[redacted] Regression

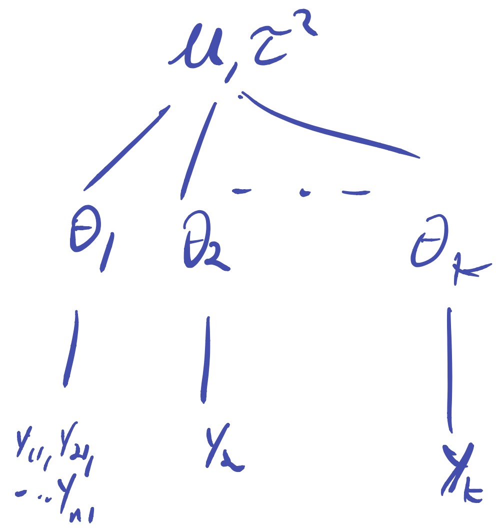
Professor Alexander Franks

2020-12-07

$$Y_j \sim N(\theta_j, \sigma_j^2)$$

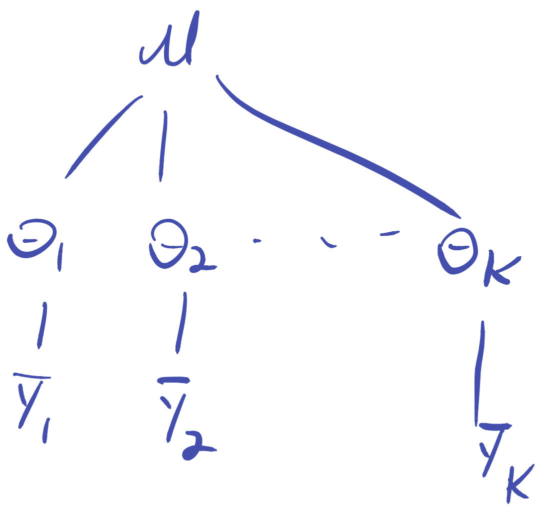
$$\theta_j \sim N(\mu, \tau^2)$$

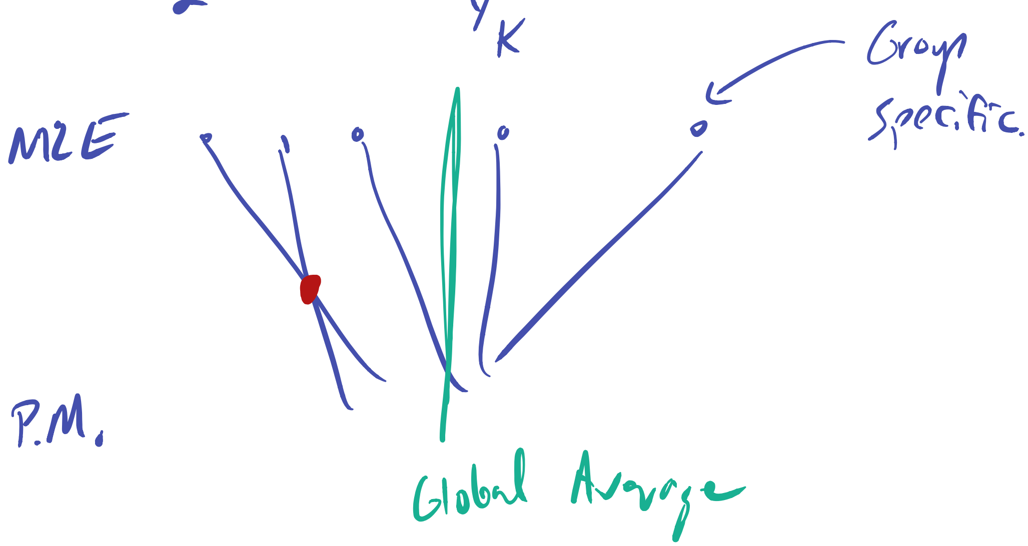
$$\mu, \tau^2 \sim \text{prior}$$



$$Y_{1j}, Y_{2j}, \dots Y_{nj} \sim N(\theta_j, \sigma_j^2)$$

$$\bar{Y}_j \sim N\left(\theta_j, \frac{\sigma^2}{n_j}\right)$$





Shrin~~a~~kge Methods

- Bias-variance trade-off has been an essential concept in this course
- One way to control bias/variance is to *penalize* model complexity

- Shrinkage methods ~~peanlize~~ penalize large values to reduce variance
  - Usually add bias (it's a trade-off afterall)
  - Frequentnists think of this as a "regularizer" or penalty
  - Bayesian think of this as a prior distribution!

Polynomial regression

- Polynomial regression: $Y = \beta_0 + \beta_1 X + \ldots + \beta_p X^p + \epsilon$
- If $p = n$, the polynomial regression will perfectly fit the training data perfectly
- Large p means higher variance, but lower bias
- High variance can manifest itself in terms of very large coefficients β

Regularization (AKA "shrinkage")

- The idea behind regularization is to reduce variance by "shrinking"" coefficients toward 0
- Keep all p predictors but constrain complexity of model fit
- Two common examples from regression
  - "ridge" penalty
  - "lasso" penalty

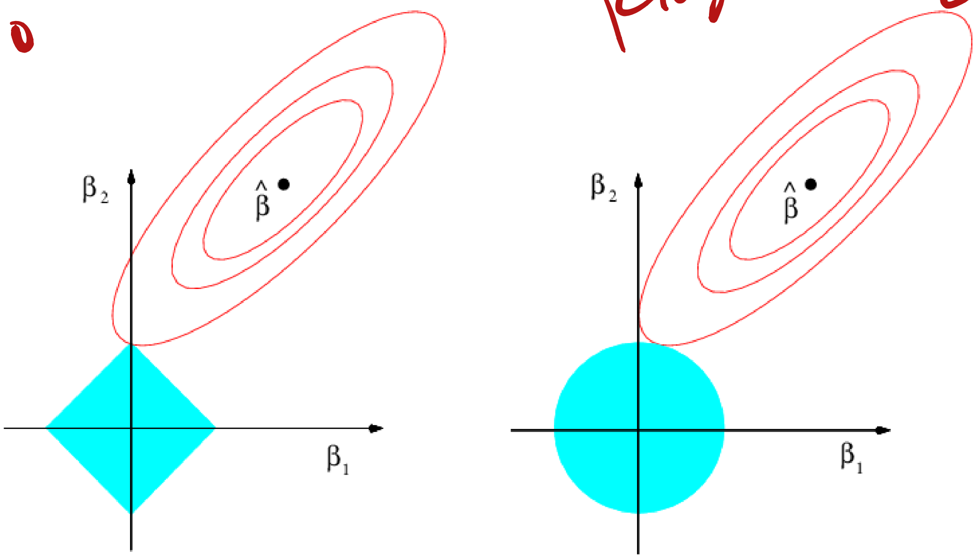
Regularized models

Sum squared Errors Penalty

- Ridge: $\min_{\beta} \left[\text{error}(\beta, X, Y) + \lambda \sum \beta_i^2\right]$
- Lasso: $\min_{\beta} \left[\text{error}(\beta, X, Y) + \lambda \sum |\beta_j|\right]$

Lasso Ridge

Contours of the likelihood (or s.s.e)



Polynomial regression with ridge penalty

- Here we will assume the polynomial order p is fixed. We are not selecting p

- Rather than select p to control overfitting, constrain the coefficients β

- Minimize

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

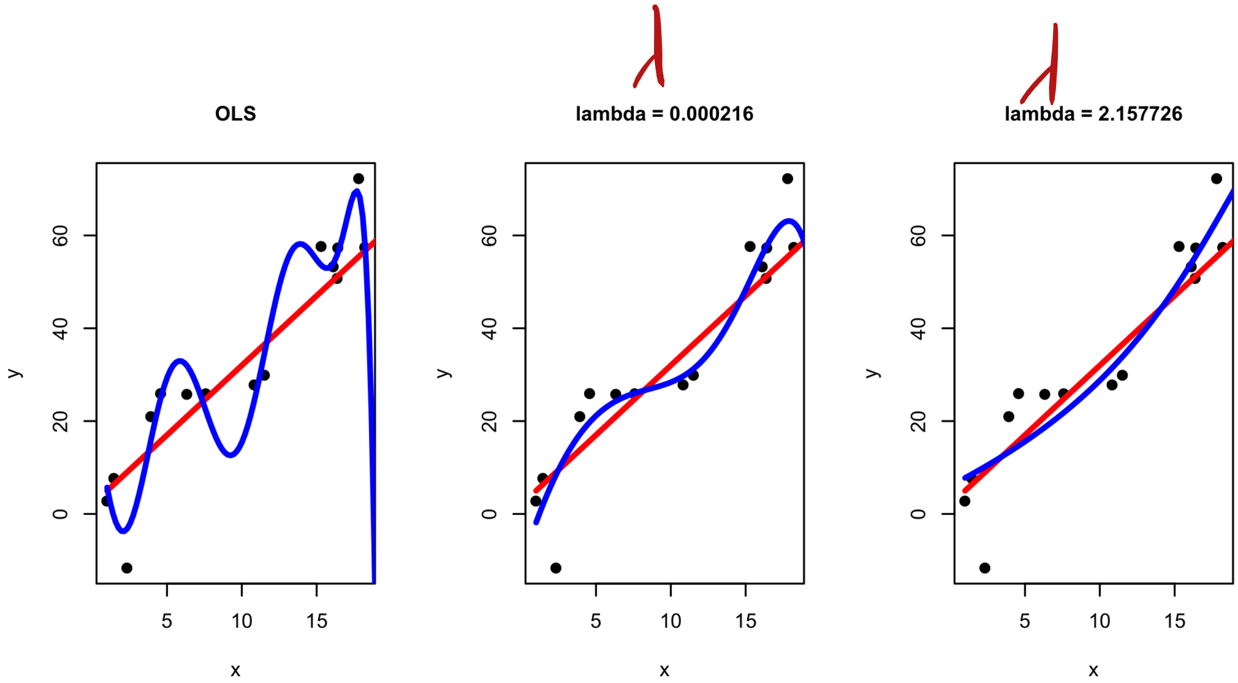
- $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x^j)^2$ is the usualize OLS objective

- $\lambda \sum_{j=1}^{p} \beta_j^2$ is the "ridge penalty" and λ is the tuning parameter determining the strength of the penalty

A simple example

- $Y = 3X + 2 + \epsilon$
- $\epsilon \sim N(0, 10)$
- Generate 10 random observations from this model
- Fit a 9th order polynomial, e..g include predictors $(x, x^2, \ldots x^9)$
- True model can be expressed as 9-th order polynomial with $(\beta_0, \beta_1, \ldots, \beta_9) = (2, 3, 0, 0, \ldots, 0)$

Ridge regression fit



$$Y_i | X_i \sim N\Big(\sum_j B_j X_{ij},\ 1\Big)$$

Linear Regression

$$E[Y] = B_1 X_1 + B_2 X_2 \dots + \varepsilon$$

$$B_j \sim N(0, \tau^2)$$

prior

$$P(B_1, B_2, \dots B_K \mid Y_1, \dots Y_n, X) \propto \underbrace{\frac{1}{\sqrt{2\pi}}\, e^{-\frac{\sum (Y_i - BX_i)^2}{2}}}_{\text{lik.}} \times \underbrace{\prod_j^K \frac{1}{\sqrt{2\pi\tau^2}}\, e^{-\frac{(B_j - 0)^2}{2\tau^2}}}_{\text{prior}}$$

$$-\log P(B_1, \dots B_K \mid Y, X) = +\underbrace{\sum (Y_i - BX_i)^2}_{\text{OLS obj.}} + \frac{1}{\tau^2}\underbrace{\sum_j B_j^2}_{\text{Ridge Penalty}}$$

$$\frac{1}{\tau^2} = \lambda$$

Posterior Mean:

$$E[\theta \mid y]$$

Post. Med.

$$\text{med}(\theta \mid y)$$

Posterior Mode:

$$\operatorname{argmax} P(\theta \mid y)$$

Mode

Lasso penalty

- Another popular alternative to ridge regression is the "LASSO"
- Minimize

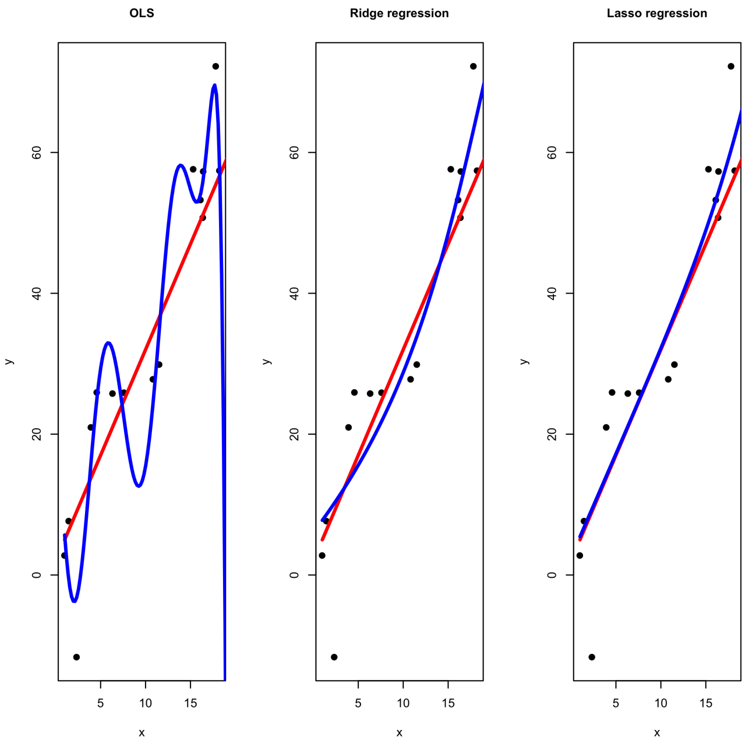
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j) + \lambda \sum_{j=1}^{p} |\beta_j|$$

- $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j)$ is the usualize OLS objective
- $\lambda \sum_{j=1}^{p} |\beta_j|$ is the "lasso penalty"
- Lasso constrains the sum of the absolute values of the coefficients
- Contrast: ridge constrains the sum of squared values of the coefficients

Lasso penalty

- Coefficients estimated with lasso have a lot more true 0's than the ridge penalty
- This is useful when many of them may not be relevant for predicting the outcome
- This is called "sparsity". Lasso estiamtes are sparse.
- A tool for variable selection

Comparing ridge and lasso



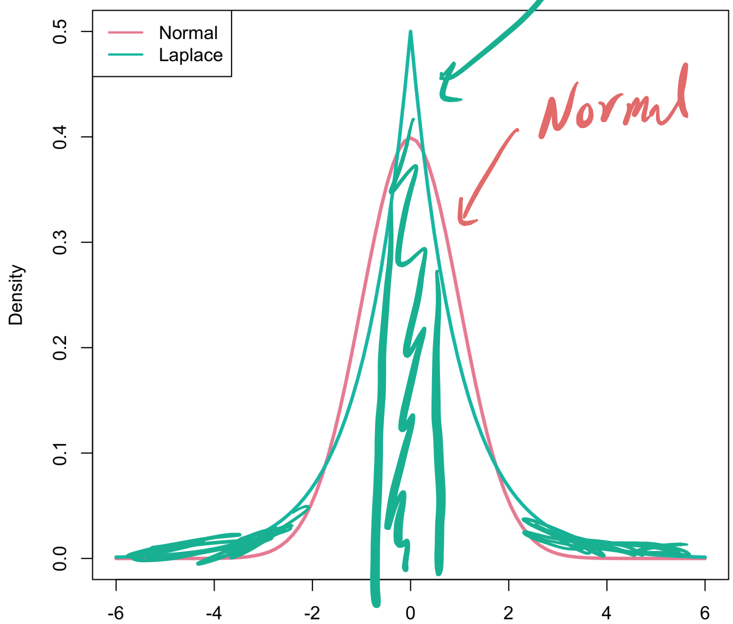
Truth is sparse so lasso works particularly well!

Laplace Random Variables

center.

Laplace density: `$p(y \mid \theta) = \frac{1}{2} e^{-|y-\theta|}$`

Laplace



$$Y|X \sim N(BX, 1)$$

Lik.

$$B_j \sim \text{Laplace}(0, \tau^2)$$

$\to$ OLS

Prior

$$P(B_1, \dots B_K) \propto \prod_j^K e^{-\frac{1}{\tau^2}|B_j|}$$

log Poster

$$\to - \log P(B_1, \dots B_K | Y, X) = \text{OLS} + \frac{1}{\tau^2} \sum |B_j|$$

LASSO = Posterior Mode of Bayesian Regression w/ Laplace Priors

"Inductive Bias" =

Model Assumptions + Prior