Module 1: Nonparametric Preliminaries

LASSO cont'd

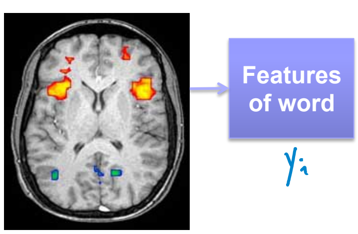
STAT/BIOSTAT 527, University of Washington

Emily Fox

April 7th, 2014

fMRI Prediction Subtask

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



y_i

x_i

$\begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \rightarrow y_{ij}$

j^{th} semantic feature

$p = \#$ voxels $\approx 20{,}000$

$\boxed{p >> n}$

n = # training examples

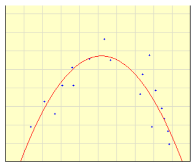
$\hat{\beta}^{LS} = (X^TX)^{-1}X^Ty$

rank deficient

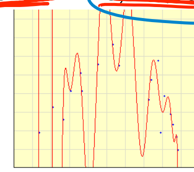
Regularization in Linear Regression

- Overfitting usually leads to very large parameter choices, e.g.:

$-2.2 + 3.1\,X - 0.30\,X^2$



$-1.1 + 4{,}700{,}910.7\,X - 8{,}585{,}638.4\,X^2 + \ldots$



even for $n > p$, p large

- ***Regularized*** or ***penalized*** regression aims to impose a "complexity" penalty by penalizing large weights
  - "Shrinkage" method

Ridge Regression

- Ameliorating issues with overfitting: penalization of weights "regularization"
- New objective:

$$\min_{\beta} \underbrace{\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta^T x_i)\right)^2}_{\text{LS obj.}} + \lambda \|\beta\|_2^2$$

don't penalize intercept

$\|\beta\|_2^2 = \beta^T\beta$

λ = strength of penalty

$$\Updownarrow$$

$$\min_{\beta} RSS(\beta) \quad \text{s.t. } \|\beta\|_2^2 \le S$$

Variable Selection

- not min this obj.
- coeff. sensitive to what's inc. in the model

- Ridge regression: Penalizes large weights
- What if we want to perform "feature selection"? (variable)
  - 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

    2^p subsets of predictors ... can't do this

    discrete
  - Stepwise procedures are sensitive to data perturbations and often include features with negligible improvement in fit ← greedy, ∃ backtracking alg.
- Try new penalty: Penalize non-zero weights
  - Penalty: L_1 $\|\beta\|_1 = \sum_j |\beta_j|$
  - 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:

$$\min_\beta \underbrace{\overbrace{\sum_{i=1}^n (y_i - (\beta_0 + \beta^T x_i))^2}^{\text{LS obj.}}}_{RSS(\beta)} + \lambda \|\beta\|_1$$

$\sum |\beta_j|$

L_1 penalty

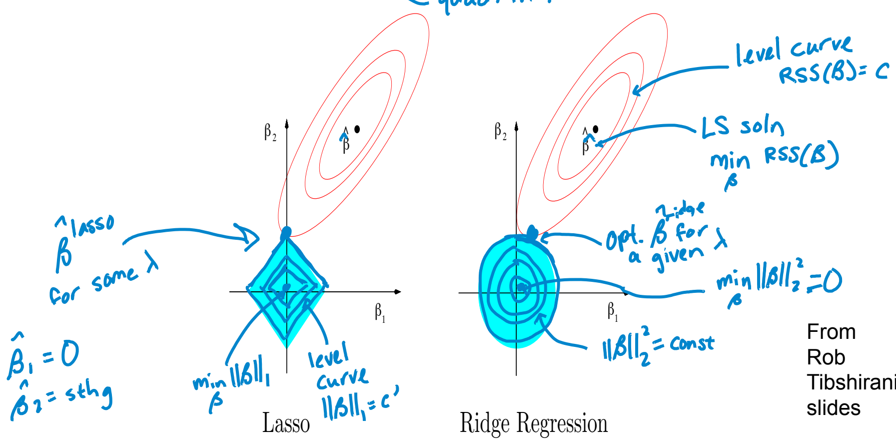
$$\Updownarrow$$

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

Geometric Intuition for Sparsity

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

quad. in β



From Rob Tibshirani slides

Soft Threshholding

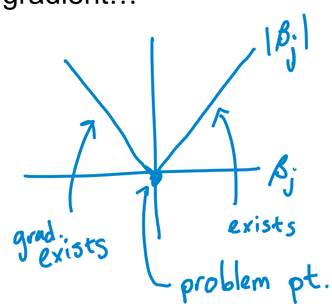
- To see why LASSO results in sparse solutions, look at conditions that must hold at optimum

  look at β_j ... do this for all $j \Rightarrow$ set of simultaneous equations

- L_1 penalty $||\beta||_1$ is not differentiable whenever $\beta_j = 0$

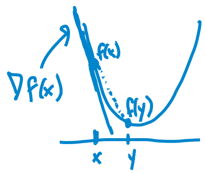
  $\Sigma|\beta_j|$

- Look at subgradient…



Subgradients of Convex Functions

- Gradients lower bound convex functions:

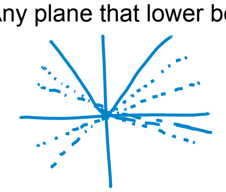


$$\frac{f(y)-f(x)}{y-x} \geq \nabla f(x)$$

$$\Rightarrow f(y) \geq f(x) + \nabla f(x)(y-x)$$

- Gradients are unique at **x** if function differentiable at **x**

- Subgradients: Generalize gradients to non-differentiable points:
  - Any plane that lower bounds function:



For $|\beta_j|$:

$v \in [-1,1]$

$$\boxed{v \in \partial f(x) \text{ subgrad. if } f(y) \geq f(x) + v(y-x)}$$

Soft Threshholding

Goal. $\nabla_{\beta_j}(\underbrace{RSS(\beta)+\lambda\|\beta\|_1}_{F(\beta)}) = 0$

- Gradient of RSS term:

$$\frac{\partial}{\partial \beta_j} RSS(\beta) = a_j \beta_j - c_j$$

a_j: $2\sum_{i=1}^{n}(x_{ij})^2$

c_j: $2\sum_{i=1}^{n} x_{ij}(y_i - \beta_{-j}^T x_{i,-j})$

β_{-j}: all β's except for β_j

$x_{i,-j}$: all cov's other than x_{ij}

$c_j \propto corr(x_j, r_{-j})$

c_j: msr of how relevant x_j is for pred y beyond what others can

r_{-j}: residuals from a model w/o j^{th} cov.

- Subgradient of full objective:

$$\partial_{\beta_j} F(\beta) = (a_j\beta_j - c_j) + \lambda\, \partial_{\beta_j}\|\beta\|_1$$

$$= \begin{cases} a_j\beta_j - c_j - \lambda & \beta_j < 0 \\ [-c_j-\lambda, -c_j+\lambda] & \beta_j = 0 \\ a_j\beta_j - c_j + \lambda & \beta_j > 0 \end{cases}$$

Soft Threshholding

- Set subgradient = 0:

$$\partial_{\beta_j} F(\beta) = \begin{cases} a_j\beta_j - c_j - \lambda & \beta_j < 0 \\ [-c_j - \lambda, -c_j + \lambda] & \beta_j = 0 \\ a_j\beta_j - c_j + \lambda & \beta_j > 0 \end{cases} = 0$$

If $\beta_j < 0$
$a_j\beta_j - c_j - \lambda = 0$
$\Rightarrow \beta_j = \frac{c_j + \lambda}{a_j} < 0 \Rightarrow c_j < -\lambda$ If strong neg. corr., then $\hat{\beta}_j < 0$

If $\beta_j > 0$
$a_j\beta_j - c_j + \lambda = 0 \Rightarrow \beta_j = \frac{c_j - \lambda}{a_j} > 0 \Rightarrow c_j > \lambda$ If strong pos. corr, then $\hat{\beta}_j > 0$

If $\beta_j = 0$ $-\lambda < c_j < \lambda$ if not strong corr, then $\hat{\beta}_j = 0$

- The value of $c_j = 2\sum_{i=1}^{N} x_j^i (y^i - \beta'_{-j} x_{-j}^i)$ constrains β_j

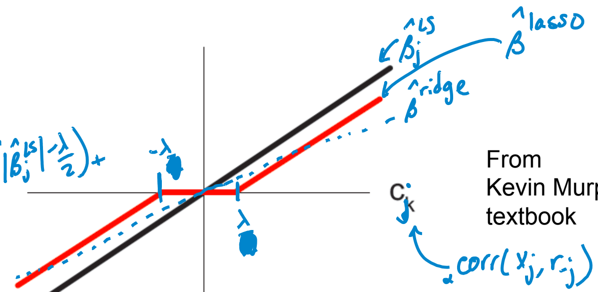
Soft Threshholding

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

If $X^T X = I$

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

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



From Kevin Murphy textbook

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

Coordinate Descent

- Given a function F (β)
  - Want to find minimum $\beta^* = \min_\beta F(\beta) \leftarrow F(\beta_1, \dots, \beta_p)$
- Often, hard to find minimum for all coordinates, but easy for one coordinate

  1-d optimization problem ... just solved for the lasso
- Coordinate descent:

  while not converged

  pick coord. j

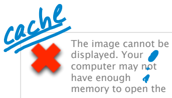
  $\beta_j \leftarrow \min_b F(\beta_1, \dots, \beta_{j-1}, b, \beta_{j+1}, \dots, \beta_p)$
- How do we pick a coordinate?

  Round robin, randomly, smartly, ...
- When does this converge to optimum?

  e.g. strongly convex, separability

Stochastic Coordinate Descent for LASSO (aka Shooting Algorithm)

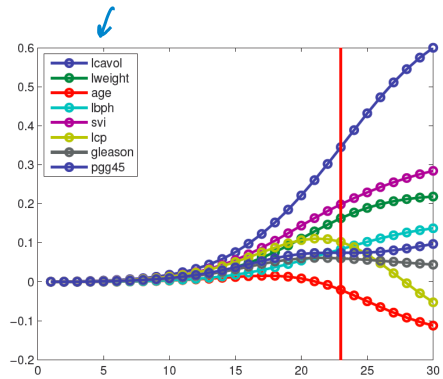
- Repeat until convergence
  - Pick a coordinate *j* at random
    - Set: $$\hat\beta_j = \begin{cases} (c_j+\lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j-\lambda)/a_j & c_j > \lambda \end{cases} = \text{sign}(c_j)\frac{(|c_j|-\lambda)_+}{a_j}$$
    - Where: cache



$$c_j = 2\sum_{i=1}^{N} x_j^i (y^i - \beta'_{-j} x_{-j}^i)$$

  - For convergence rates, see Shalev-Shwartz and Tewari 2009
- Other common technique = LARS
  - Least angle regression and shrinkage, Efron et al. 2004

Recall: *Ridge Coefficient Path*

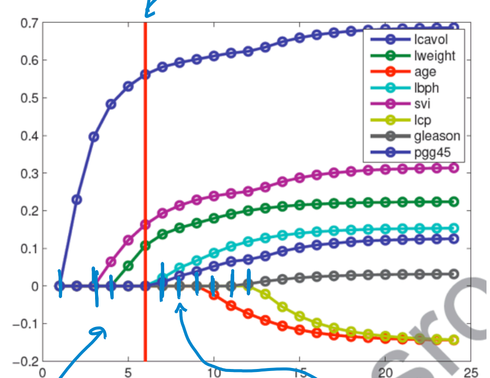


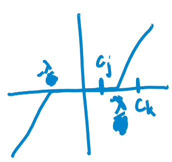
From
Kevin Murphy
textbook

- Typical approach: select λ using cross validation

Now: *LASSO Coefficient Path*

CV soln





From
Kevin Murphy
textbook

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

inc. λ B

solns are sparse for any given λ

only a few critical values of λ where support changes

LASSO Example

$\hat{b}$ CV solns

	Term	Least Squares	Ridge	Lasso
β_0	Intercept	2.465	2.452	2.468
β_1	lcavol	0.680	0.420	0.533
β_2	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	
β_p	pgg45	0.267	0.133	

not in the model

(sparse solns)

From Rob Tibshirani slides

Sparsistency

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

$\hat{\theta} \rightarrow \theta^*$? true param as $n \rightarrow \infty$

- Here we want to examine *p* >> *n* domains
- Let both model size *p* and sample size *n* go to infinity!
  - Hard case: $\boxed{n = k \log p}$

n grows slowly relative to p

Sparsistency

- Rescale LASSO objective by *n*:

$$\min_{\beta} \frac{1}{n} RSS(\beta) + \lambda_n \sum_j |\beta_j|$$

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

$$\lambda_n > \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log p}{n}}$$

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

$$1-4\exp(-c_1 n \lambda_n^2) \longrightarrow 1 :$$

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

Comments

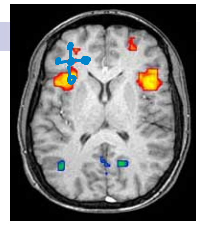
- In general, can't solve analytically for GLM (e.g., logistic reg.)
  - Gradually decrease λ and use efficiency of computing $\hat{\beta}(\lambda_k)$ from $\hat{\beta}(\lambda_{k-1})$ = warm-start strategy
  - See Friedman et al. 2010 for coordinate ascent + warm-starting strategy

- 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

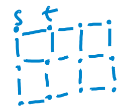
$$\|y - X\beta\|_2^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \|\beta\|_2^2$$

(still some issues, but other solns)

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:



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

↑ in edge set

penalizing $\beta_s \neq \beta_t$

A Bayesian Formulation

- Consider a model with likelihood

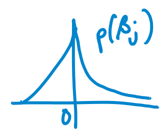
$$y_i \mid \beta \sim N(\beta_0 + x_i^T \beta, \sigma^2)$$

  and prior

$$\beta_j \sim \mathrm{Lap}(\beta_j; \lambda)$$

  where

$$\mathrm{Lap}(\beta_j; \lambda) = \frac{\lambda}{2} e^{-\lambda|\beta_j|}$$



$p(\beta_j = 0) = 0$
but posterior mode

- For large λ
  more peaked around 0
- LASSO solution is equivalent to the ***mode*** of the posterior
- Note: posterior mode ≠ posterior mean in this case
  any given posterior sample is not sparse, but it will be penalized like in ridge.
- There is no closed-form for the posterior. Rely on approx. methods.
  spike + slab priors as alternatives

Reading

- Hastie, Tibshirani, Friedman: 3.4, 3.8.6

What you should know

- LASSO objective
- Geometric intuition for differences between ridge and LASSO solns
- How LASSO performs soft threshholding
- Shooting algorithm
- Idea of sparsistency
- Ways in which other L1 and L1-Lp objectives can be encoded
  - Elastic net
  - Fused LASSO