STAT 380:
Variable Selection, Ridge, and Lasso Regression

An Example using .... Data

United Arab Emirates University

Penalized Regression: idea behind

- They are also known as shrinkage methods or regularization models.
- We would like to continue using linear regression models, but we need to **adjust** them to be usable with big or high-dimensional datasets.
- We introduce a **penalty** for too many or too large coefficients.
- 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.

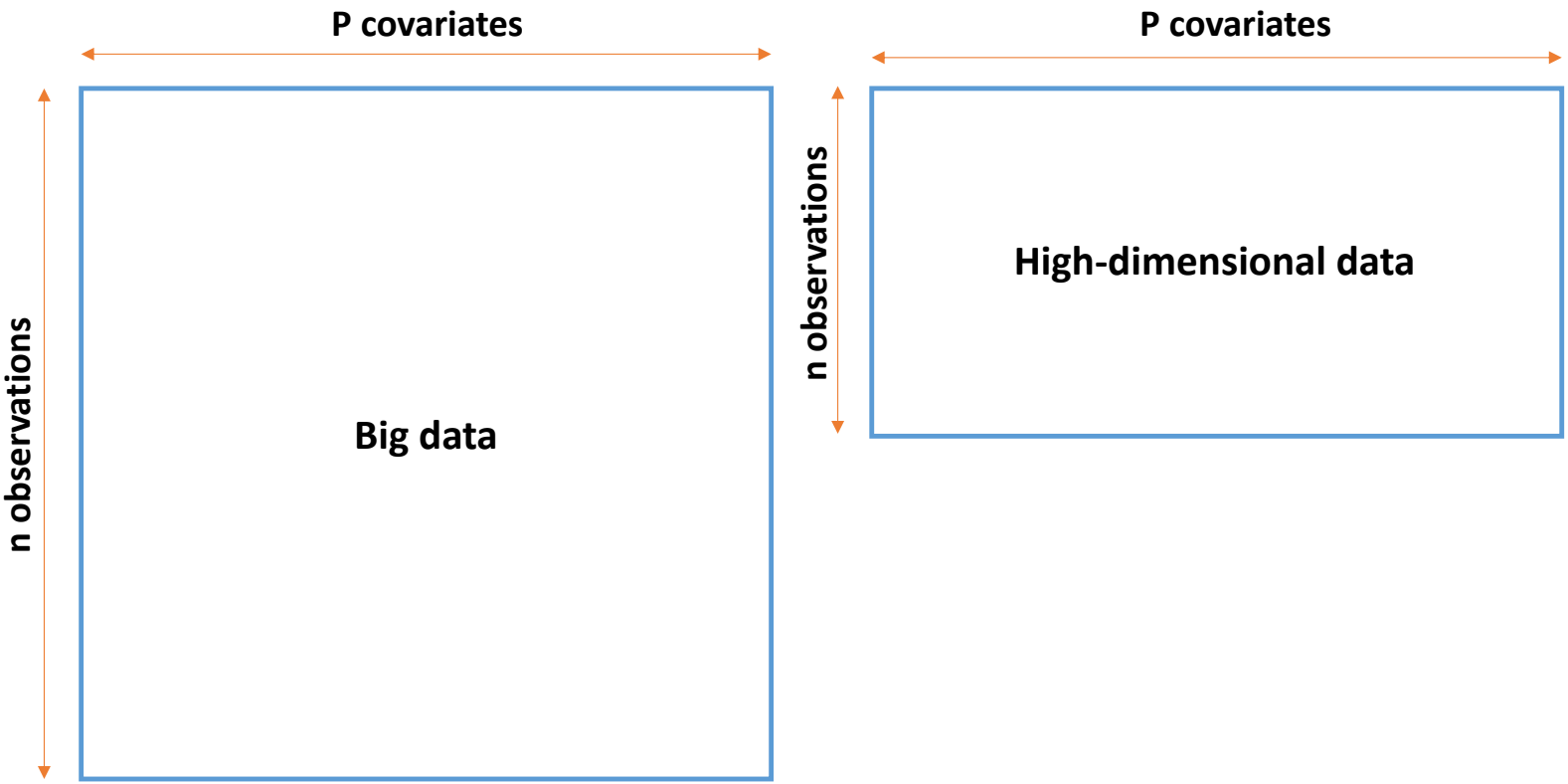
Penalized regression: when?

Regularization methods can be used when at least one of the following conditions is met

- large number of variables
- more variables than observations $n \ll p$
- strong multicollinearity
- a sparse solution is wanted/needed (feature selection)
- "The word 'high-dimensional' refers to a situation where the number of unknown parameters which are to be estimated is one or several orders of magnitude larger than the number of samples in the data."[2]

[2] Peter Bühlmann, Sara van de Geer - Statistics for High-Dimensional Data, Springer 2011

Big data vs. high-dimensional data



Examples of high-dimensional data

Typically, high-dimensional data arise in a number of settings:

- genomics (microarrays, proteomics)
- signal processing
- image analysis
- market basket data and portfolio allocation
- industry (3d-printing)

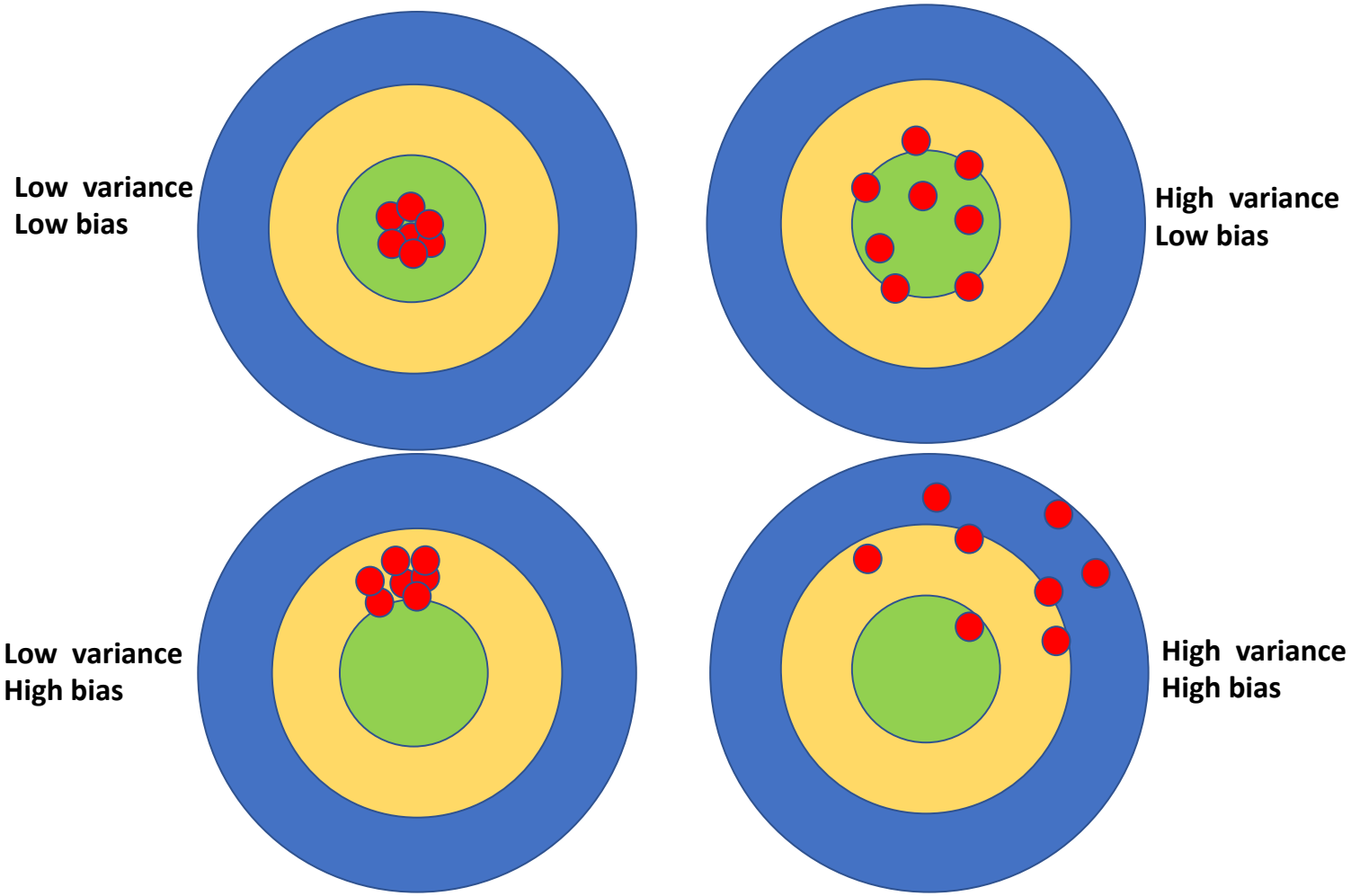
MSE of a predictor: remeber

We use the MSE together with cross validation to assess our model fit.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$$

Or more exactly, the mean squared prediction error.

Bias-variance trade-off



Geneset MicroArray Data

- The covariates are the allele frequencies of 200 Gene sets for 120 subjects. (Scheetz et al., (2006)

  It represents the data of 120 rats with 200 gene probes.

  Response a 120-dimensional vector of, which represents the expression level of 'TRIM32' gene.

  We want to idensity which of the other genes are significantly responsible for the gene counts of the 'TRIM32'.

- Therefore, In terms of the regression terminology: $n = 120$ and $p = 200$.

Thank You