Case Study 3: fMRI Prediction

fMRI Prediction Task, Ridge, LASSO Review Fused LASSO, LARS

Machine Learning for Big Data
CSE547/STAT548, University of Washington

So far large/streaming N,
now big-p domain

Emily Fox

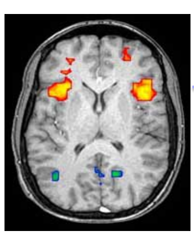
January 30th, 2014

fMRI Prediction Task

- **Goal:** Predict word stimulus from fMRI image
- **Challenges:** # of voxels = # params
  - p >> N (feature dimension >> sample size)
  - Cost of fMRI recordings is high
  - Only have a few training examples for each word

many more params than obs.

What can we do?



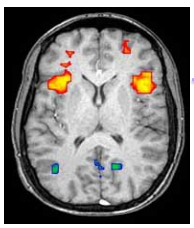
Classifier
(logistic regression, kNN, …)

~~HAMMER~~
or
HOUSE

Zero-Shot Classification

- **Goal:** Classify words not in the training set
- **Challenges:**
  - Cost of fMRI recordings is high
  - Can't get recordings for every word in the vocabulary

Never showed "giraffe" in scanner



Classifier
(logistic regression, kNN, …)

~~HAMMER~~
or
HOUSE

Semantic Features

Google Trillion word corpus

Semantic feature values: "**celery**"
0.8368, eat
0.3461, taste
0.3153, fill
0.2430, see
0.1145, clean
0.0600, open
0.0586, smell
0.0286, touch
…
…
0.0000, drive
0.0000, wear
0.0000, lift
0.0000, break
0.0000, ride

co-occurrence

Semantic feature values: "**airplane**"
0.8673, ride
0.2891, see
0.2851, say
0.1689, near
0.1228, open
0.0883, hear
0.0771, run
0.0749, lift
…
…
0.0049, smell
0.0010, wear
0.0000, taste
0.0000, rub
0.0000, manipulate

Zero-Shot Classification

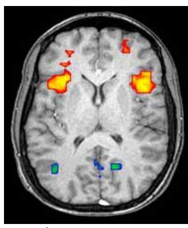
image
A= { [image] → "dog" }
few examples

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

B= { [⋮] → "dog" }
semantic feature vecs
many examples

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

Training: { [image] → [⋮] → "dog" } N examples... N small
use both A + B



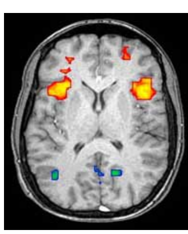
Features of word → **Classifier** (logistic regression, kNN, …) → ~~HAMMER~~ or HOUSE

new image
using B
Predict : [image] → [⋮] → "giraffe"
S learned on training data

©Emily Fox 2014 5

fMRI Prediction Subtask

- **Goal:** Predict semantic features from fMRI image



Features of word

©Emily Fox 2014 6

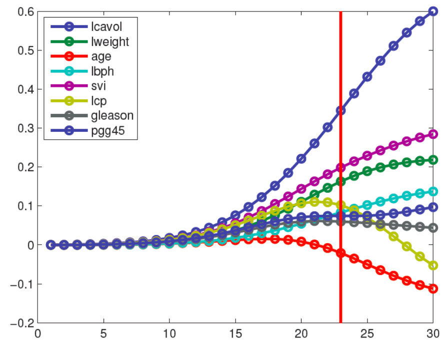
Linear Regression

- Model:
- MLE: $\hat{\theta} = \arg\max_{\theta} \; \log p(D \mid \theta)$
- Minimizing RSS= least squares regression

Ridge Regression

- Ameliorating issues with overfitting:
- New objective:
  - Solution:

Ridge Coefficient Path



From
Kevin Murphy
textbook

- Typical approach: select λ using cross validation

Case Study 3: fMRI Prediction

fMRI Prediction Results

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
January 30th, 2014

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

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

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**

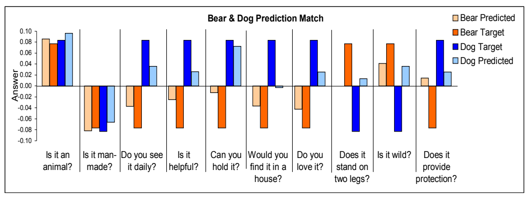


Figure 1: Ten semantic features from the human218 knowledge base for the words *bear* and *dog*. The true encoding is shown along with the predicted encoding when fMRI images for bear and dog were left out of the training set.

fMRI Prediction Results

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

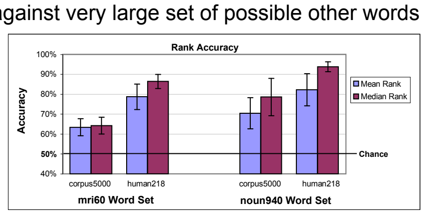


Figure 2: The mean and median rank accuracies across nine participants for two different semantic feature sets. Both the original 60 fMRI words and a set of 940 nouns were considered.

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

Case Study 3: fMRI Prediction

LASSO Review

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
January 30th, 2014

Variable Selection

- Ridge regression: Penalizes large weights
- What if we want to perform "feature selection"?
  - E.g., Which regions of the brain are important for word prediction?
  - Can't simply choose predictors with largest coefficients in ridge solution
  - Computationally impossible to perform "all subsets" regression
  - Stepwise procedures are sensitive to data perturbations and often include features with negligible improvement in fit
- Try new penalty: Penalize non-zero weights
  - Penalty:
  - Leads to sparse solutions
  - Just like ridge regression, solution is indexed by a continuous param λ

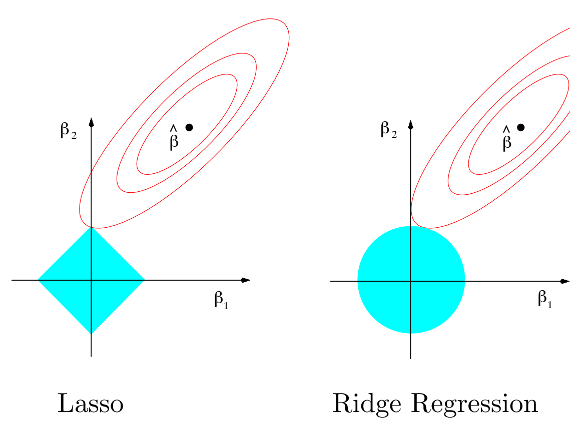
LASSO Regression

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

- New objective:

©Emily Fox 2014 17

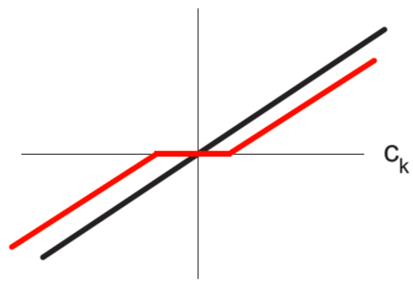
Geometric Intuition for Sparsity



©Emily Fox 2014 18

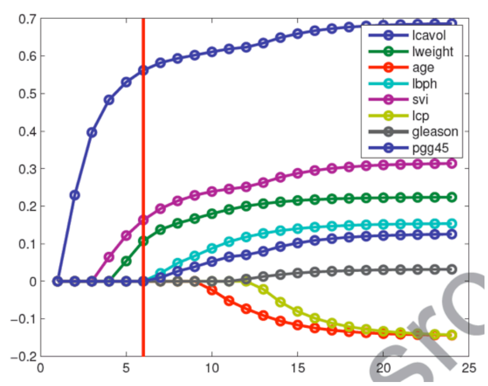
Soft Threshholding

$$\hat{\beta}_j = \begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases}$$



From
Kevin Murphy
textbook

LASSO Coefficient Path

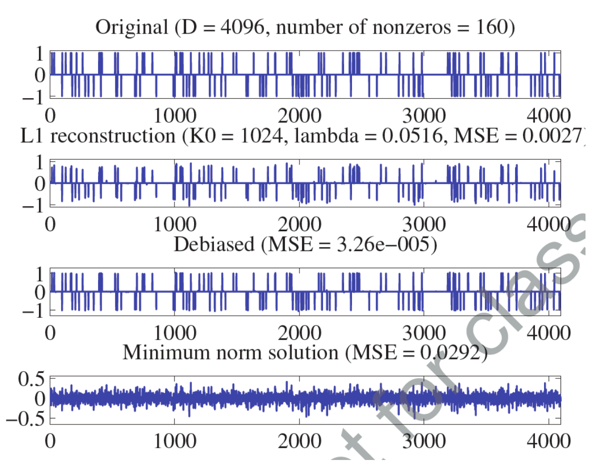


From
Kevin Murphy
textbook

LASSO Example

Term	Least Squares	Ridge	Lasso
Intercept	2.465	2.452	2.468
lcavol	0.680	0.420	0.533
lweight	0.263	0.238	0.169
age	−0.141	−0.046	
lbph	0.210	0.162	0.002
svi	0.305	0.227	0.094
lcp	−0.288	0.000	
gleason	−0.021	0.040	
pgg45	0.267	0.133	

Debiasing



From Kevin Murphy textbook

Sparsistency

- Typical Statistical Consistency Analysis:
  - Holding model size (*p*) fixed, as number of samples (*N*) goes to infinity, estimated parameter goes to true parameter

- Here we want to examine *p* >> *N* domains
- Let both model size *p* and sample size *N* go to infinity!
  - Hard case: *N* = *k* log *p*

Sparsistency

- Rescale LASSO objective by *N*:

- Theorem (Wainwright 2008, Zhao and Yu 2006, …):
  - Under some constraints on the design matrix *X*, if we solve the LASSO regression using

  Then for some c_1>0, the following holds with at least probability

    - The LASSO problem has a unique solution with support contained within the true support
    - If $\min_{j \in S(\beta^*)} |\beta_j^*| > c_2 \lambda_n$ for some c_2>0, then $S(\hat{\beta}) = S(\beta^*)$

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…

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

Case Study 3: fMRI Prediction

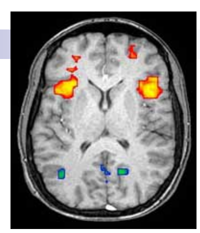
Fused LASSO

Machine Learning for Big Data
CSE547/STAT548, University of Washington

Emily Fox

January 30^{th}, 2014

Fused LASSO



- Might want coefficients of neighboring voxels to be similar
- 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:

Generalized LASSO

- Assume a structured linear regression model:

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

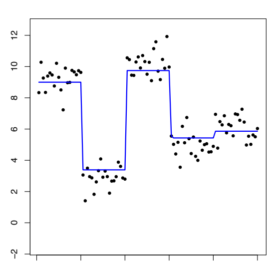
- Otherwise, not equivalent

- For solution path, see
  Ryan Tibshirani and Jonathan Taylor, “The Solution Path of the Generalized Lasso.” Annals of Statistics, 2011.

Generalized LASSO

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

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.



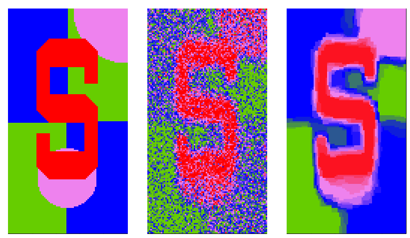
Generalized LASSO

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

Suppose D gives "adjacent" differences in β:

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

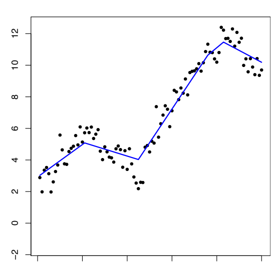
where adjacency is defined according to a graph $\mathcal{G}$. For a 2d grid, this is the 2d fused lasso.



Generalized LASSO

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

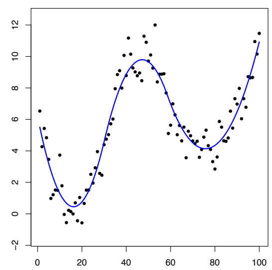
Let $D = \begin{bmatrix} -1 & 2 & -1 & 0 & \ldots \\ 0 & -1 & 2 & -1 & \ldots \\ 0 & 0 & -1 & 2 & \ldots \\ \vdots & & & & \end{bmatrix}$. This is linear trend filtering.



Generalized LASSO

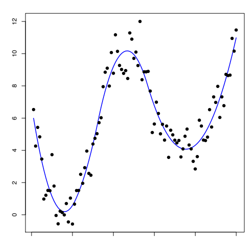
$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|D\beta\|_1$$

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.

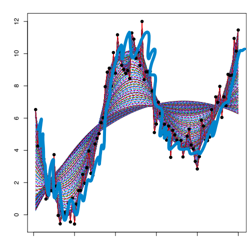


Generalized LASSO

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



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



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

Case Study 3: fMRI Prediction

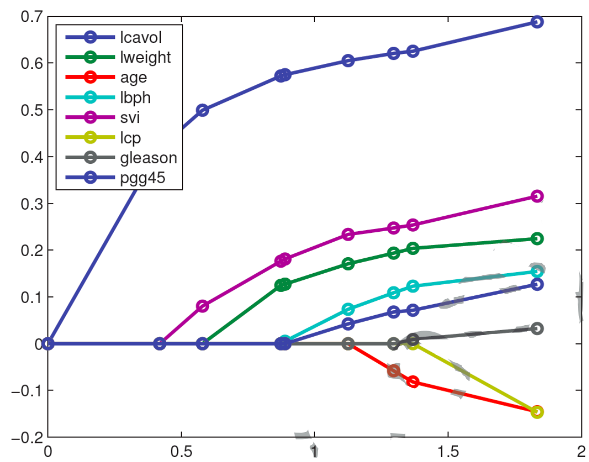
LASSO Solvers – Part 1: LARS

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
January 30th, 2014

LARS – Efron et al. 2004

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

LASSO Coefficient Path

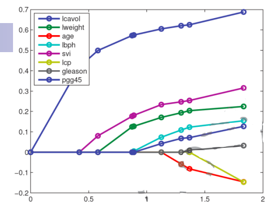


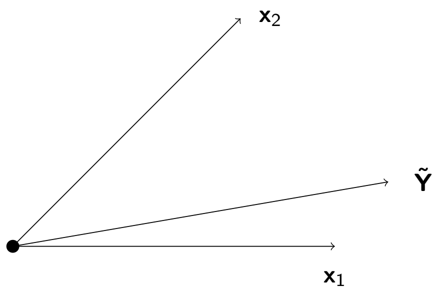
From
Kevin Murphy
textbook

LARS – Algorithm Overview

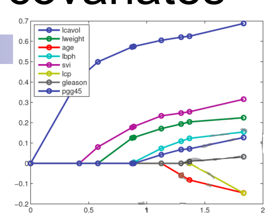
- Start with all coefficient estimates
- Let $\mathcal{A}$ be the "active set" of covariates most correlated with the "current" residual
- Initially, $\mathcal{A} = \{x_{j_1}\}$ for some covariate x_{j_1}
- Take the largest possible step in the direction of x_{j_1} until another covariate x_{j_2} enters $\mathcal{A}$
- 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

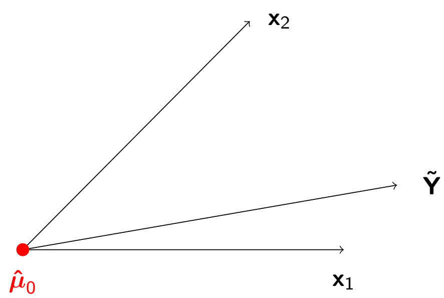
LARS – Illustration for *p*=2 covariates



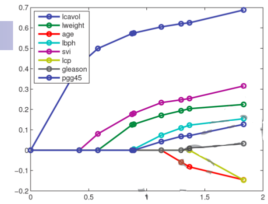


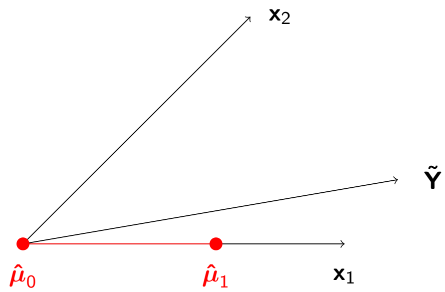
LARS – Illustration for *p*=2 covariates



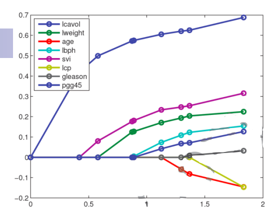


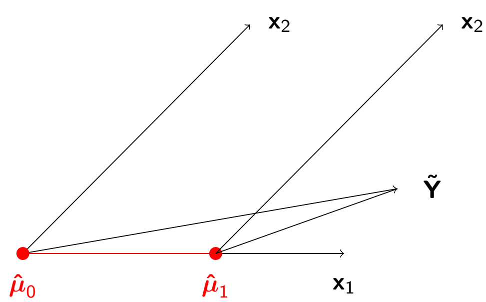
LARS – Illustration for *p*=2 covariates



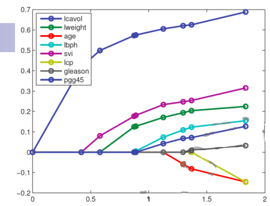


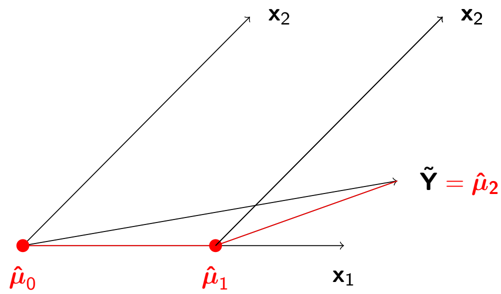
LARS – Illustration for *p*=2 covariates





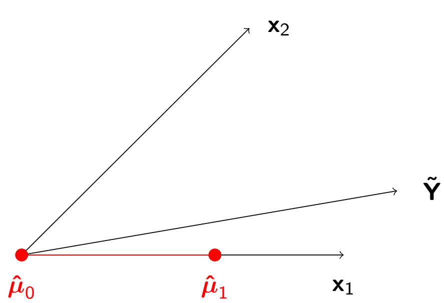
LARS – Illustration for *p*=2 covariates





LARS-LASSO Relationship

- Let $\mu(\gamma) = X\beta(\gamma)$ with
- We showed that for active covariate *j*: $\text{sign}(\hat{\beta}_j) = \text{sign}(x_j'(y - \hat{\mu}))$

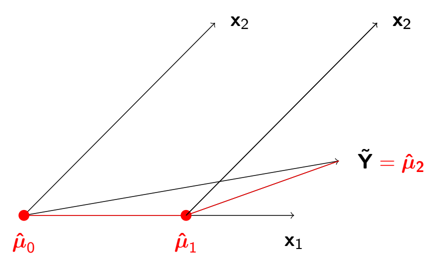


LARS-LASSO Relationship

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

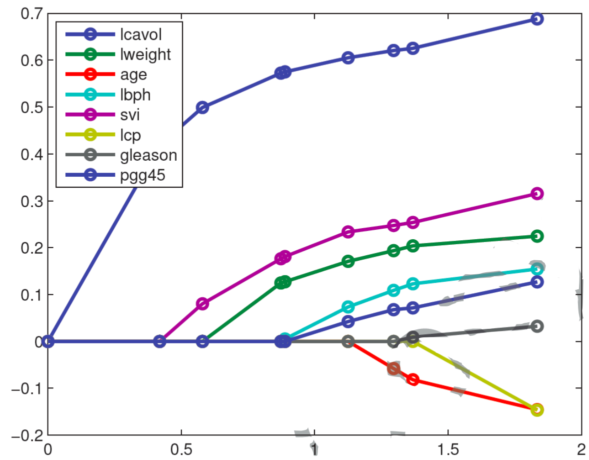
LARS-LASSO Relationship

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



- **LASSO modification:**

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.

Acknowledgements

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