DSO530 Statistical Learning Methods

Lecture 6 part II : Shrinkage Methods

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California

Bias-variance trade-off

- Suppose the true relation between x and y is

$$y = f(x) + \varepsilon$$

- Based on training set $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, we construct $\hat{f}$ to estimate f
- for a new pair of observation (x_0, y_0), we predict y_0 using $\hat{f}(x_0)$, and the discrepancy is $y_0 - \hat{f}(x_0)$
- Then the expected test MSE at x_0 is

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon)$$

  - where $Var(\hat{f}(x_0)) = E[\hat{f}(x_0) - E(\hat{f}(x_0))]^2$
  - and $Bias(\hat{f}(x_0)) = E[\hat{f}(x_0) - f(x_0)]$

- The overall expected test MSE is computed by averaging $E\left(y_0 - \hat{f}(x_0)\right)^2$ over all possible values of x_0 in the test set.

Shrinkage Methods

- We can fit a model containing all p predictors using a technique that *constrains* or *regularizes* the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- Shrinking the coefficient estimates can significantly reduce their variance (with some cost in bias).
- Best known shrinkage methods: ridge regression and the lasso.
- The ridge regression coefficient estimates $\hat{\beta}^R_\lambda$ are the values that minimize

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

  where $\lambda \geq 0$ is a tuning parameter, and x_{ij} is the jth coordinate of the ith observation x_i.
- Lasso: find $\hat{\beta}^L_\lambda$ that minimizes

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

- In contrast to the usual least squares approach, standardizing the predictors is important in shrinkage methods. Why?
- To standardize the predictors:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

- λ is an important penalty parameter that controls the amount of shrinkage.
- What happens when $\lambda = 0$ and $\lambda \to \infty$?
- How do we choose the tuning parameter λ ? Cross-validation
- Lasso tends to give sparser models compared to ridge (better for model interpretability), and it tends to perform better when the true model is sparse.
- But we do not know *a priori* which is better for prediction accuracy.

- In contrast to the usual least squares approach, standardizing the predictors is important in shrinkage methods. Why?
- To standardize the predictors:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

- λ is an important penalty parameter that controls the amount of shrinkage.
- What happens when $\lambda = 0$ and $\lambda \to \infty$?
- How do we choose the tuning parameter λ ? Cross-validation
- Lasso tends to give sparser models compared to ridge (better for model interpretability), and it tends to perform better when the true model is sparse.
- But we do not know *a priori* which is better for prediction accuracy.

Ridge regression does NOT give you sparse models

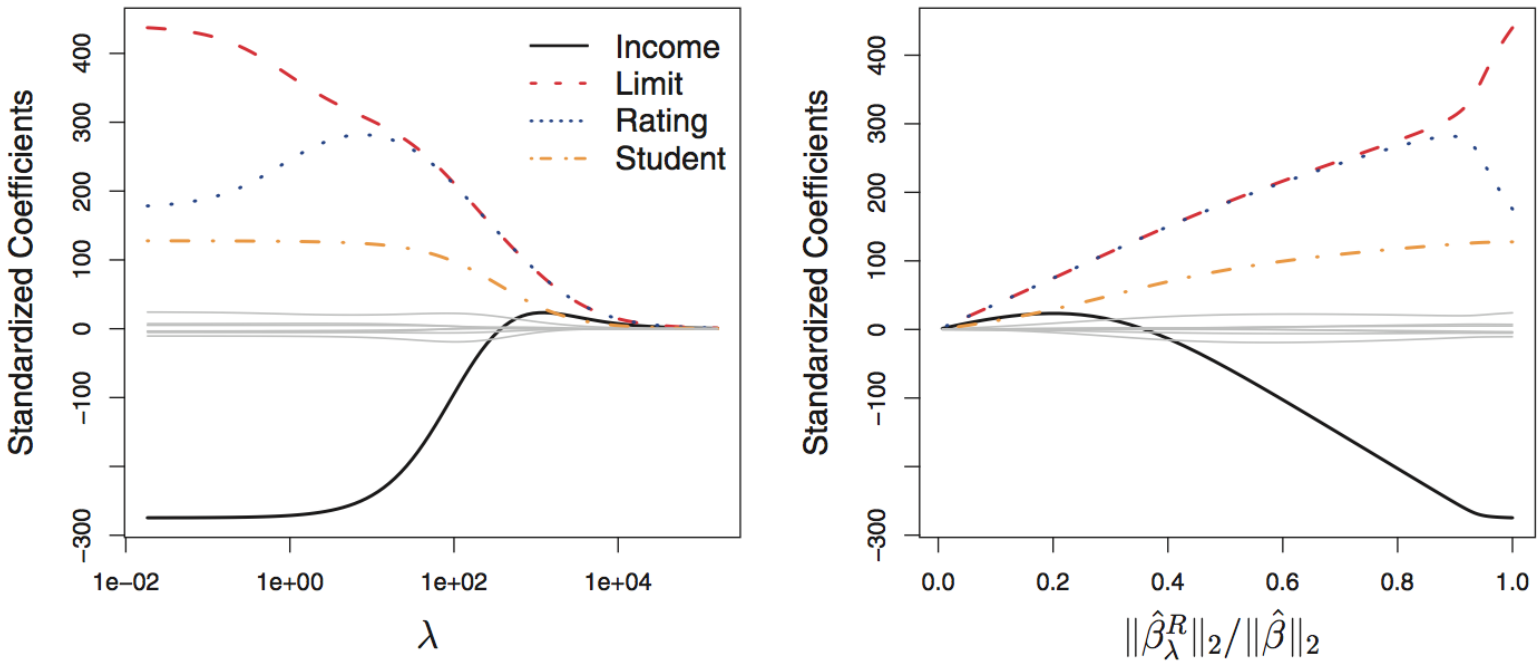

FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Figure 1

- Note $\|\beta\|_2 = \sqrt{\beta_1^2 + \cdots + \beta_p^2}$
- $\hat{\beta}$ denotes the vector of least squares estimate

Bias-variance trade-off for ridge regression

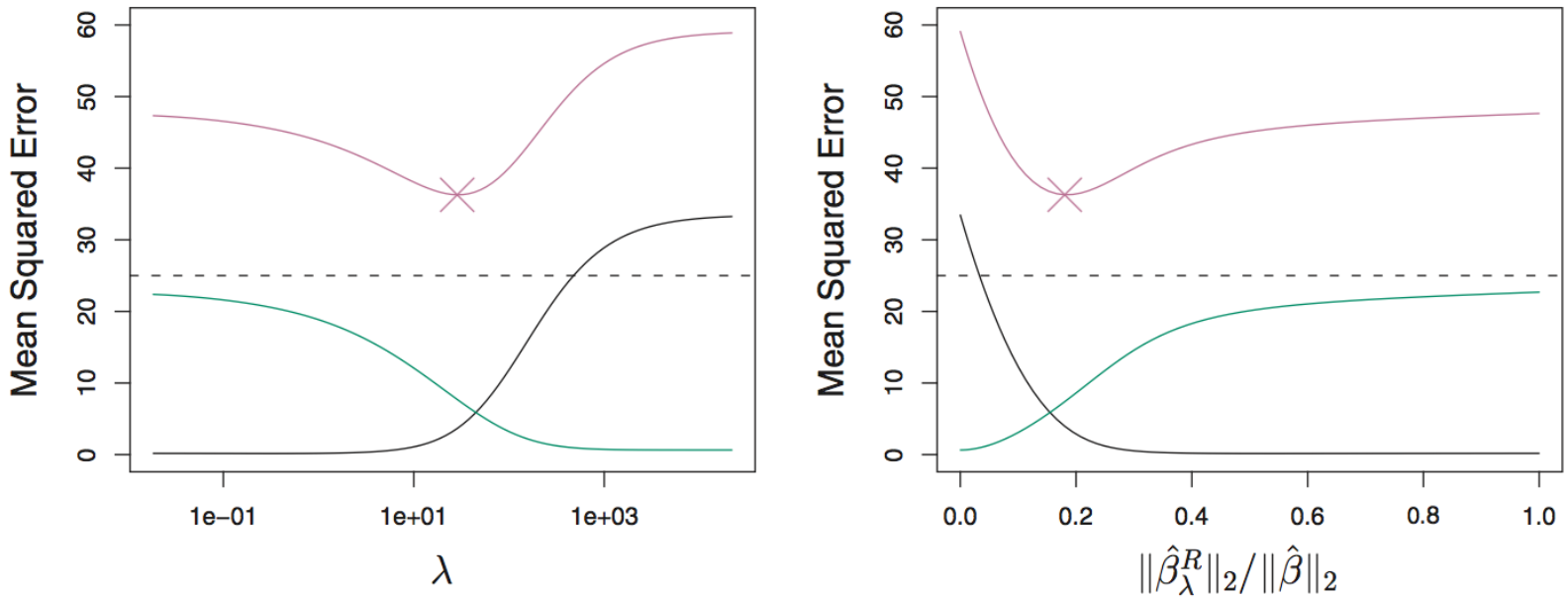

FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple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

Figure 2

Lasso encourages sparse models

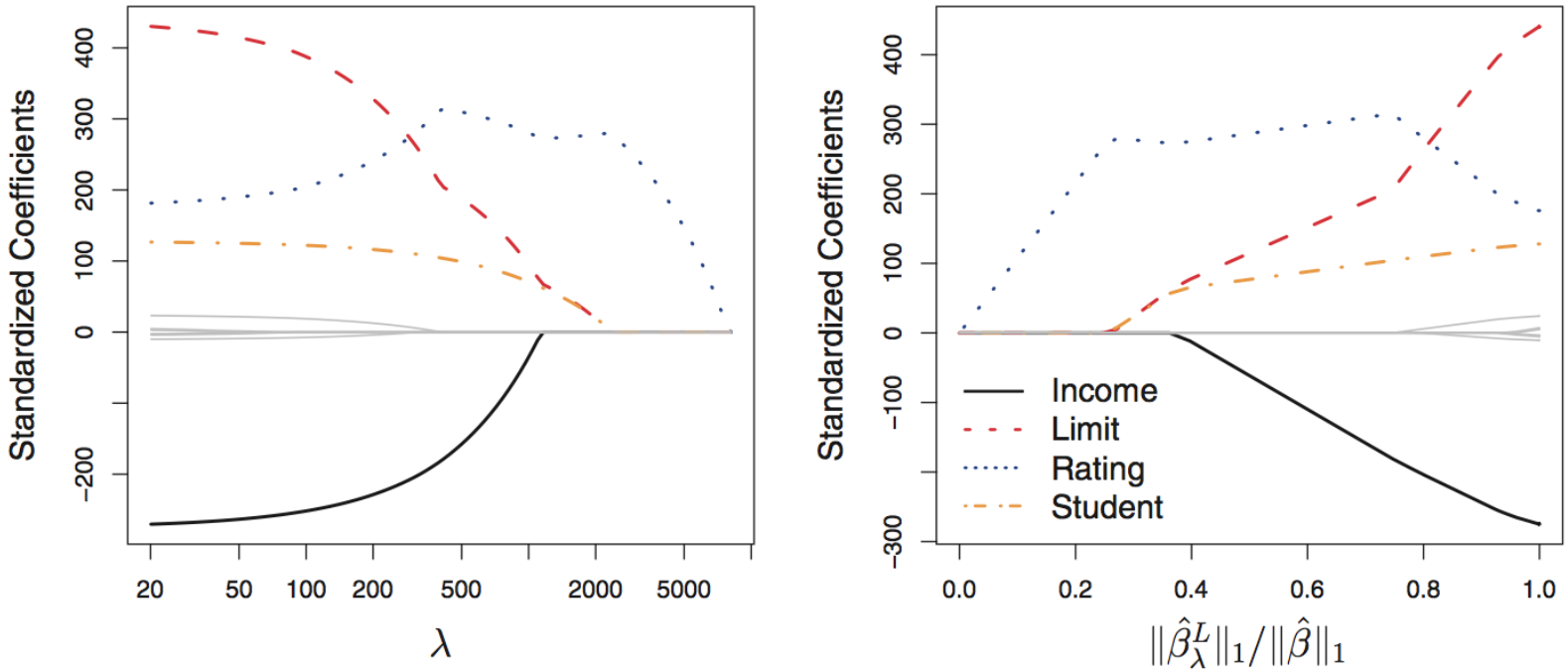

FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Figure 3

-Note $\|\beta\|_1 = |\beta_1| + \cdots |\beta_p|$

Another fomulation of ridge and lasso

- The Lasso's sparsity is better interpreted by an alternative formulation of Lasso and ridge regression
- λ and s has some corresponding relationships.

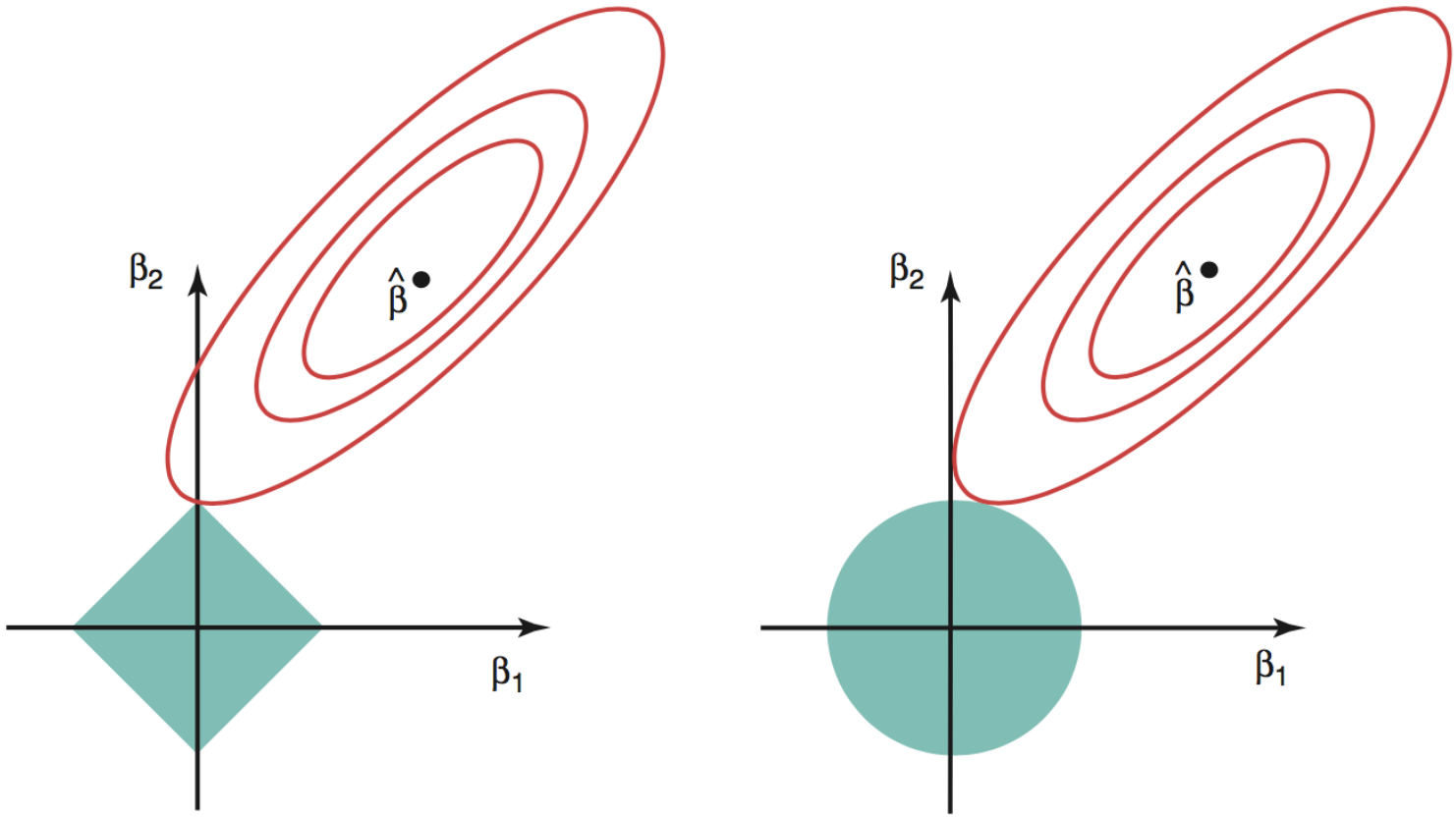

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

A simple special case for ridge regression and the Lasso

- Consider a simple case with $n = p$, and X is a diagonal matrix with 1's on the diagonal and 0's in all off-diagonal elements
- Assume that we perform regression without an intercept
- Under these assumptions, the usual least squares approach amounts to finding $\beta_1, \cdots, \beta_p$ that minimize

$$\sum_{j=1}^{p}(y_j - \beta_j)^2$$

- The least squares solution is given by

$$\hat{\beta}_j = y_j$$

- In this setting, ridge regression amounts to finding $\beta_1, \cdots, \beta_p$ to minimize

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- the ridge regression estimates take the form

$$\hat{\beta}_j^R = y_j/(1 + \lambda)$$

A simple special case for ridge regression and the Lasso (cont')

- The lasso amounts to finding coefficients to minimize

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- The lasso estimate takes the form

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2 & \text{if } y_j < \lambda/2 \\ 0 & \text{if } |y_j| \leq \lambda/2 \end{cases}$$

A simple special case for ridge regression and the Lasso (cont')

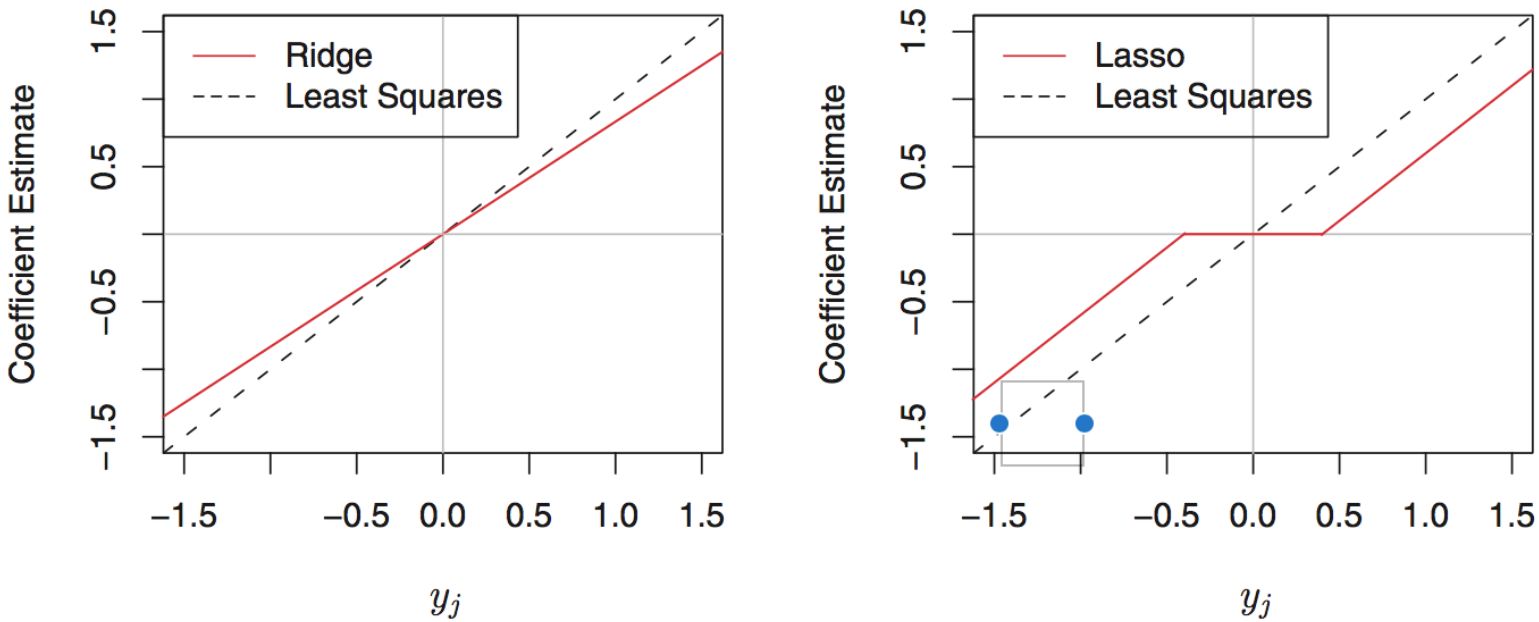

FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with* $n = p$ *and* $\mathbf{X}$ *a diagonal matrix with* 1*'s on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

Figure 4

High-dimensional setting

- High-dimensional settings: the scenarios where the number of predictors p is bigger than the sample size n
- A situation common in modern biology and medical sciences, but less so in business
- Including more variables into the regression, we potentially might find some useful features, but this benefit needs to be weighted against including many noise features.
- Example: Suppose 20 features are useful to predict a numerical outcome, and we fix sample size (say at $n = 50$). Please compare the following three scenarios
  - i) Use all 20 features for prediction
  - ii) Use 18 of the above 20 features for prediction
  - iii) Use all 20 features, plus 100 noise features for prediction
- i) is clearly better than iii) (no decrease in bias, but increase in variance). It is hard to compare i) and ii) only based on this abstract description (think about variance and bias again).

High-dimensional setting

- High-dimensional settings: the scenarios where the number of predictors p is bigger than the sample size n
- A situation common in modern biology and medical sciences, but less so in business
- Including more variables into the regression, we potentially might find some useful features, but this benefit needs to be weighted against including many noise features.
- Example: Suppose 20 features are useful to predict a numerical outcome, and we fix sample size (say at $n = 50$). Please compare the following three scenarios
  - i) Use all 20 features for prediction
  - ii) Use 18 of the above 20 features for prediction
  - iii) Use all 20 features, plus 100 noise features for prediction
- i) is clearly better than iii) (no decrease in bias, but increase in variance). It is hard to compare i) and ii) only based on this abstract description (think about variance and bias again).