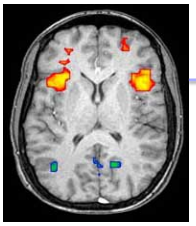
Module 1: Nonparametric Preliminaries

Review of Regression, Linear Smoothers

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 4th, 2013

fMRI Prediction Subtask

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



Linear Regression – *review*

- Model:
- ***Design matrix***:
- Rewrite in matrix form:

Linear Regression – *review*

- Least squares estimation:
  - Minimize ***residual sum of squares***
  - Take gradient and set = 0
- In Gaussian case, LS est. = maximum likelihood est.

Linear Regression – *review*

- ***Fitted values***

- Number of parameters

- For any *x*, we can write

Linear Smoothers

- Definition:
  $\hat{f}_n$ of f is a ***linear smoother*** if, for each *x*, there exists
  $$\ell(x) = (\ell_1(x), \ldots, \ell_n(x))^T$$
  such that

- Matrix form
  - Fitted values
  - Smoothing or "hat" matrix

- Effective degrees of freedom:

Linear Smoothers

- Note 1:

  A linear smoother does ***not*** imply that $f(x)$ is linear in *x*

- Note 2:

  If $Y_i = c$ for all *i*, then $\hat{f}_n(x) = c$ for all *x*

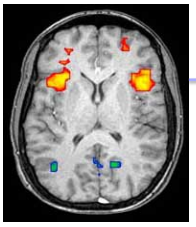
Module 1: Nonparametric Preliminaries

Overfitting, Ridge Regression, LASSO

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 4th, 2013

fMRI Prediction Subtask

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

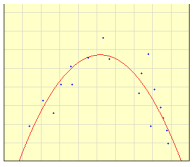


Features of word

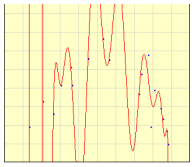
Regularization in Linear Regression

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

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



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



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

Ridge Regression

- Ameliorating issues with overfitting:
- New objective:

Ridge Regression

- New objective:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta^T x_i))^2 + \lambda ||\beta||_2^2$$

  - Reformulate:
  - Set gradient = 0

- Linear smoother!!

Ridge Regression

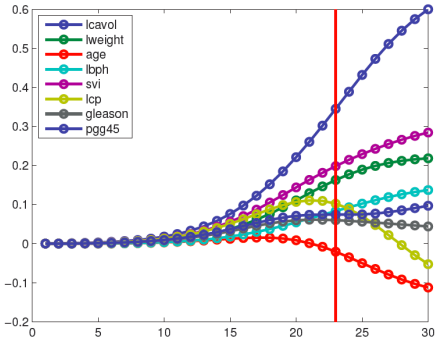
- Solution is indexed by the regularization parameter λ
- Larger λ
- Smaller λ
- As λ → 0
- As λ →∞

Shrinkage Properties

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

- If orthogonal covariates $X^T X = I$
- Effective degrees of freedom:

Ridge Coefficient Path



From
Kevin Murphy
textbook

- Typical approach: select λ using cross validation

A Bayesian Formulation

- Consider a model with likelihood

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

and prior

$$\beta \sim N\left(0, \frac{\sigma^2}{\lambda} I_p\right)$$

- For large λ

- The posterior is

$$\beta \mid y \sim N\left(\hat{\beta}^{ridge}, \sigma^2 (X^T X + \lambda I)^{-1} X^T X \sigma^2 (X^T X + \lambda I)^{-1}\right)$$

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

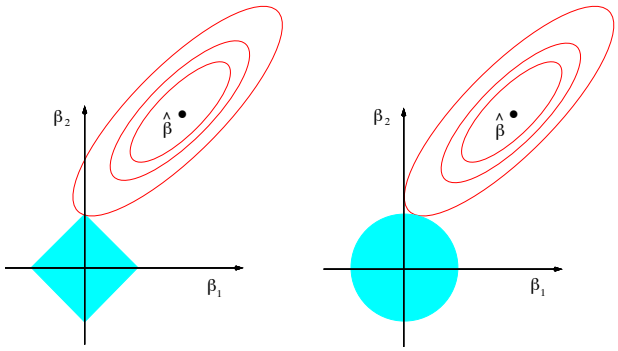
LASSO Regression

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

LASSO Solutions

- The LASSO solution is **nonlinear** in *y*…*not a linear smoother*
  - Degrees of freedom cannot be computed as before
  - Many recent studies on this (e.g., Zou et al. 2007, Tibshirani & Taylor 2011)
  - Standard errors via the bootstrap
- Efficient algorithms exist for solving
  - Will return to this in a few slides

Geometric Intuition for Sparsity



Lasso Ridge Regression

From Rob Tibshirani slides

Soft Threshholding

- To see why LASSO results in sparse solutions, look at conditions that must hold at optimum
- L_1 penalty $||\beta||_1$ is not differentiable whenever $\beta_j = 0$
- Look at subgradient…

Subgradients of Convex Functions

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

Soft Threshholding

- Gradient of RSS term:

- Subgradient of full objective:

Soft Threshholding

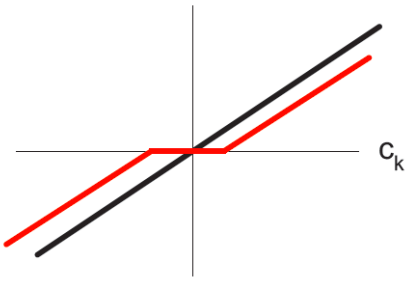
- Set subgradient = 0:

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

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

Soft Threshholding

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



From
Kevin Murphy
textbook

Coordinate Descent

- Given a function F
  - Want to find minimum
- Often, hard to find minimum for all coordinates, but easy for one coordinate
- Coordinate descent:
- How do we pick a coordinate?
- When does this converge to optimum?

Stochastic Coordinate Descent for LASSO (aka Shooting Algorithm)

- Repeat until convergence
  - Pick a coordinate *j* at random
    - Set:

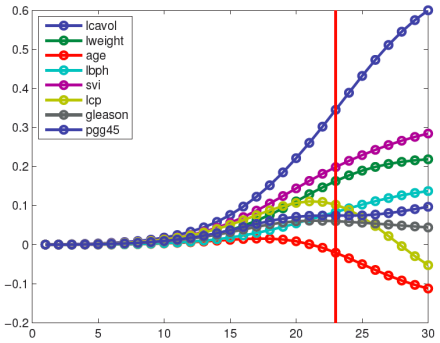
$$\hat{\beta}_j = \begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases}$$

    - Where:

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

- Other common technique = LARS
  - Least angle regression and shrinkage, Efron et al. 2004

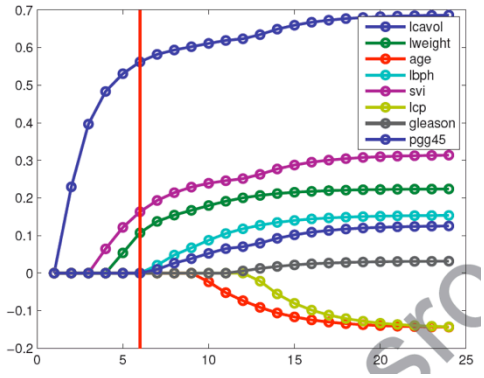
Recall: *Ridge Coefficient Path*



From Kevin Murphy textbook

- Typical approach: select λ using cross validation

Now: *LASSO Coefficient Path*



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	

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
- 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

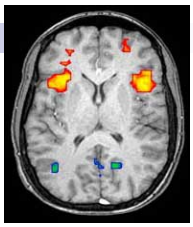
- Rescale LASSO objective by *n*:
- Theorem (Wainwright 2008, Zhao and Yu 2006, …):
  - Under some constraints on the design matrix *X*, if we solve the LASSO regression using

    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 $S(\hat{\beta}) = S(\beta^*)$

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

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:

A Bayesian Formulation

- Consider a model with likelihood

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

and prior

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

where

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

- For large λ
- LASSO solution is equivalent to the ***mode*** of the posterior
- Note: posterior mode ≠ posterior mean in this case
- There is no closed-form for the posterior. Rely on approx. methods.