# Mathematical techniques in data science

Lecture 12: Shrinkage methods

March 15th, 2019

# Schedule

| Week | Chapter |
|---|---|
| 1 | Chapter 2: Intro to statistical learning |
| 3 | Chapter 4: Classification |
| 4 | Chapter 9: Support vector machine and kernels |
| 5, 6 | Chapter 3: Linear regression |
| 7 | Chapter 8: Tree-based methods + Random forest |
| 8 | |
| 9 | Neural network |
| 12 | PCA $\rightarrow$ Manifold learning |
| 11 | Clustering: K-means $\rightarrow$ Spectral Clustering |
| 10 | Bootstrap + Bayesian methods + UQ |
| 13 | Reinforcement learning/Online learning/Active learning |
| 14 | Project presentation |

# Chapter 3 & 6: Topics on Linear regression

- Linear regression
- Subset selection
- Shrinkage methods

Note: Homework 2 is uploaded. Due on 03/29 at 5pm.

# Linear model: settings

- Linear model

$$Y = \beta^{(0)} + \beta^{(1)} X^{(1)} + \beta^{(2)} X^{(2)} + \dots \beta^{(p)} X^{(p)} + \epsilon$$

- Equivalent to

$$\mathbf{Y} = \mathbf{X}\beta, \qquad \beta = \begin{bmatrix} \beta^{(0)} \\ \beta^{(1)} \\ \dots \\ \beta^{(n)} \end{bmatrix}$$

- Least squares regression

$$\hat{\beta}^{LS} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

# $\ell_0$ regularization

- $\ell_0$ regularization

$$\hat{\beta}^0 = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{i=1}^{p} \mathbf{1}_{\beta^{(i)} \neq 0}$$

  where $\lambda > 0$ is a parameter
- pay a fixed price $\lambda$ for including a given variable into the model
- variables that do not significantly contribute to reducing the error are excluded from the model (i.e., $\beta_i = 0$)
- problem: difficult to solve (combinatorial optimization). Cannot be solved efficiently for a large number of variables.

# $\ell_2$ (Tikhonov) regularization

- Ridge regression/ Tikhonov regularization

$$\hat{\beta}^{RIDGE} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} [\beta^{(j)}]^2$$

  where $\lambda > 0$ is a parameter
- shrinks the coefficients by imposing a penalty on their size
- penalty is a smooth function.
- easy to solve (solution can be written in closed form)
- can be used to regularize a rank deficient problem ($n < p$)

# $\ell_2$ (Tikhonov) regularization

$$\frac{\partial \left( \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|^2 \right)}{\partial \beta} = 2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) + 2\lambda\beta$$

- The critical point satisfies

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^T\mathbf{Y}$$

- Note: $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is positive definite, and thus invertible
- Thus

$$\hat{\beta}^{RIDGE} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

# Compare to classical inverse problem

- Typical inverse problem

$$\min_u \|F(u) - G\|_2^2 + \lambda \|u\|_2^2$$

  At some point we need to let $\lambda \to 0$.

- Ridge regression

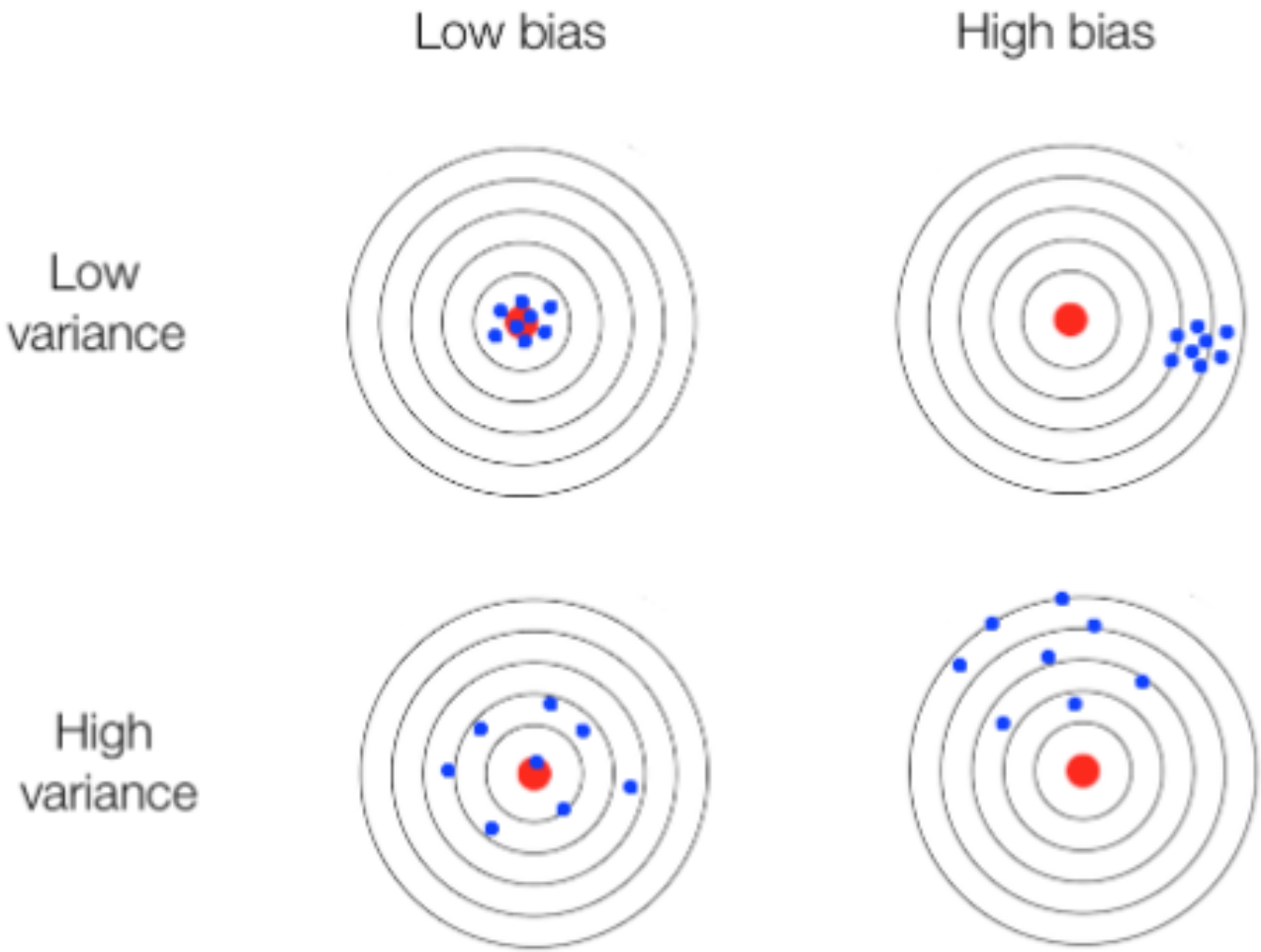
$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

  What we really want to maximize is

$$\mathbb{E}_{(X,Y)\sim P}[\|Y - X\beta\|^2]$$

  We may keep $\lambda$ away from 0.

# Bias-variance decomposition




$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (bias)^2$$

# Stein's phenomenon

- Given i.i.d. $X_1, \ldots, X_n$ samples from $\mathcal{N}(\mu, I_p)$ $(p \geq 3)$, we wish to estimate $\mu$
- The accuracy of an estimator is measured by the risk function

$$MSE(\hat{\mu}) = E[\|\hat{\mu} - \mu\|^2]$$

- The standard estimate is

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

which minimizes

$$\min_c \sum_{i=1}^{n} \|X_i - c\|^2$$

# Stein's phenomenon

- The standard estimate is

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

which minimizes

$$\min_c \sum_{i=1}^{n} \|X_i - c\|^2$$

- James-Stein's estimator

$$\mu^{JS} = \left(1 - \frac{p-2}{n\|\bar{X}\|^2}\right)\bar{X}$$

is a strictly better estimator than the sample mean $\bar{X}$

# $\ell_2$ (Tikhonov) regularization

$$\hat{\beta}^{RIDGE} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

- When $\lambda > 0$, the estimator is defined even when $n < p$
- When $\lambda = 0$ and $n > p$, we recover the usual least squares solution

# $\ell_2$ (Tikhonov) regularization

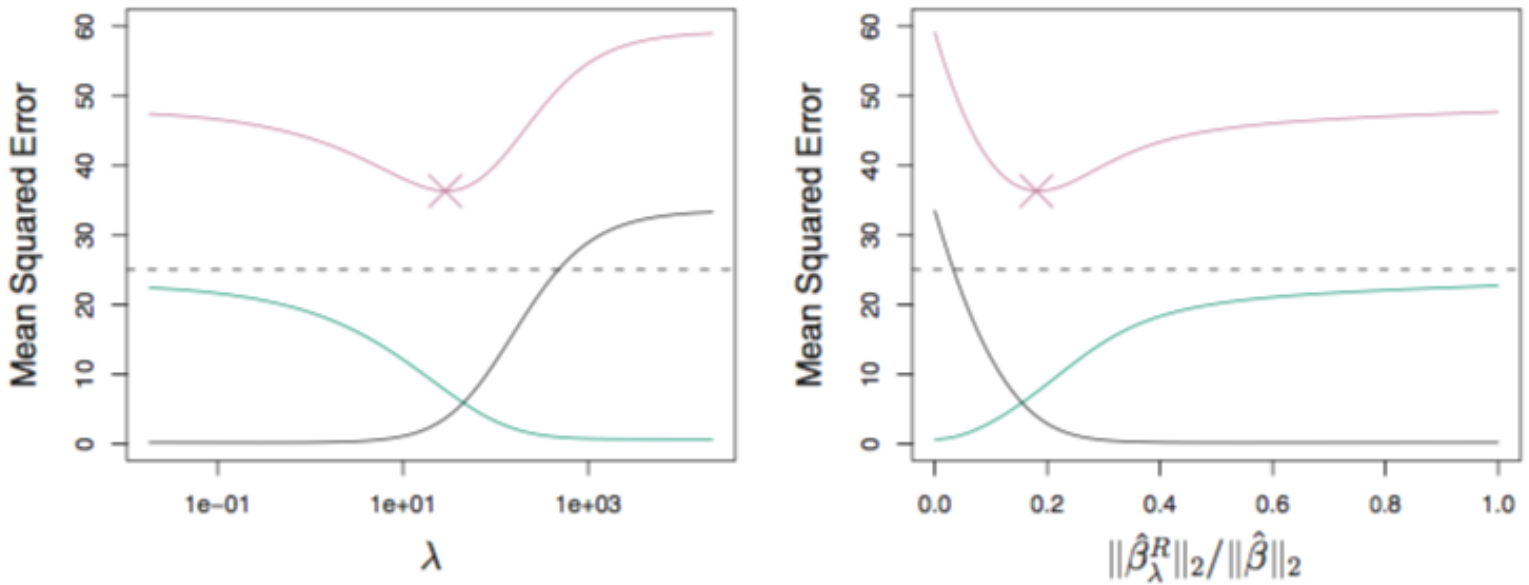



**FIGURE 6.5.** *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* $\lambda$ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

# The Lasso

# Lasso

- The Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}^{lasso} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta^{(j)}|$$

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero
- However, the $\ell_1$ penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when $\lambda$ is sufficiently large
- the lasso performs variable selection $\rightarrow$ models are easier to interpret

# Lasso: alternative form

Alternative form of lasso (using the Lagrangian and min-max argument)

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta^{(j)}| \leq s$$

# Lasso: alternative form

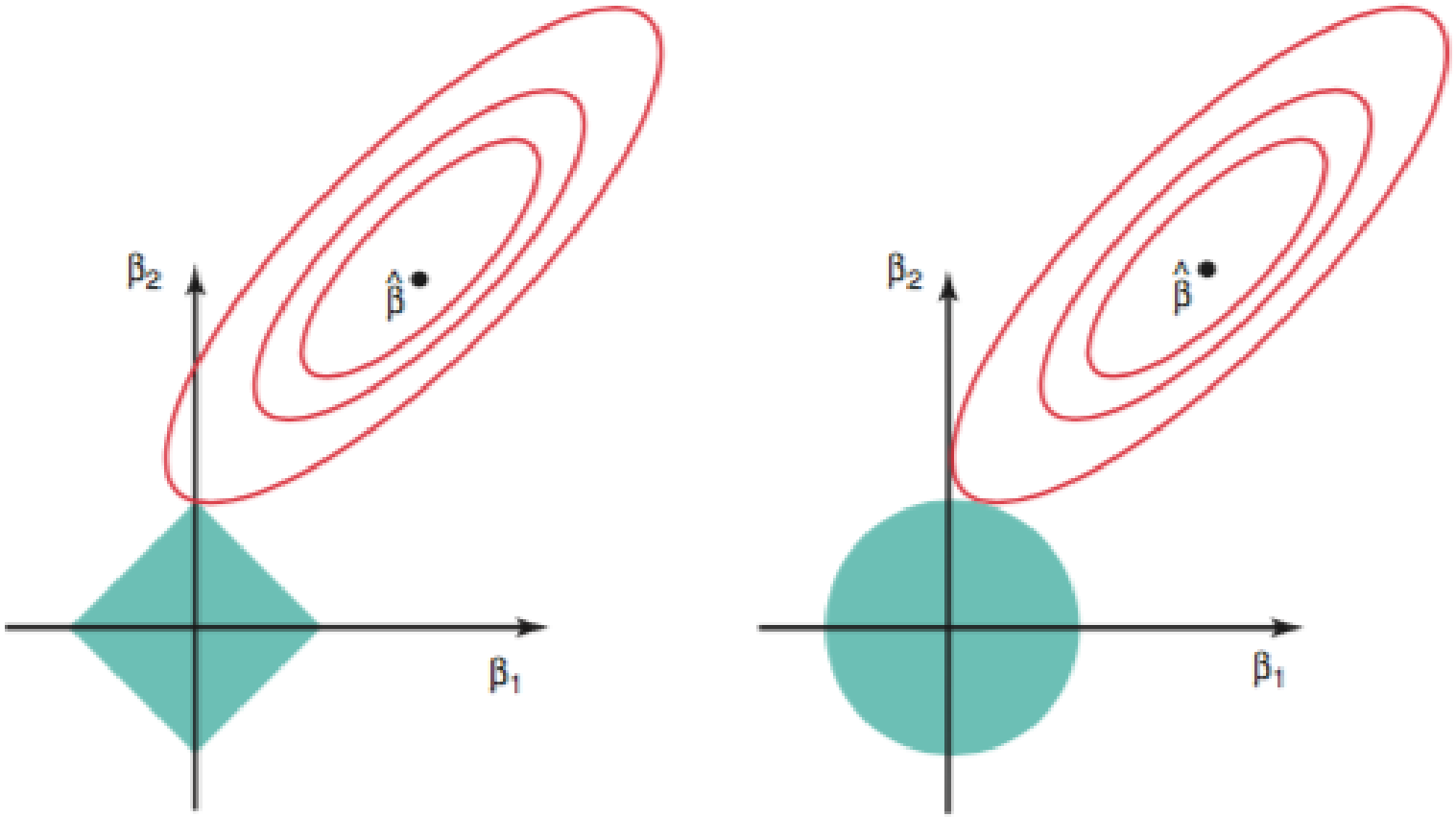



**FIGURE 6.7.** *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

# Lasso

- The Lasso:

$$\hat{\beta}^{lasso} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta^{(j)}|$$

- More "global" approach to selecting variables compared to previously discussed greedy approaches
- Can be seen as a convex relaxation of the $\hat{\beta}^0$ problem
- No closed form solution, but can solved efficiently using convex optimization methods.
- Performs well in practice
- Very popular. Active area of research

# Other shrinkage methods

- $\ell_q$ regularization ($q \geq 0$):

$$\hat{\beta} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} [\beta^{(j)}]^q$$

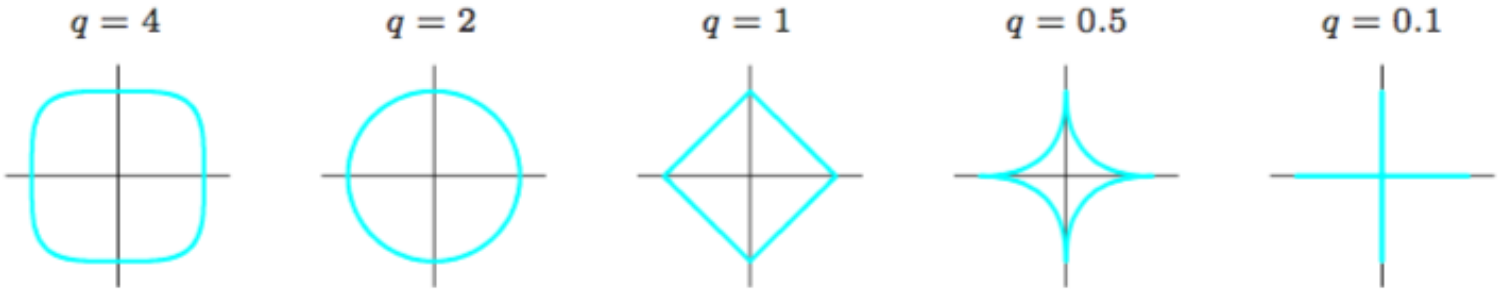



**FIGURE 3.12.** *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* $q$.

# Other shrinkage methods

- Elastic net

$$\lambda \sum_{j=1}^{p} \alpha[\beta^{(j)}]^2 + (1-\alpha)|\beta^{(j)}|$$

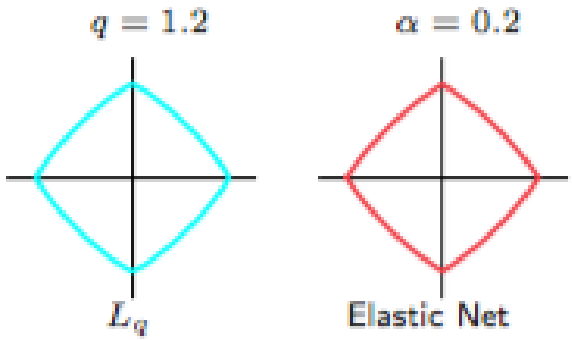



**FIGURE 3.13.** *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

# Other methods

- Least angle regression (LAR)
- The Dantzig Selector
- The grouped lasso