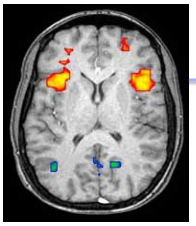
Module 1: Nonparametric Preliminaries

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



Features of word

y_i

x_i

x_{i1} … x_{ip} → y_{i1} … y_{id}

Voxels → Semantic features

Examine every output dim. ind.

Linear Regression – *review*

- Model: $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \quad i = 1, \ldots, n$

  $= x_i^T \beta + \epsilon_i$ ($\beta = (\beta_1, \ldots, \beta_p)^T$)

  $E[\epsilon_i] = 0 \quad \text{var}(\epsilon_i) = \sigma^2$

  $x_{i1} = 1$ for intercept

- ***Design matrix***:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Rewrite in matrix form:

$$Y = X\beta + \epsilon$$

Linear Regression – *review*

- Least squares estimation:
  - Minimize ***residual sum of squares***

$$\hat{\beta} = \underset{\beta}{\text{argmin}} (Y - X\beta)^T (Y - X\beta) = \sum_{i=1}^{n} (y_i - \sum_j x_{ij} \beta_j)^2$$

$\text{RSS}(\beta)$

$$\frac{1}{2}\text{RSS}(\beta) = \frac{1}{2}\beta^T (X^T X)\beta - \beta^T X^T y + \text{const.}$$ (wrt β)

  - Take gradient and set = 0

$$\nabla_\beta \frac{1}{2}\text{RSS}(\beta) = X^T X \beta - X^T y = 0$$

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

- In Gaussian case, LS est. = maximum likelihood est.

Linear Regression – *review*

- ***Fitted values***

$\hat{f}_n = X\hat{\beta}^{LS} = Ly$ $\quad L = X(X^TX)^{-1}X^T$ "hat matrix"

- Number of parameters

$p = tr(L)$ $\quad$ proof: $tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_p)$

- For any x, we can write

$\hat{f}_n(x) = \ell(x)^T y = \sum_{i=1}^n \ell_i(x) y_i$

where $\ell(x) = x(X^TX)^{-1}X^T$ ← all x_{ij}'s from training data

Linear Smoothers

- Definition:

$\hat{f}_n$ of f is a ***linear smoother*** if, for each x, there exists

$$\ell(x) = (\ell_1(x), \ldots, \ell_n(x))^T \quad \text{with} \quad \sum_{i=1}^n \ell_i(x) = 1$$

such that $\hat{f}_n(x) = \sum_{i=1}^n \ell_i(x) y_i$

- Matrix form
  - Fitted values $\hat{f} = Ly$
  - Smoothing or "hat" matrix $L_{ij} = \ell_j(x)$
- Effective degrees of freedom: $\nu = tr(L)$

Linear Smoothers

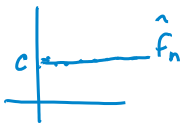
- Note 1:

  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*

bc $\sum_{i=1}^{n} \ell_i(x) = 1$



Module 1: Nonparametric Preliminaries

Overfitting, Ridge Regression, LASSO

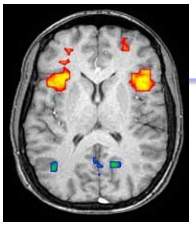
STAT/BIOSTAT 527, University of Washington

Emily Fox

April 4th, 2013

fMRI Prediction Subtask

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



Features of word

y_i

x_i

$x_{i1} \ldots x_{ip} \rightarrow y_{ij}$ semantic feature

p = # voxels = 20,000

only small n

$p \gg n$

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

$(X^T X)^{-1}$: $p \times p$, rank deficient

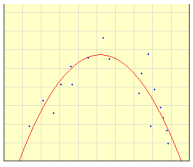
X^T: $p \times n$

y: n

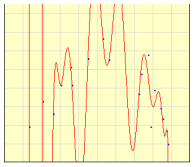
Regularization in Linear Regression

- 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 + …



even for $n \approx p$, 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:

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

don't penalize the intercept; $\beta^T\beta$; strength of the penalty

$$\Updownarrow$$

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

Ridge Regression

- New objective:

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

$F(\beta)$; $RSS(\beta)$ as before

  - Reformulate:

$$\frac{1}{2}F(\beta) = \frac{1}{2}\beta^T(X^TX)\beta - \beta^T X^T y + \text{const.} + \frac{1}{2}\lambda\beta^T\beta$$

as before

$$= \frac{1}{2}\beta^T(X^TX + \lambda I)\beta - \beta^T X^T y + \text{const.}$$

  - Set gradient = 0

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

- Linear smoother!!

$$\hat{f}_n^{ridge} = X\hat{\beta}^{ridge} = Ly \qquad L = X(X^TX + \lambda I)^{-1}X^T$$

Ridge Regression

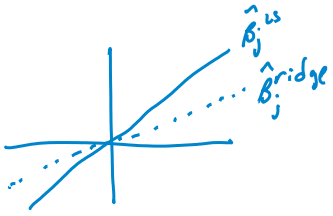
- Solution is indexed by the regularization parameter λ
- Larger λ high reg.
- Smaller λ low reg.
- As λ → 0 $\hat{\beta}^{ridge} \rightarrow \hat{\beta}^{LS}$
- As λ →∞ $\hat{\beta}^{ridge} \rightarrow 0$

Shrinkage Properties

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

- If orthogonal covariates $X^T X = I$

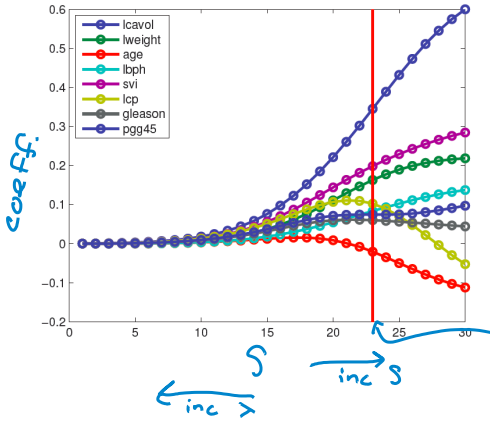
$$\hat{\beta}_j^{ridge} = \frac{\hat{\beta}_j^{LS}}{1+\lambda} \quad \forall j$$



- Effective degrees of freedom:

$$\nu = tr(L) = \left(X\left(X^T X + \lambda I\right)^{-1} X^T\right)$$

Ridge Coefficient Path



From Kevin Murphy textbook

$\|\beta\|_2^2 \leq s$

- Typical approach: select λ using cross validation (CV)

A Bayesian Formulation

- Consider a model with likelihood

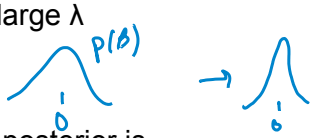
If $\epsilon \sim N(0,\sigma^2)$

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

and prior

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

- For large λ



prior is peaked around $\beta = 0$

$\Leftrightarrow$ penalizing β far from 0

- The posterior is

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

↑ easy to show

$\mathrm{Var}(\hat{\beta}^{ridge})$

Variable Selection

- 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.
- 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

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

discrete

LASSO Regression

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

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

$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta^T x_i)\right)^2 = RSS(\beta)$

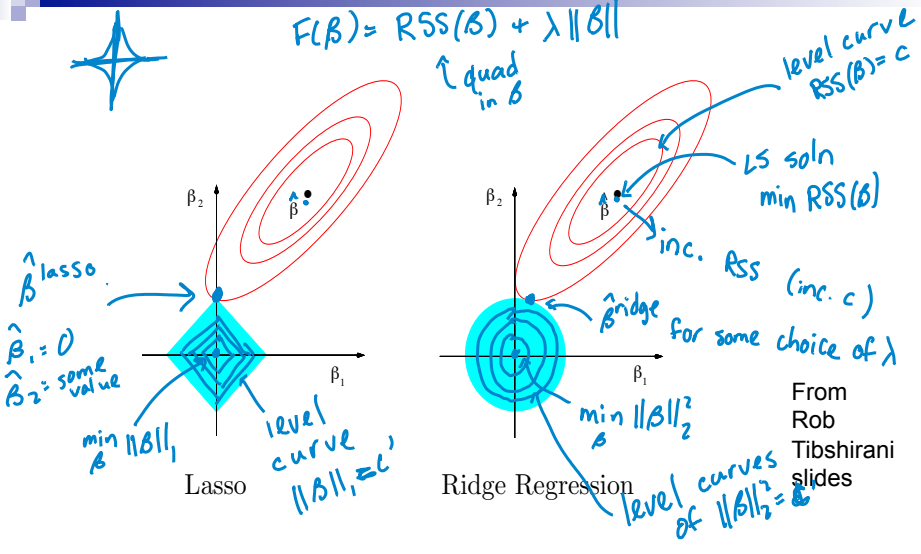
$\Updownarrow$

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

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



From Rob Tibshirani slides