kevin Murphy
textbook

LARS – Algorithm

- Assumptions: standardize
  - Response has 0 mean

$$\sum_i y^i = 0$$

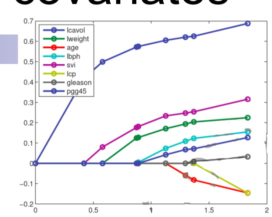
  - Covariates are normalized

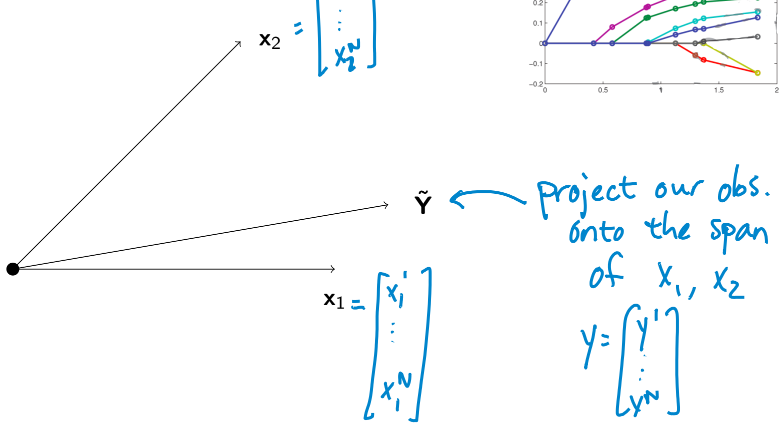
$$\sum_i x_j^i = 0 \qquad \sum_i (x_j^i)^2 = 1 \qquad j = 1, \dots, p$$

LARS – Algorithm Overview

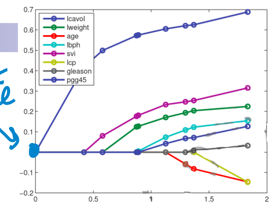
- Start with all coefficient estimates $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_p = 0$
- Let $\mathcal{A}$ be the "active set" of covariates most correlated with the "current" residual ← based on covariates already in model
- Initially, $\mathcal{A} = \{x_{j_1}\}$ for some covariate x_{j_1}
- Take the largest possible step in the direction of x_{j_1} until another covariate x_{j_2} enters $\mathcal{A}$
- Continue in the direction equiangular between x_{j_1} and x_{j_2} until a third covariate x_{j_3} enters $\mathcal{A}$
- Continue in the direction equiangular between $x_{j_1}, x_{j_2}, x_{j_3}$ until a fourth covariate x_{j_4} enters $\mathcal{A}$
- This procedure continues until all covariates are added at which point

LARS – Illustration for *p*=2 covariates



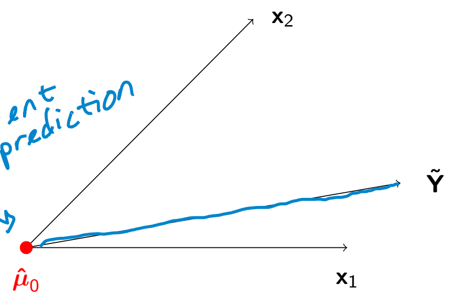


LARS – Illustration for *p*=2 covariates



start here

current prediction



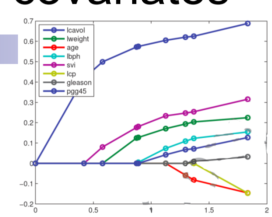
$\tilde{y} - \hat{\mu}_0$ = "current" residual

most corr. covariate

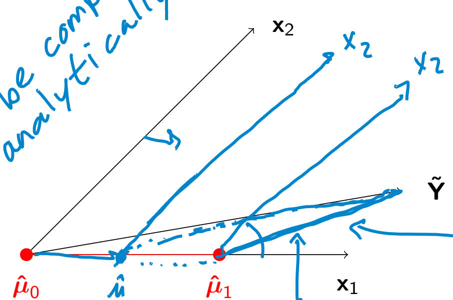
$\hat{\mu} = X\hat{\beta}$

prediction based on $\hat{\beta}$

LARS – Illustration for *p*=2 covariates



$\hat{\gamma}_1$ can be computed analytically



"current prediction"

$\tilde{y} - \hat{\mu}_1$

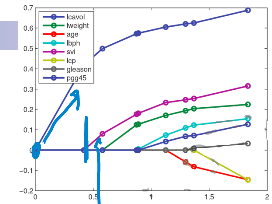
keep going

bisects ∠ bt x_1, x_2

⇒ x_1, x_2 are equally corr. w/ residual

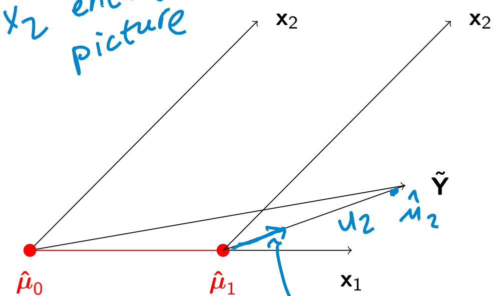
$\boxed{\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 x_1}$

LARS – Illustration for *p*=2 covariates



x_2 enters picture

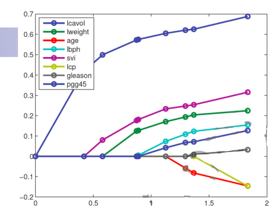
x_2 enters

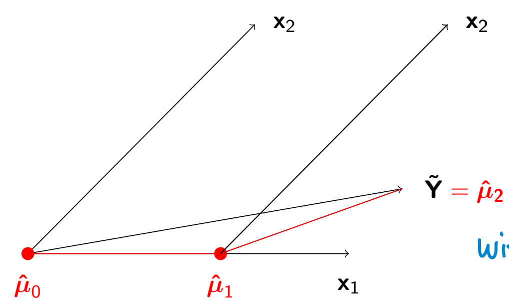


walk in dir "equiangular" bt x_1, x_2

$\hat{\mu}_2 = \hat{\mu}_1 + \underline{\underline{\hat{\gamma}_2}}\, u_2$

LARS – Illustration for *p*=2 covariates



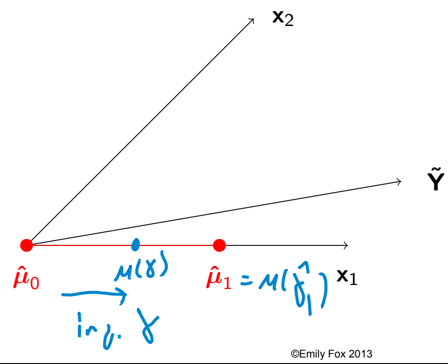


with 2 cov, when both are in the model, $\hat{\mu}_2 = \tilde{Y}$

LARS-LASSO Relationship

- Let $\mu(\gamma) = X\beta(\gamma)$ with $\beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j$ ← comes from LS soln based on active set
- We showed that for active covariate *j*: $\text{sign}(\hat{\beta}_j) = \text{sign}(x_j'\underbrace{(y - \hat{\mu})}_{c_j})$

c_j corr. bt x_j and our residual w/o x_j in model



Soft Threshholding

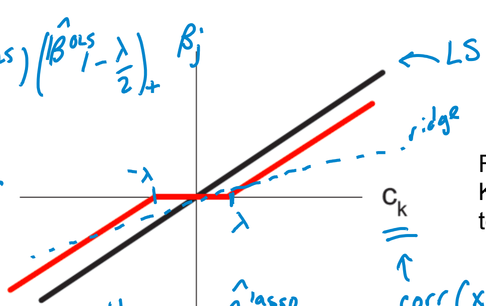
FROM LAST TIME

$$\boxed{\hat{\beta}_j} = \begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases} = \boxed{\text{sign}\left(\frac{c_j}{a_j}\right)}\left(\frac{|c_j|}{a_j} - \frac{\lambda}{a_j}\right)_+$$

If $X^TX = I$

$$\hat{\beta}_j^{lasso} = \text{sign}(\hat{\beta}_j^{OLS})\left(|\hat{\beta}_j^{OLS}| - \frac{\lambda}{2}\right)_+$$

$$\hat{\beta}_j^{ridge} = \frac{\hat{\beta}_j^{OLS}}{1+\lambda}$$



From Kevin Murphy textbook

In Lasso, all coeff $\hat{\beta}_j^{lasso}$ are shrunk relative to $\hat{\beta}_j^{OLS}$

LARS-LASSO Relationship

- Let $\mu(\gamma) = X\beta(\gamma)$ with $\beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j$
- We showed that for active covariate *j:* $\text{sign}(\hat{\beta}_j) = \text{sign}(x_j'(y - \hat{\mu}))$
- $\beta_j(\gamma)$ changes sign at $\beta_j(\gamma) = 0 \Rightarrow \gamma = -\frac{\hat{\beta}_j}{\hat{d}_j}$

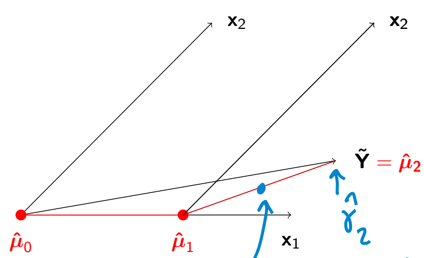
violated

- 1st sign change occurs at $\tilde{\gamma} = \min_{\gamma_j > 0}\{\gamma_j\}$ for covariate $\tilde{j}$

1st cov. to change signs

LARS-LASSO Relationship

- If $\tilde{\gamma}$ occurs before $\hat{\gamma}$, then next LARS step is not a LASSO solution

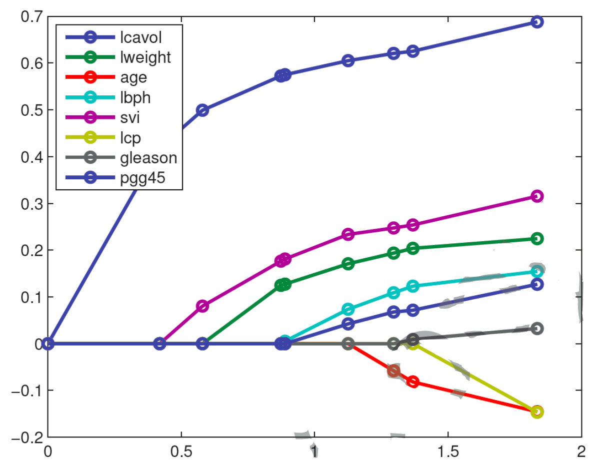


$\tilde{\gamma}$ stop here

- **LASSO modification:**

If $\tilde{\gamma} < \hat{\gamma}$, then stop LARS at $\gamma = \tilde{\gamma}$ and remove $\tilde{j}$ from our calc. of equiangular dir.
(remove $\beta_{\tilde{j}}$ from model + restart)

LASSO Coefficient Path



From
Kevin Murphy
textbook

Comments

- LARS increases $\mathcal{A}$, but LASSO allows it to decrease
- Only involves a single index at a time
- If $p > N$, LASSO returns at most N variables
- If group of variables are highly correlated, LASSO tends to choose one to include rather arbitrarily
  - Straightforward to observe from LARS algorithm….Sensitive to noise.

beware of interpreting the variables included

Comments

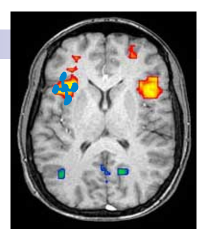
- In general, can't solve analytically for GLM (e.g., logistic reg.)
  - Gradually decrease λ and use efficiency of computing $\hat{\beta}(\lambda_k)$ from $\hat{\beta}(\lambda_{k-1})$ = warm-start strategy
  - See Friedman et al. 2010 for coordinate ascent + warm-starting strategy

- If $N > p$, but variables are correlated, ridge regression tends to have better predictive performance than LASSO (Zou & Hastie 2005)
  - Elastic net is hybrid between LASSO and ridge regression

$$\|y - X\beta\|_2^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \|\beta\|_2^2$$

(there some issues . . . details KM book)

Fused LASSO



- Might want coefficients of neighboring voxels to be similar

  discover regions of importance

- How to modify LASSO penalty to account for this?

- Graph-guided fused LASSO
  - Assume a 2d lattice graph connecting neighboring pixels in the fMRI image
  - Penalty:

$$\|y - X\beta\|_2^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_{(s,t)\in E} |\beta_s - \beta_t|$$

penalize these, taking diff vals

has edge in graph

Generalized LASSO

- Assume a structured linear regression model:

$$\|y - X\beta\|_2^2 + \lambda \|D\beta\|_1$$

$$D \in \mathbb{R}^{m \times p}$$

- If D is invertible, then get a new LASSO problem if we substitute

$$\alpha = D^{-1}\beta$$

- Otherwise, not equivalent
- For solution path, see
Ryan Tibshirani and Jonathan Taylor, "The Solution Path of the Generalized Lasso." Annals of Statistics, 2011.

Generalized LASSO

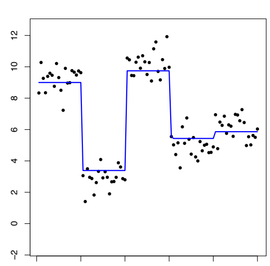
signal approximation scenario

$X = I$

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda \|D\beta\|_1$$

Let $D = \begin{bmatrix} -1 & 1 & 0 & 0 & \dots \\ 0 & -1 & 1 & 0 & \dots \\ 0 & 0 & -1 & 1 & \dots \\ \vdots & & & & \end{bmatrix}$. This is the 1d fused lasso.

$$\lambda \sum_j |\beta_j - \beta_{j-1}|$$



encourage piecewise const. ~~linear~~

Generalized LASSO

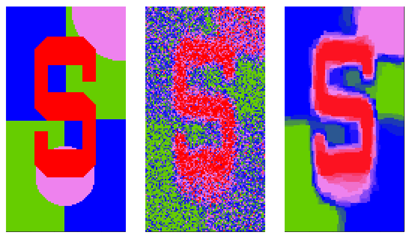
$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|D\beta\|_1$$

Suppose D gives "adjacent" differences in β:

$$D_i = (0, 0, \ldots -1, \ldots, 1, \ldots 0),$$

$\lambda \sum_{(s,t)\in E} |\beta_t - \beta_s|$

where adjacency is defined according to a graph $\mathcal{G}$. For a 2d grid, this is the 2d fused lasso.

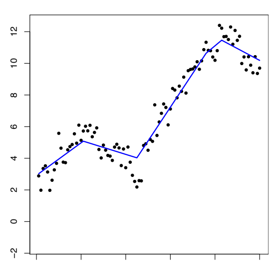


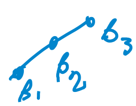
encourages constant regions

Generalized LASSO

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|D\beta\|_1$$

Let $D = \begin{bmatrix} -1 & 2 & -1 & 0 & \ldots \\ 0 & -1 & 2 & -1 & \ldots \\ 0 & 0 & -1 & 2 & \ldots \\ \vdots & & & & \end{bmatrix}$. This is linear trend filtering.





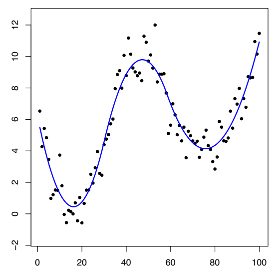
$\beta_3 - \beta_2 = \beta_2 - \beta_1$

$\Rightarrow 2\beta_2 - \beta_1 - \beta_3 = 0$

Generalized LASSO

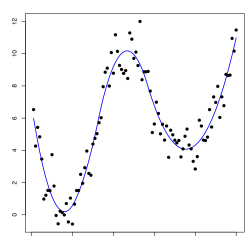
$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda \|D\beta\|_1$$

Let $D = \begin{bmatrix} -1 & 3 & -3 & 1 & \dots \\ 0 & -1 & 3 & -3 & \dots \\ 0 & 0 & -1 & 3 & \dots \\ \vdots & & & & \end{bmatrix}$. Get quadratic trend filtering.

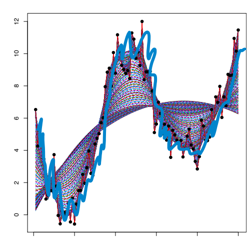


Generalized LASSO

- Tracing out the fits as a function of the regularization parameter



$\hat{\beta}_\lambda$ for $\lambda = 25$



$\hat{\beta}_\lambda$ for $\lambda \in [0, \infty]$

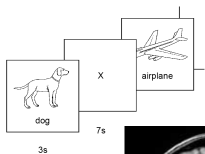
fMRI Prediction Task

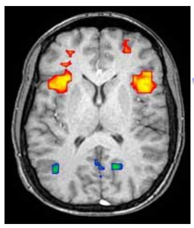
- **Goal:** Predict word stimulus from fMRI image

Can we read your brain?

$p >> N$

yes!





Classifier (logistic regression, kNN, …) → ~~HAMMER~~ or HOUSE

Zero-Shot Classification

- From training data, learn two mappings:
  - Key → S: input image → semantic features
  - L: semantic features → word

A = { → "dog" } few

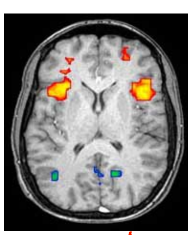
B = { [ ] → "dog" } many

- Can use "cheap" co-occurrence data to help learn L

from B

Training: { → [ ] → "dog" } N examples … N small

use both A + B



Features of word → **Classifier** (logistic regression, kNN, …) → ~~HAMMER~~ or HOUSE

Predict: new image → [ ] → "giraffe"

S ← learned from training data

using B

Semantic Features

Google Trillion word corpus

Semantic feature values: "**celery**"	Semantic feature values: "**airplane**"
0.8368, eat	0.8673, ride
0.3461, taste	0.2891, see
0.3153, fill	0.2851, say
0.2430, see	0.1689, near
0.1145, clean	0.1228, open
0.0600, open	0.0883, hear
0.0586, smell	0.0771, run
0.0286, touch	0.0749, lift
…	…
…	…
0.0000, drive	0.0049, smell
0.0000, wear	0.0010, wear
0.0000, lift	0.0000, taste
0.0000, break	0.0000, rub
0.0000, ride	0.0000, manipulate

co-occurrence

fMRI Prediction Results

- Palatucci et al., "Zero-Shot Learning with Semantic Output Codes", NIPS 2009

- fMRI dataset:
  - 9 participants
  - 60 words (e.g., *bear, dog, cat, truck, car, train*, …)
  - 6 scans per word
  - Preprocess by creating 1 "time-average" image per word

- Knowledge bases
  - Corpus5000 – semantic co-occurrence features with 5000 most frequent words in Google Trillion Word Corpus
  - human218 – Mechanical Turk (Amazon.com)
    218 semantic features (*"is it manmade?", "can you hold it?"*,…)
    Scale of 1 to 5

fMRI Prediction Results

- **First stage:** Learn mapping from images to semantic features
- Ridge regression $X \in \mathbb{R}^{N \times p} \longrightarrow F \in \mathbb{R}^{N \times d} \leftarrow$ # sem. features

From training data $\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T F \leftarrow$ stack up solns for each sem. feature

$\hat{f}^{new} = X^{new} \hat{\beta}^{ridge}$

- **Second stage:** 1-NN classification using knowledge base

look for word w/ f closest to $\hat{f}^{new}$

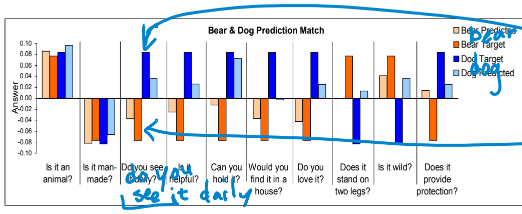
fMRI Prediction Results

- Leave-two-out-cross-validation
  - Learn ridge coefficients using 58 fMRI images
  - Predict semantic features of 1st heldout image
  - Compare whether semantic features of 1st or 2nd heldout image are closer

Table 1: Percent accuracies for leave-two-out-cross-validation for 9 fMRI participants (labeled P1-P9). The values represent classifier percentage accuracy over 3,540 trials when discriminating between two fMRI images, both of which were omitted from the training set.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	Mean
corpus5000	79.6	67.0	69.5	56.2	77.7	65.5	71.2	72.9	67.9	**69.7**
human218	90.3	82.9	86.6	71.9	89.5	75.3	78.0	77.7	76.2	**80.9**

$\leftarrow$ stat. sig.



yes, see dogs daily

"bear" ... no

Figure 1: Ten semantic features from the human218 knowledge base for the words *bear* and *dog*. The true encoding is shown along with the predicted encoding when fMRI images for bear and dog were left out of the training set.

fMRI Prediction Results

- Leave-one-out-cross-validation
  - Learn ridge coefficients using 59 fMRI images
  - Predict semantic features of heldout image
  - Compare whether very large set of possible other words

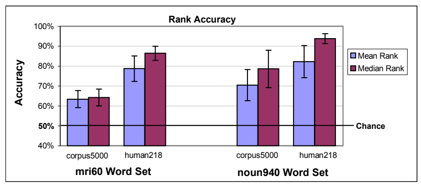


Figure 2: The mean and median rank accuracies across nine participants for two different semantic feature sets. Both the original 60 fMRI words and a set of 940 nouns were considered.

Table 2: The top five predicted words for a novel fMRI image taken for the word in bold (all fMRI images taken from participant P1). The number in the parentheses contains the rank of the correct word selected from 941 concrete nouns in English.

Bear	**Foot**	**Screwdriver**	**Train**	**Truck**	**Celery**	**House**	**Pants**
(1)	(1)	(1)	(1)	(2)	(5)	(6)	(21)
bear	*foot*	*screwdriver*	*train*	jeep	beet	supermarket	clothing
fox	feet	pin	jet	*truck*	artichoke	hotel	vest
wolf	ankle	nail	jail	minivan	grape	theater	t-shirt
yak	knee	wrench	factory	bus	cabbage	school	clothes
gorilla	face	dagger	bus	sedan	*celery*	factory	panties

Acknowledgements

- Some material in this lecture was based on slides provided by:
  - Tom Mitchell – fMRI
  - Rob Tibshirani – LASSO
  - Ryan Tibshirani – Fused LASSO