Module 1: Nonparametric Preliminaries

LASSO cont'd

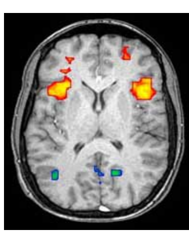
STAT/BIOSTAT 527, University of Washington

Emily Fox

April 9th, 2013

fMRI Prediction Subtask

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



x_i

Features of word

y_i

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

semantic feature

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

only small n

$\boxed{p \gg n}$

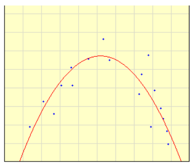
$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T y$$

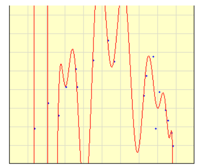
rank deficient

Regularization in Linear Regression

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

$-2.2 + 3.1\,X - 0.30\,X^2$ $\qquad$ $-1.1 + 4{,}700{,}910.7\,X - 8{,}585{,}638.4\,X^2 + \ldots$





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:

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

don't penalize the intercept

$\beta^T \beta$

strength of the penalty

$$\Updownarrow$$

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

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

  - Stepwise procedures are sensitive to data perturbations and often include features with negligible improvement in fit ← greedy, ∃ backtracking alg.

  discrete

- Try new penalty: Penalize non-zero weights
  - Penalty: $\|\beta\|_1 = \sum_j |\beta_j|$
  - Leads to sparse solutions
  - Just like ridge regression, solution is indexed by a continuous param λ

LASSO Regression

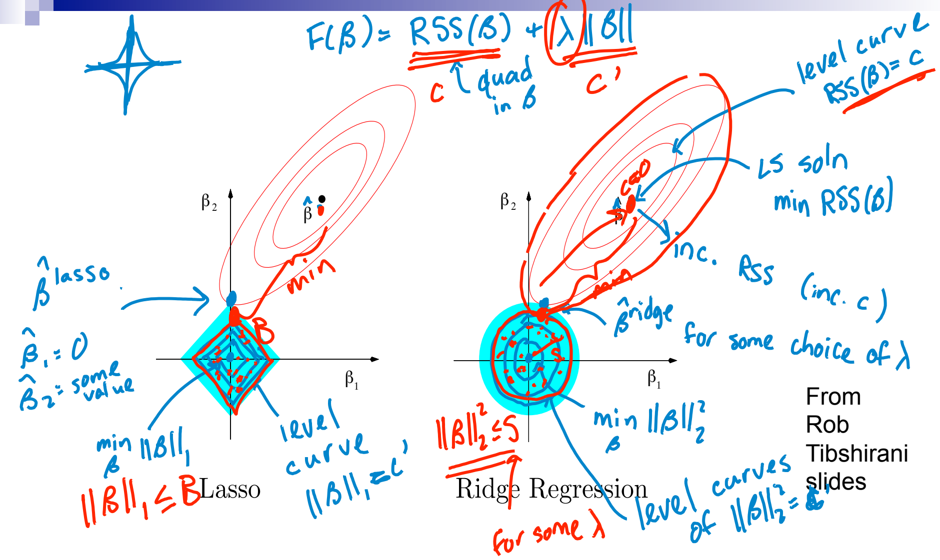
- **LASSO:** least absolute shrinkage and selection operator
- New objective:

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

$$\Updownarrow$$

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

Geometric Intuition for Sparsity



Soft Threshholding

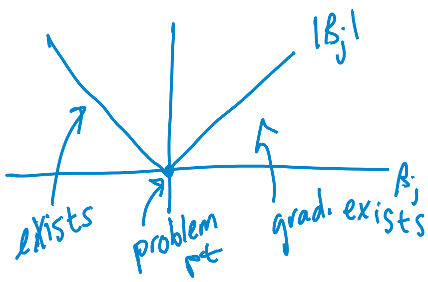
- To see why LASSO results in sparse solutions, look at conditions that must hold at optimum

  look at β_j … do this for all j ⇒ set of simult. eqns

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

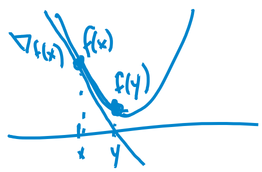
  $\sum |\beta_j|$

- Look at subgradient…



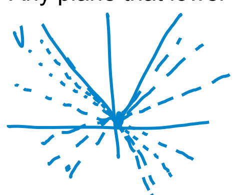
Subgradients of Convex Functions

- Gradients lower bound convex functions:



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

- Gradients are unique at **x** if function differentiable at **x**
- Subgradients: Generalize gradients to non-differentiable points:
  - Any plane that lower bounds function:



For $|\beta_j|$:
$V \in [-1, 1]$

$\boxed{\begin{array}{l} V \in \partial f(x) \text{ subgradient} \\ \text{if} \\ f(y) \geq f(x) + V(y-x) \end{array}}$

Soft Threshholding

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

- Gradient of RSS term:

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

c_j: $2\sum_{i=1}^{n} x_j^i (y_i - \beta_{-j}^T x_{-j}^i)$ all cov other than x_j; all β's except for β_j

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

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

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

residuals from model w/o jth covariate

- 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 $\beta_j < 0$

$a_j\beta_j - c_j - \lambda = 0$

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

If $\beta_j > 0$

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

If $\beta_j = 0$ $-\lambda < c_j < \lambda$ if not strong corr., $\beta_j = 0$

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

©Emily Fox 2013 11

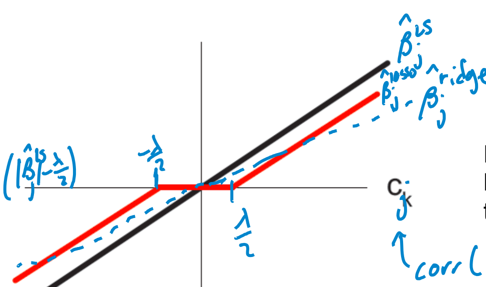
Soft Threshholding

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

If $X^TX = I$

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

$\hat\beta_j^{lasso} = \text{sign}(\hat\beta_j^{LS})(|\hat\beta_j^{LS}| - \frac{\lambda}{2})$



From Kevin Murphy textbook

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

©Emily Fox 2013 12

Coordinate Descent

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

  1-d optimization problem... Just solved this for lasso
- Coordinate descent:

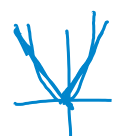
  while not converged

  Pick coord j.

  $\beta_j \leftarrow \min_b F(\beta_1, \ldots, \beta_{j-1}, b, \beta_{j+1}, \ldots, \beta_p)$
- 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:

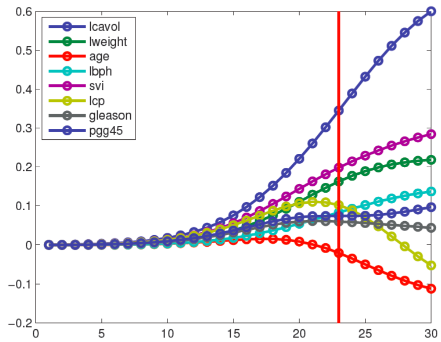
$$\hat{\beta}_j = \begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases} = \mathrm{sign}(c_j)\, \frac{(|c_j| - \lambda)_+}{a_j}$$

    - Where: (cache)

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

  - 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*

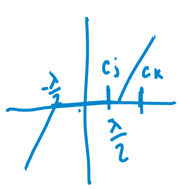


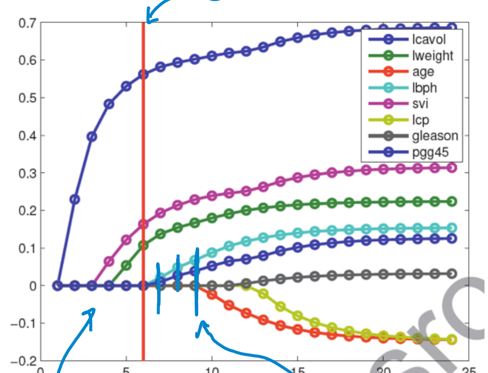
From
Kevin Murphy
textbook

- Typical approach: select λ using cross validation

Now: *LASSO Coefficient Path*

CV soln





From
Kevin Murphy
textbook

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

only a few critical values of λ where support changes

B

inc. λ

solns are sparse for any given λ

LASSO Example

$\hat{\beta}$ CV solns

Term	Least Squares	Ridge	Lasso
$\hat{\beta}_0$ Intercept	2.465	2.452	2.468
$\hat{\beta}_1$ 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	
$\hat{\beta}_p$ pgg45	0.267	0.133	

From Rob Tibshirani slides

not in the model

sparse solns

Sparsistency

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

$\hat{\theta} \rightarrow \theta^*$?

- 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

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

  - 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

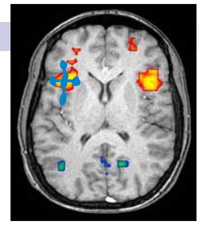
- In general, can't solve analytically for GLM (e.g., logistic reg.)
  - Gradually decrease λ and use efficiency of computing $\hat{\beta}(\lambda_k)$ from $\hat{\beta}(\lambda_{k-1})$ = warm-start strategy
  - See Friedman et al. 2010 for coordinate ascent + warm-starting strategy

- 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



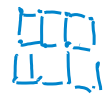
- Might want coefficients of neighboring voxels to be similar

  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

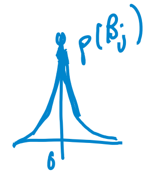
$$y_i \mid \beta \sim N(\beta_0 + x_i^T \beta, \sigma^2)$$

and prior

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

where

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



- For large λ

  more peaked around 0

  $p(\beta_j = 0) = 0$

  but look at the post. mode

- 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 prior as an alternative