

# Lecture 6

## Non-parametric Regression and Shrinkage Methods

Reading: JWHT Chapters 7 and 6.2; Faraway, 2014 Chapter  
11.3-11.4

*DSA 8020 Statistical Methods II*

February 8-12, 2021

Whitney Huang  
Clemson University

# Agenda

Non-parametric  
Regression and  
Shrinkage Methods



Non-parametric  
Regression

Ridge Regression

LASSO

1 Non-parametric Regression

2 Ridge Regression

3 LASSO

## Moving Away From Linear Regression

- We have mainly focused on **linear regression** so far

**Model:**  $Y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$

**Data:**  $\mathbf{y}$  (response vector);  $\mathbf{X}$  (design matrix)

- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}; \hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\mathbf{H}: \text{"Hat" matrix}} \mathbf{y}$

- $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$

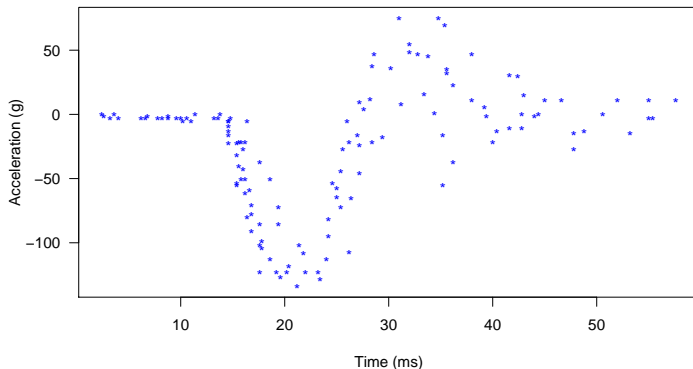
- In this lecture we are going to discuss **non-parametric regression modeling**

**Model:**  $Y = f(\mathbf{x}) + \varepsilon \Rightarrow E[Y|\mathbf{x}] = f(\mathbf{x})$

- The (smooth) function  $f(\mathbf{x})$  must be represented somehow
- The degree of smoothness of  $f(\mathbf{x})$  must be made controllable
- Some means for estimating the most appropriate degree of smoothness from data is required

# Non-parametric Regression

Let's use the motor-cycle impact data as an illustrative example. This data set is taken from a simulated motor-cycle crash experiment in order to study the efficacy of crash helmets.



# Representing a Smooth Function using Basis Functions

Non-parametric  
Regression and  
Shrinkage Methods



Non-parametric  
Regression

Ridge Regression

LASSO

- Basis function representation:  $f(x) = \sum_{j=1}^K b_j(x)\beta_j$

# Representing a Smooth Function using Basis Functions

Non-parametric  
Regression and  
Shrinkage Methods



Non-parametric  
Regression

Ridge Regression

LASSO

- Basis function representation:  $f(x) = \sum_{j=1}^K b_j(x)\beta_j$
- There are many basis functions to choose from:  
Polynomials, Fourier Series, Radial Basis Functions...

# Representing a Smooth Function using Basis Functions

Non-parametric  
Regression and  
Shrinkage Methods



Non-parametric  
Regression

Ridge Regression

LASSO

- Basis function representation:  $f(x) = \sum_{j=1}^K b_j(x)\beta_j$
- There are many basis functions to choose from:  
Polynomials, Fourier Series, Radial Basis Functions...
- We are going to focus on **Splines**: piecewise polynomials joined together to make a single smooth curve

# An Example of a Cubic Spline Function

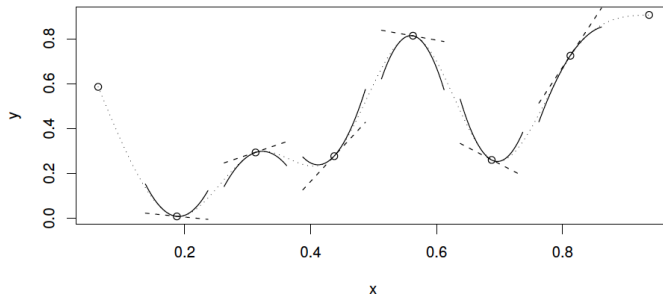


Figure 3.3 A cubic spline is a curve constructed from sections of cubic polynomial joined together so that the curve is continuous up to second derivative. The spline shown (dotted curve) is made up of 7 sections of cubic. The points at which they are joined ( $\circ$ ) (and the two end points) are known as the knots of the spline. Each section of cubic has different coefficients, but at the knots it will match its neighbouring sections in value and first two derivatives. Straight dashed lines show the gradients of the spline at the knots and the curved continuous lines are quadratics matching the first and second derivatives at the knots: these illustrate the continuity of first and second derivatives across the knots. This spline has zero second derivatives at the end knots: a 'natural spline'.

**Source:** Simon Wood, *Generalized Additive Models*, p. 122, Fig. 3.3



- Choose  $K$  knot points to partition the range of  $x$  to form the spline basis  $X$
- Techniques from linear regression can be used to carry out estimation and inference
- However, the model fit tends to depend strongly on  $K$ , the number of knots, and  $\{\xi_k\}_{k=1}^K$ , the knot locations
  - Few knots: Resulting class of functions may be too restrictive (**bias**)
  - Many knots: We run the risk of overfitting (**variance**)

- Regression splines are not truly “nonparametric” as the choices regarding  $K$  and  $\{\xi_k\}_{k=1}^K$  are fundamentally parametric choices and have a large effect on the fit
- Model selection (i.e, choosing the degree of smoothing) is not straightforward
- An alternative approach to controlling smoothness is **penalization**

$$\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int [f''(x)]^2 dx$$

- The first term captures the fit to the data, while the second penalizes curvature
- $\lambda$  is the **smoothing parameter**, and it controls the tradeoff between the two terms:
  - $\lambda = 0$  imposes no restrictions and  $f$  will therefore interpolate the data
  - $\lambda = \infty$  returning us to ordinary linear regression

Selecting an appropriate  $\lambda$  is crucial

# Natural Cubic Splines Solve the Penalized Least Squares!

Non-parametric  
Regression and  
Shrinkage Methods



Non-parametric  
Regression

Ridge Regression

LASSO

**Theorem:** Out of all twice-differentiable functions, the one that minimizes

$$\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int [f''(x)]^2 dx$$

is a natural cubic spline with knots at every unique value of  $\{x_i\}$

This penalized approach leads to the framework of [smoothing splines](#), introduced by [Grace Wahba](#) to statisticians

## Smoothing Splines (Cont'd)

Let  $\{N_j\}_{j=1}^n$  denote the collection of natural cubic spline basis functions and  $N$  denote the  $n \times n$  design matrix consisting of the basis functions evaluated at  $\{x_i\}$ :

- $f(x) = \sum_{j=1}^n N_j \beta_j$ , where  $N_{ij} = N_j(x_i) \Rightarrow f(\mathbf{x}) = N\boldsymbol{\beta}$
- We can show that the objective function for penalized splines is

$$(\mathbf{y} - N\boldsymbol{\beta})^T (\mathbf{y} - N\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta},$$

where  $\Omega_{jk} = \int N_j''(x) N_k''(x) dx$

- The minimizer is

$$\hat{\boldsymbol{\beta}} = (N^T N + \lambda \boldsymbol{\Omega})^{-1} N^T \mathbf{y}$$

- From last slide we have

$$\hat{\beta} = (N^T N + \lambda \Omega)^{-1} N^T y$$

- Therefore we have

$$\hat{y} = \hat{f}(x) = N (N^T N + \lambda \Omega)^{-1} N^T y = L_{\lambda} y,$$

$\Rightarrow$  a linear smoother

- $tr(L_{\lambda})$  is a measure of the effective number of degrees of freedom

## Choosing $\lambda$ by Cross-Validation (CV)

### Main idea:

Sequentially leave each observation out and predict it using the rest of the data. Find the  $\lambda$  that gives the best out of sample predictions.

- CV residual:

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{(1 - \mathbf{L}_{\lambda,i,i})}$$

- $CV(\lambda)$ :

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{-i})^2 = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{(1 - \mathbf{L}_{\lambda,i,i})^2}$$

- Generalized Cross-Validation (GCV):

$$\frac{1}{n} \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{(1 - \frac{\text{tr}(\mathbf{L}_{\lambda})}{n})^2}$$

# Non-parametric Regression Fits

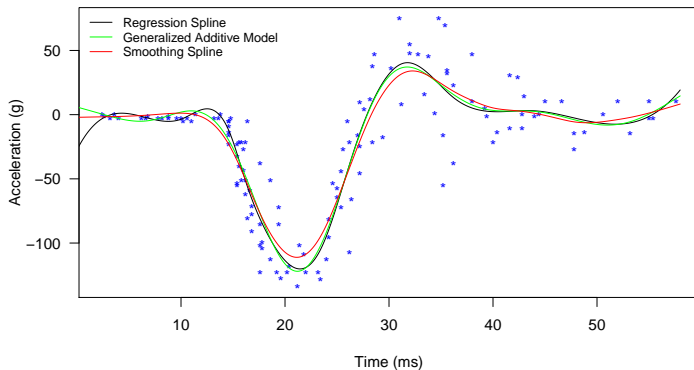
Non-parametric  
Regression and  
Shrinkage Methods



Non-parametric  
Regression

Ridge Regression

LASSO





$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$x_1, x_2, \dots, x_{p-1}$  are the predictors.

**Question:** What if we have too many predictors (i.e.,  $p$  is “large”)?

- Explanation can be difficult due to collinearity
- Can lead to overfitting by using too many predictors

We will look at two methods, namely **Ridge regression** and **LASSO**, that allow us to “shrink” the information contained in all the predictors into a more useful form

Ridge regression assumes that the regression coefficients (after normalization) should not be very large

- The ridge regression estimate chooses the  $\beta$  that minimizes:

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2,$$

where  $\lambda \geq 0$  is a **tuning parameter** to be determined via cross-validation

- The ridge regression estimates:

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

- Ridge regression is particularly effective when the model matrix is collinear

# Graphical Illustration of Ridge Regression

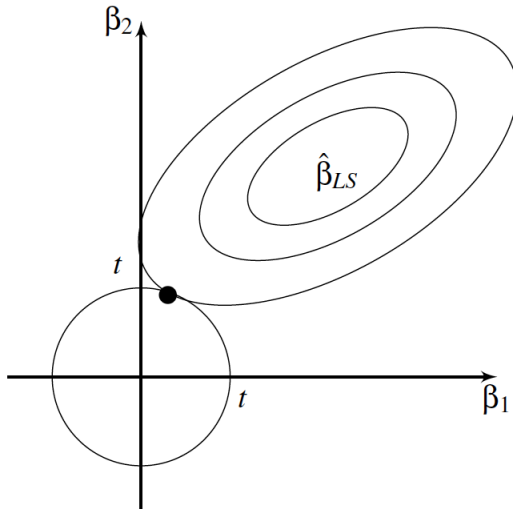
Non-parametric  
Regression and  
Shrinkage Methods

CLEMSON  
UNIVERSITY

Non-parametric  
Regression

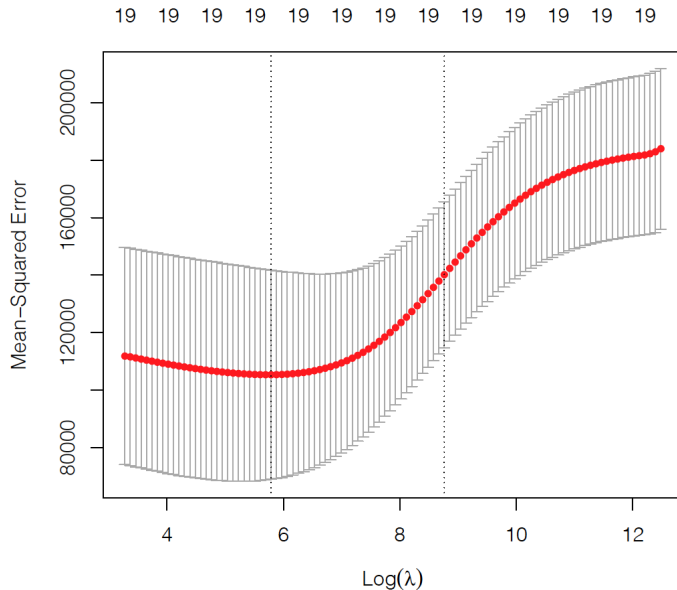
Ridge Regression

LASSO



**Source:** p. 175, Fig. 11.9 *Linear Models with R*, Faraway, 2014

## Choosing $\lambda$ via Cross-Validation



# Least Absolute Shrinkage and Selection Operator (LASSO)

Tibshirani, 1996

Non-parametric  
Regression and  
Shrinkage Methods

CLEMSON  
UNIVERSITY

Non-parametric  
Regression

Ridge Regression

LASSO

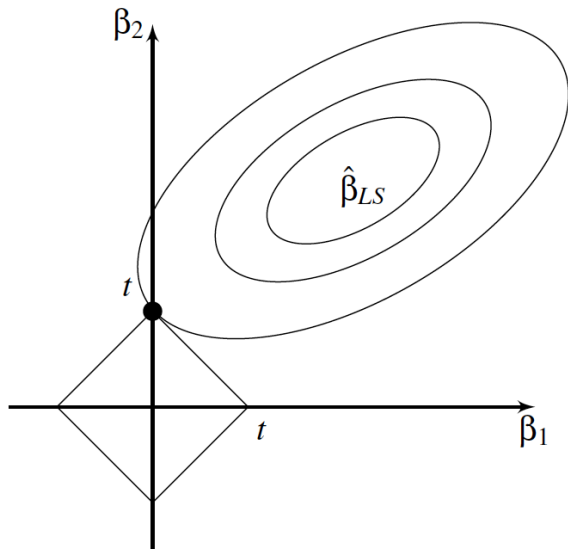
LASSO assumes the effects are **sparse** in that the response can be explained by a small number of predictors with the rest having no effect

- LASSO choose  $\hat{\beta}$  to minimize:

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p-1} |\beta_j|$$

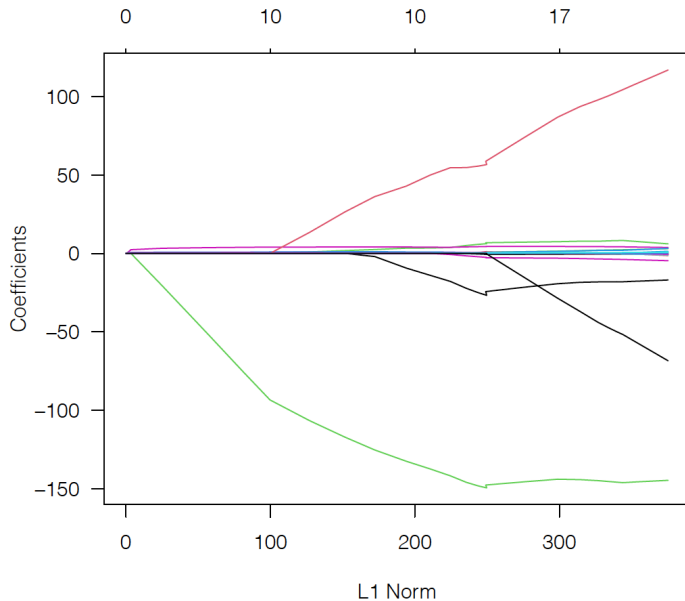
- No explicit solution to this minimization problem
- The penalty term has the effect of forcing some of the coefficient estimates to be zero when the tuning parameter  $\lambda$  is “large”  $\Rightarrow$  performs **shrinkage** and **variable selection**

## Graphical Illustration of LASSO



**Source:** p. 175, Fig. 11.9 *Linear Models with R*, Faraway, 2014

# LASSO Path



# Selecting $\lambda$ via Cross-Validation

