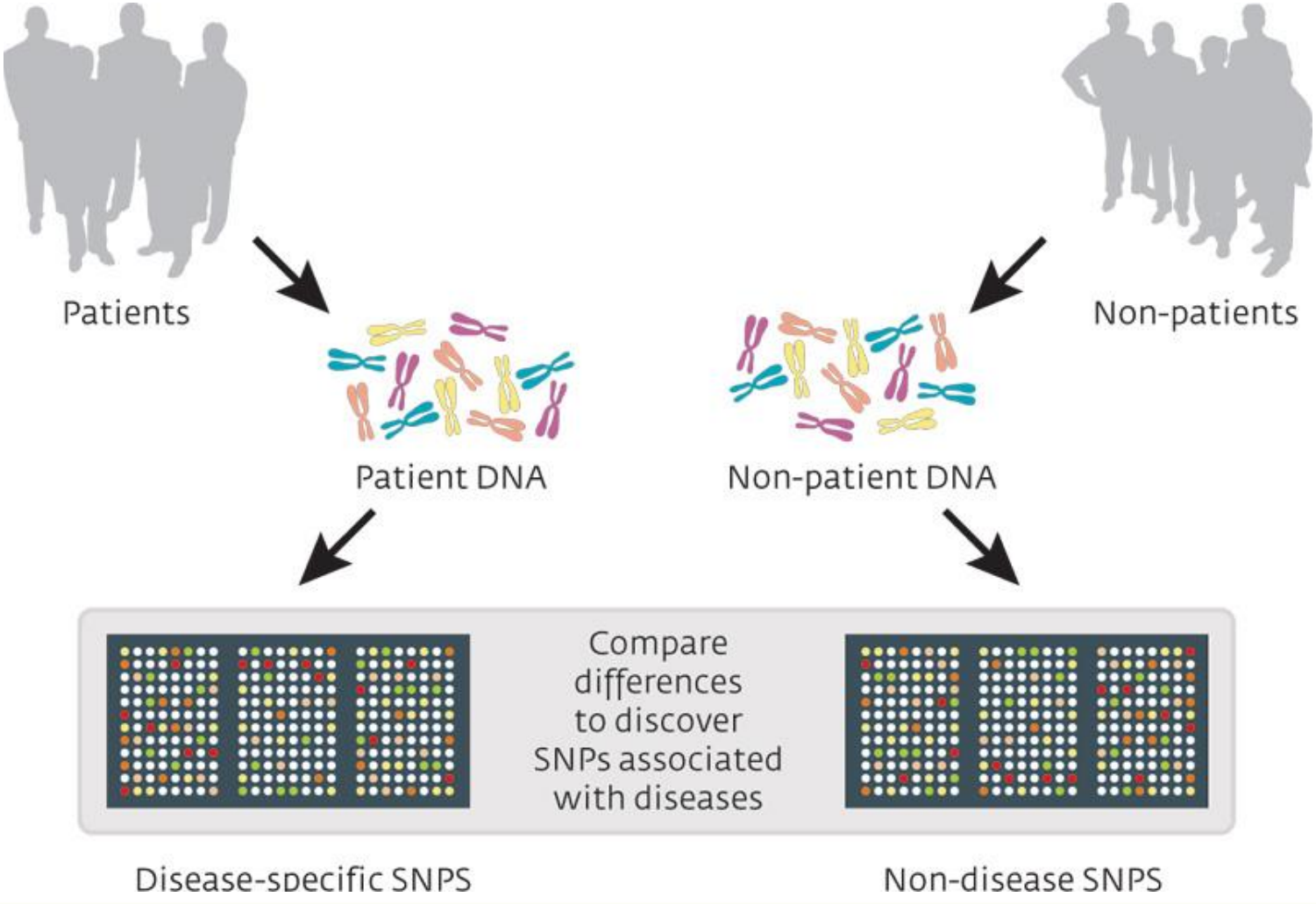
Predictive Modelling

Lecture 6: Linear model selection and regularization

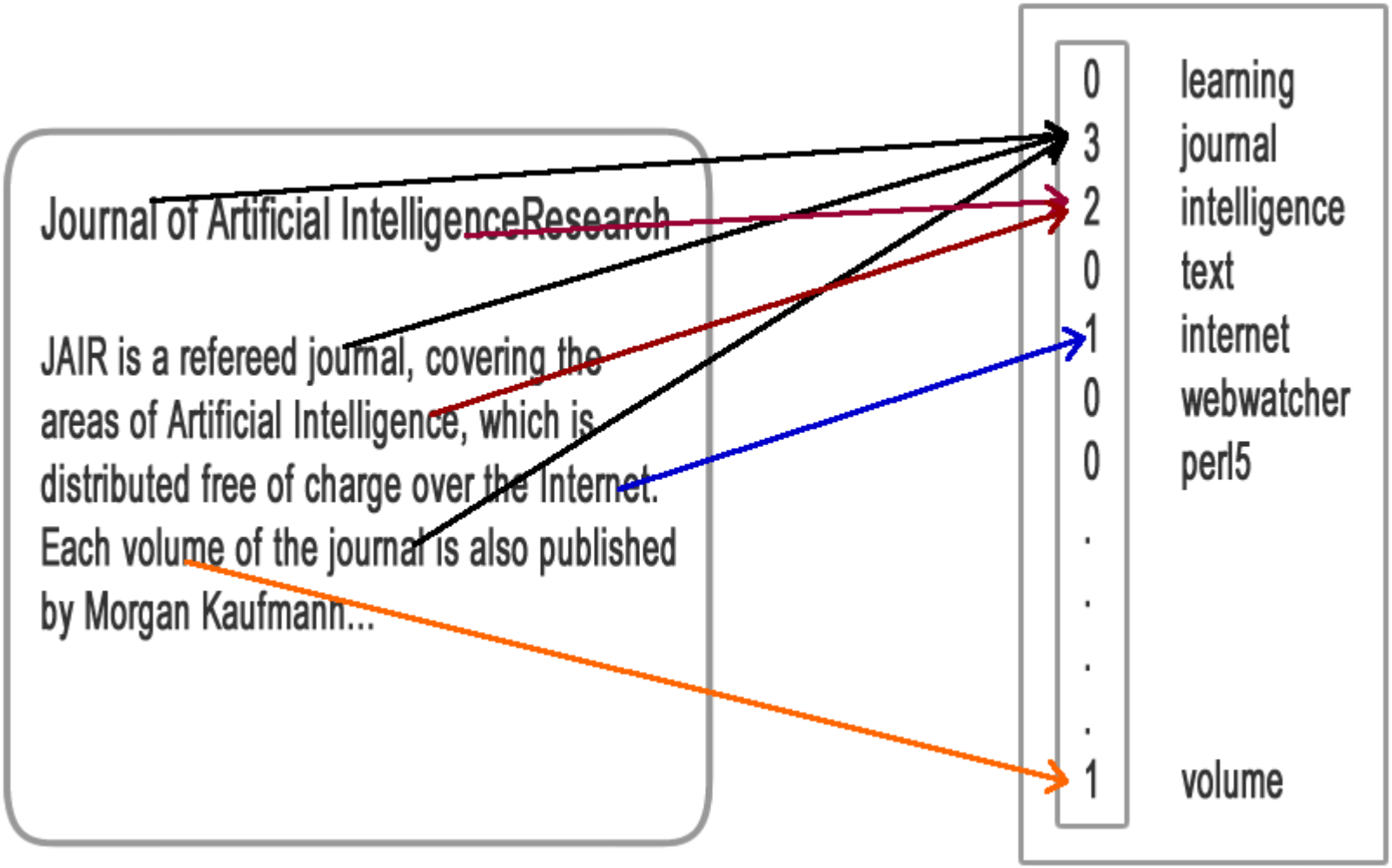
Overview of this lecture

- Chapter 6
  - 6.1: Subset selection
  - 6.2: Shrinkage methods
  - 6.4: Considerations in high dimensions

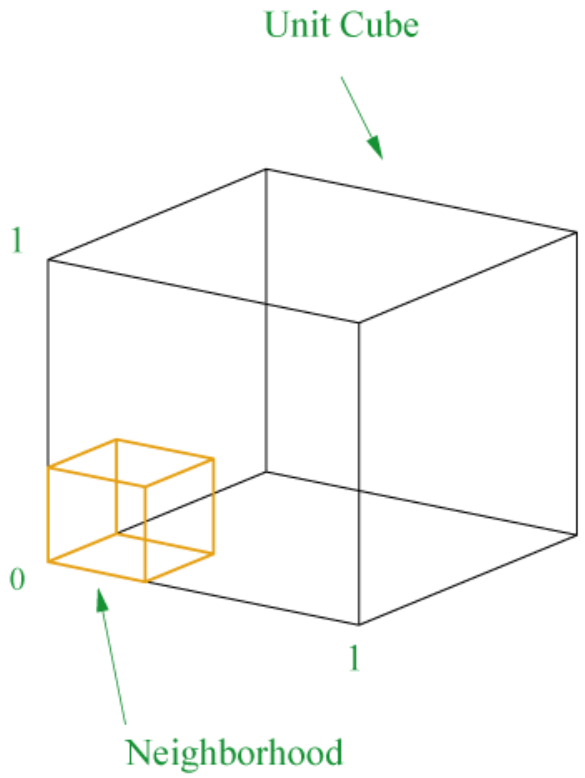
Genome-wide association studies: an example of high-dimensional data

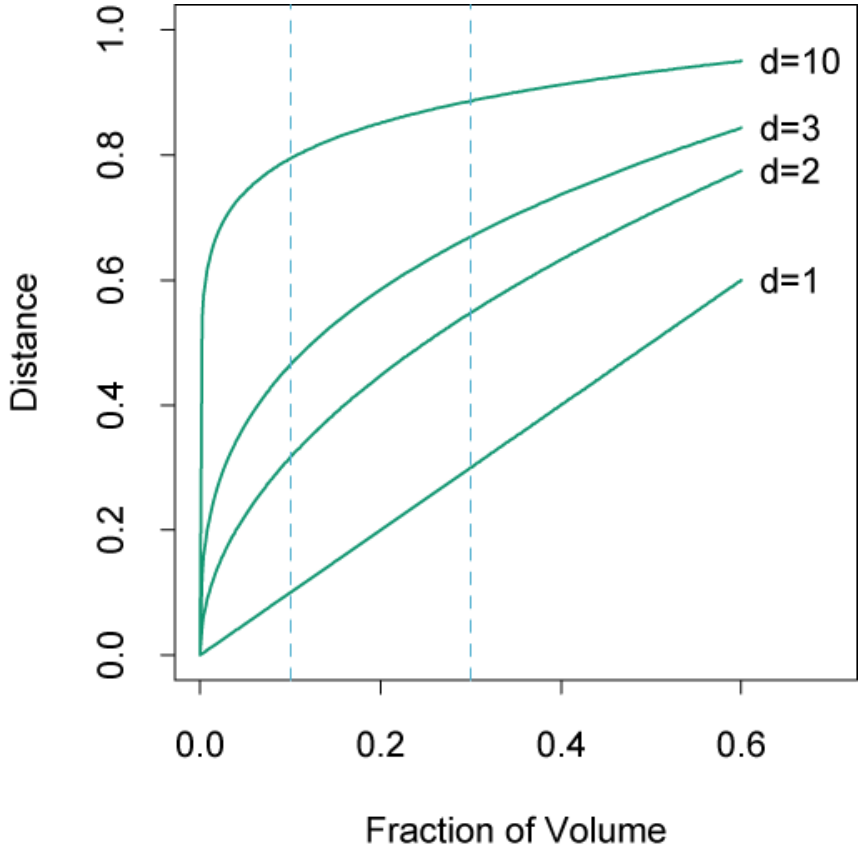


Text mining: why will k-NN fail?



An illustration of the curse of dimensionality

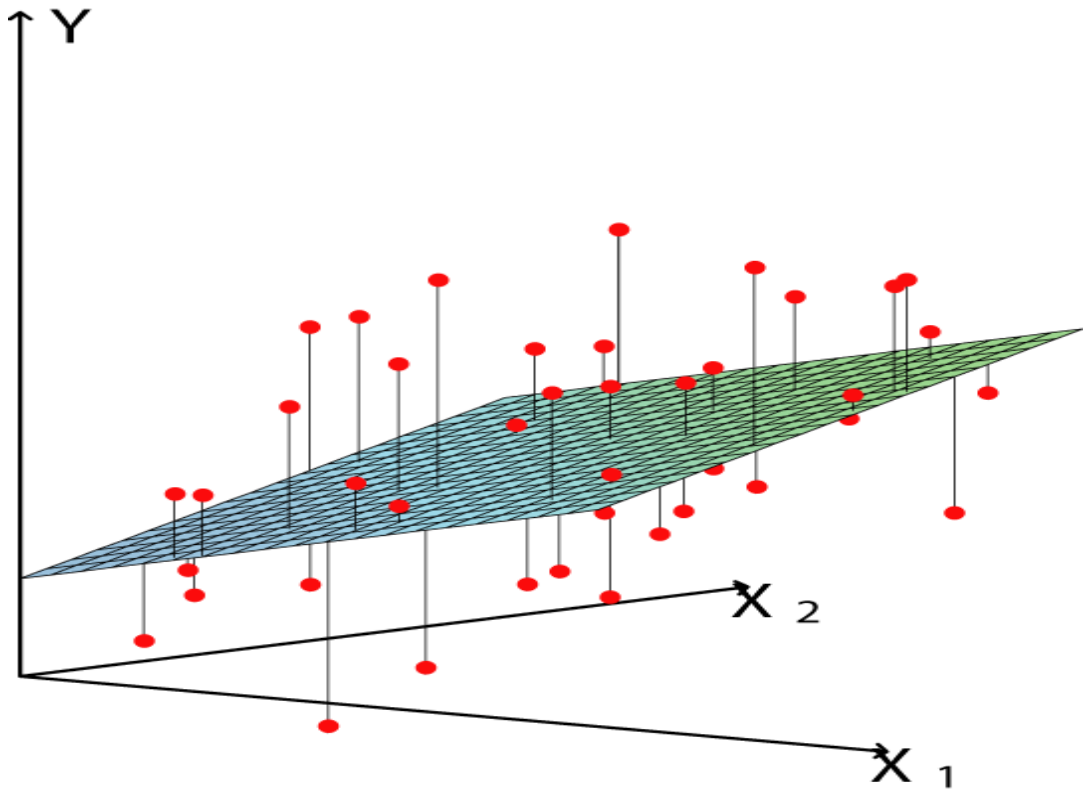




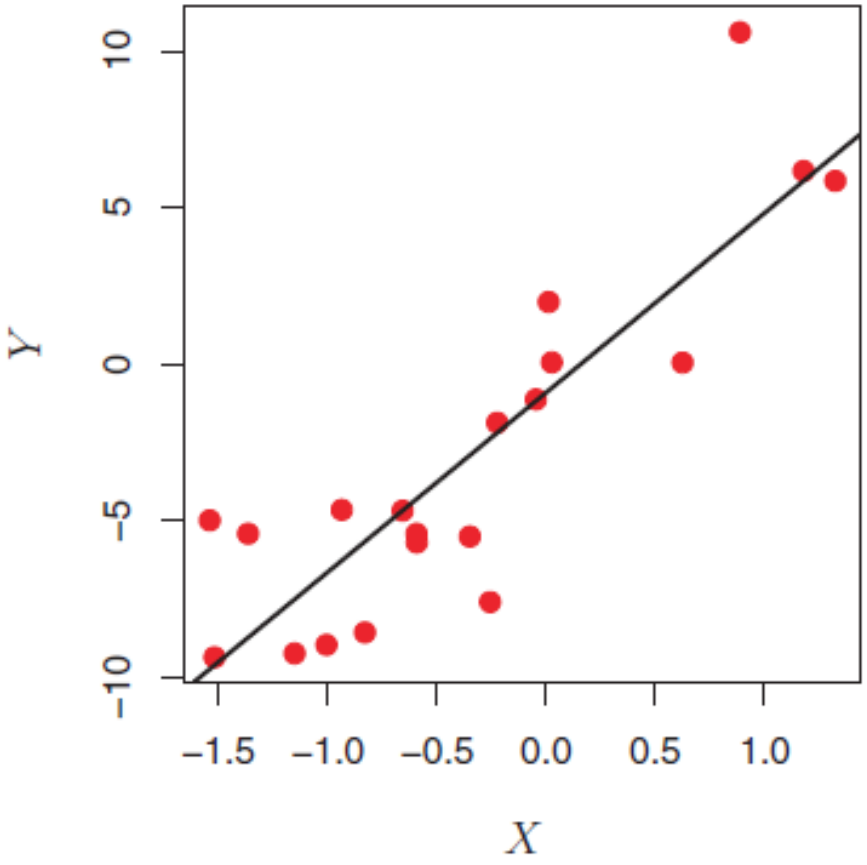
Side-length of the subcube needed to capture a fraction r of the data: 10 dims. already a side-length 0.8 to capture 10% of the data

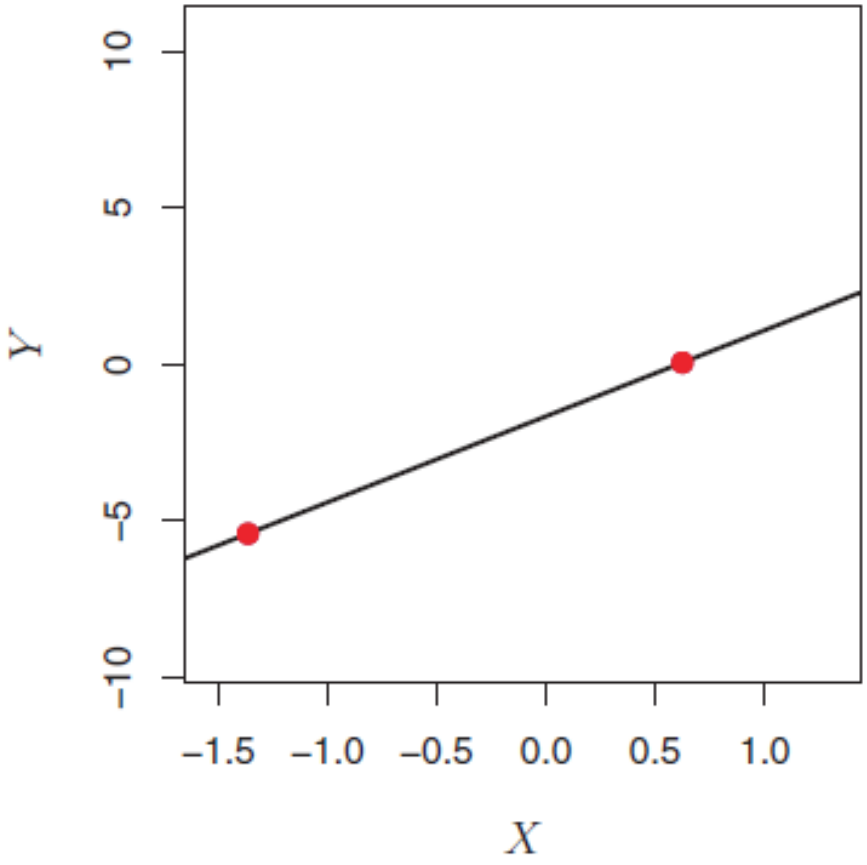
Why will linear regression fail?

$$y_i = w_0 + w_1 x_{i1} + \ldots + w_p x_{ip} + \epsilon_i = \sum_{j=0}^{p} w_j x_{ij} + \epsilon_i$$



An illustration of linear regression in high dimensions





Feature selection / Variable selection

- Different approaches:
  - **Filter methods:** select features based on a univariate statistic (e.g. Correlation with response)
  - **Wrapper methods:** use a machine learning method in a standard procedure
  - **Embedded methods:** see e.g. Lasso, Random Forests

Best subset selection

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

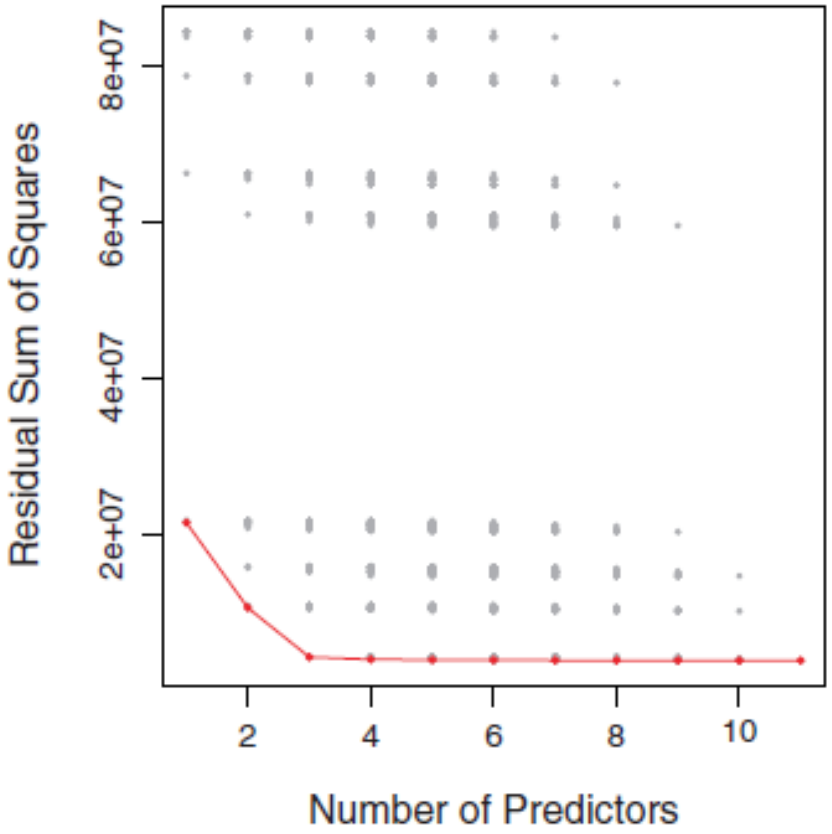
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

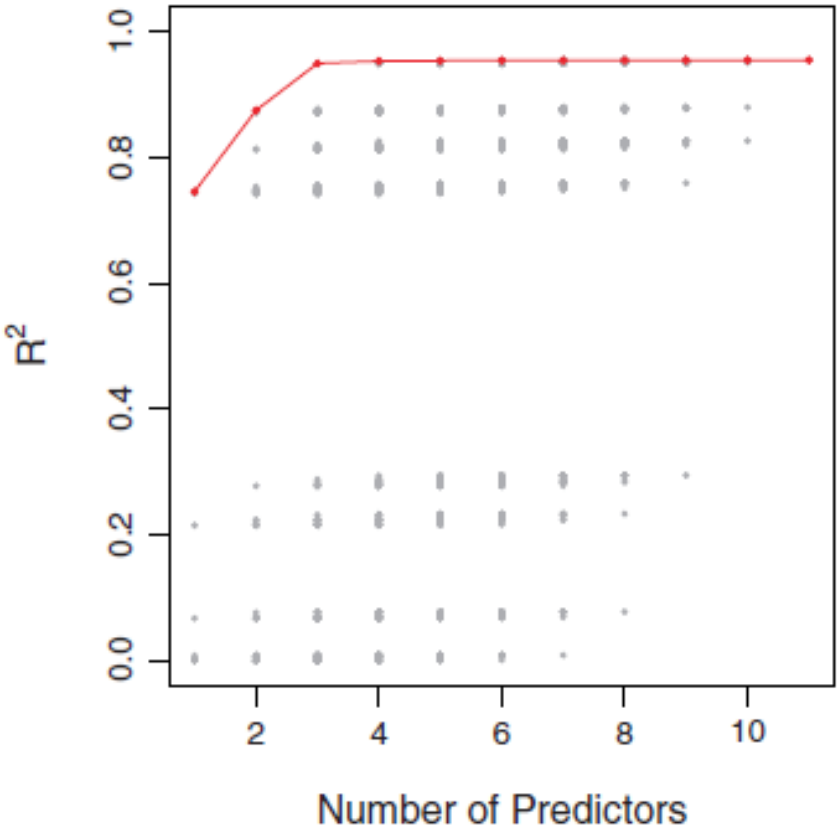
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

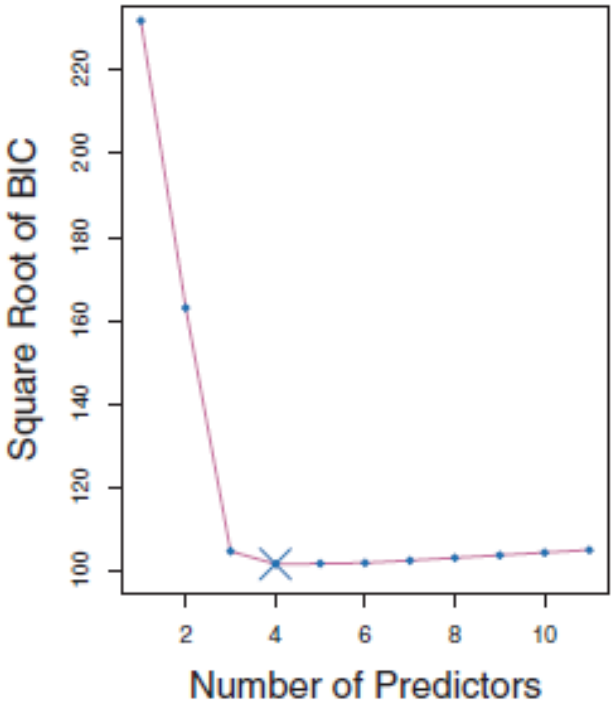
What is an obvious drawback of this method ?

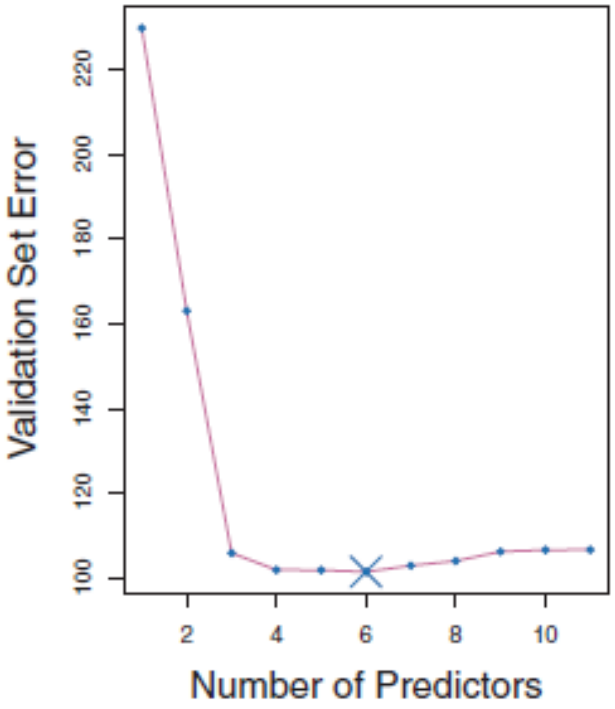
Best subset selection on the credit dataset (training data)

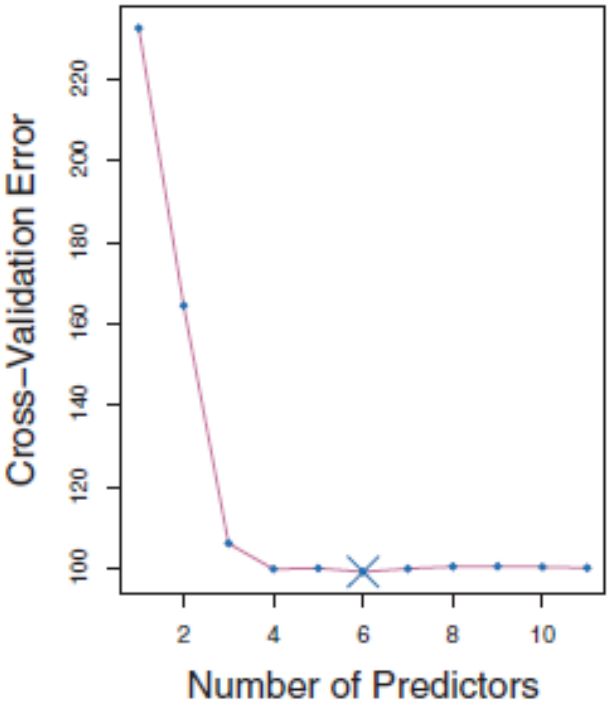




Best subset selection on the credit dataset (out-of-sample data)







Forward stepwise selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Best subset selection and forward stepwise selection compared

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

Would you prefer best or forward feature selection?

Backward stepwise selection (aka recursive feature elimination)

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

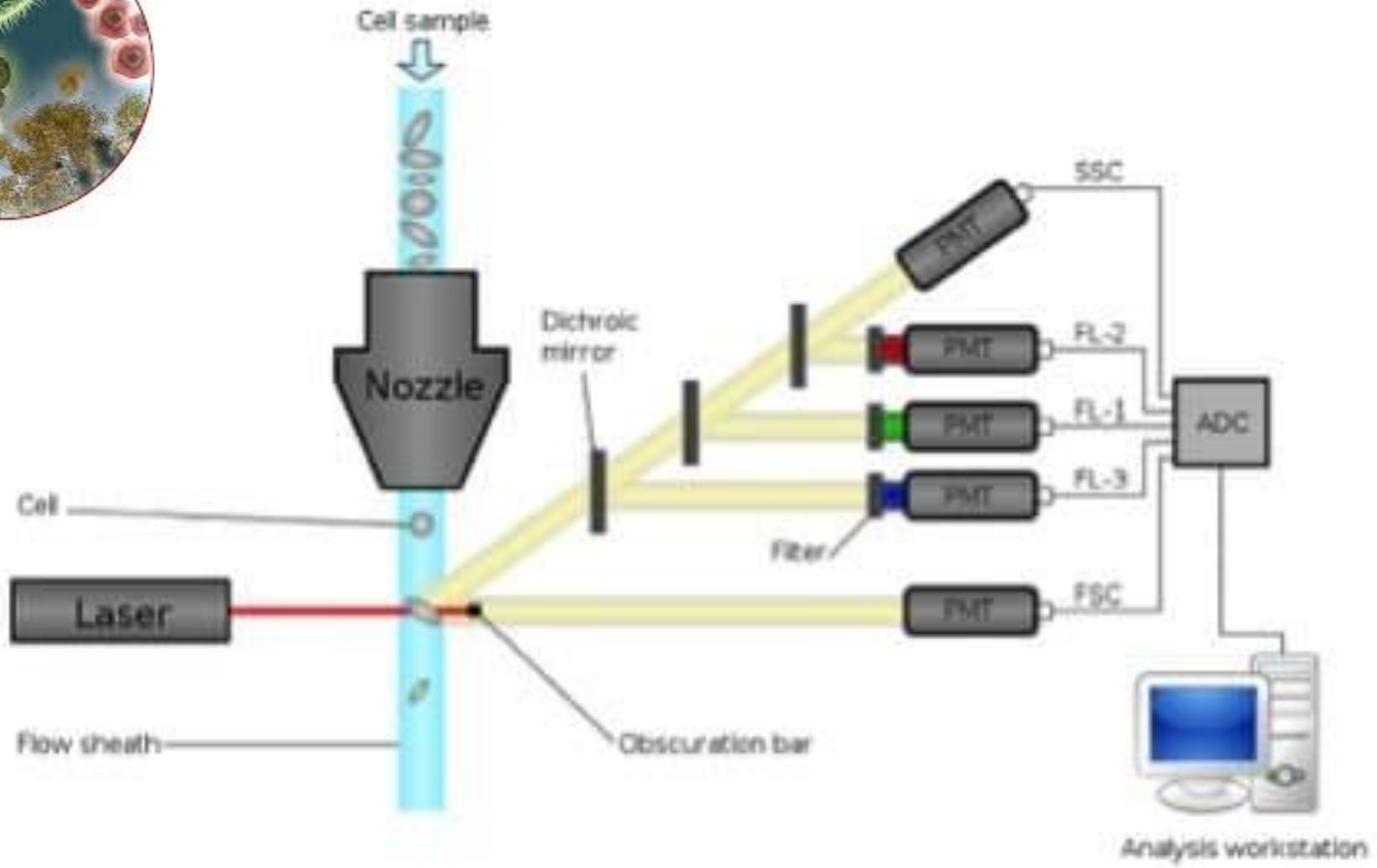
Would you prefer backward over forward selection?

Other motivations for feature selection

Interpretability

Cost efficiency

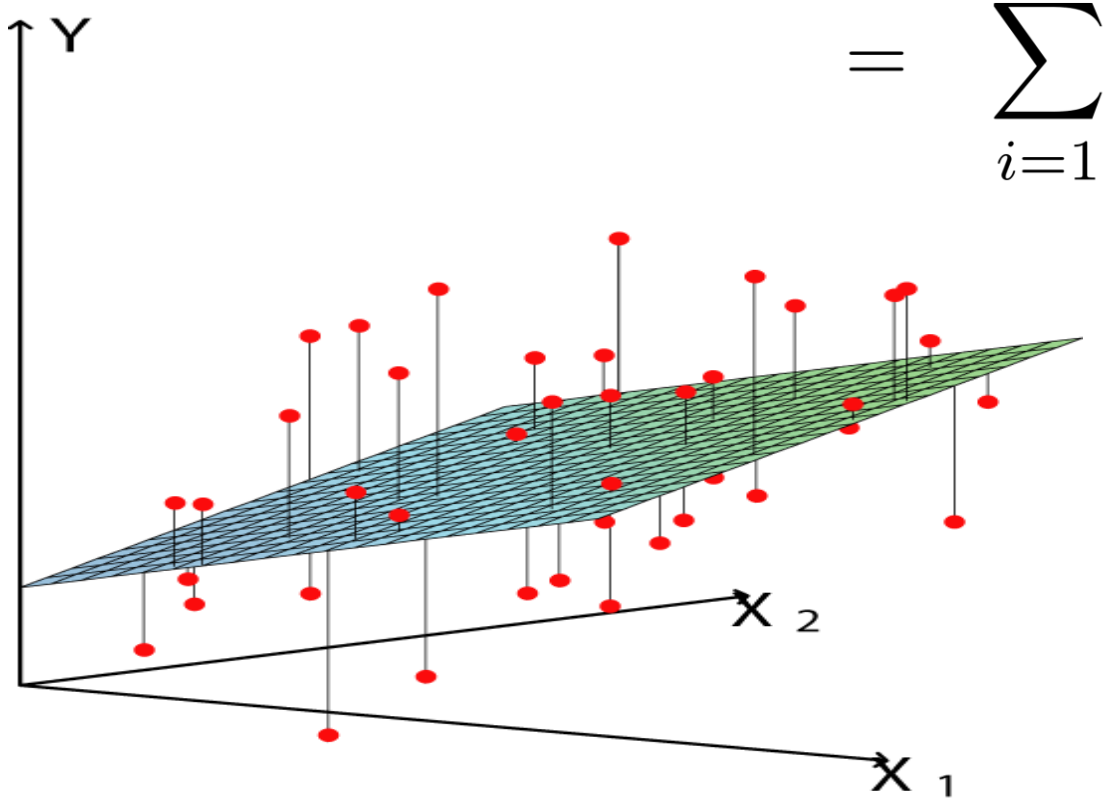
A case study: microbial flow cytometry



Least-squares minimization revisited

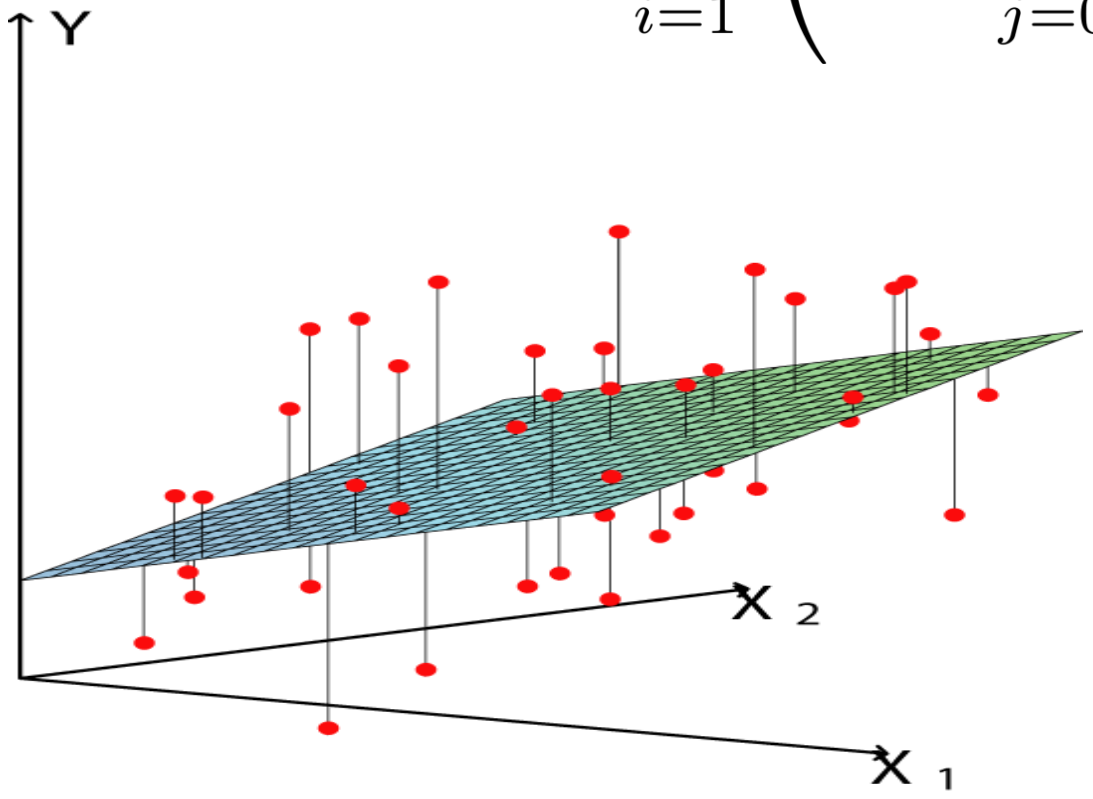
$$RSS(\boldsymbol{w}) = \epsilon_1^2 + \epsilon_2^2 + \ldots + \epsilon_n^2$$

$$= \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2$$



Ridge regression: L2-regularization of the parameter vector

$$\hat{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} w_j^2$$



What is the effect of solving this adjusted optimization problem?

$$X = \begin{bmatrix} x_{10} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n0} & \cdots & x_{np} \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}$$

$$\min_{\boldsymbol{w}} RSS(\boldsymbol{w}) \quad = \quad (\boldsymbol{y} - X\boldsymbol{w})^T(\boldsymbol{y} - X\boldsymbol{w}) + \lambda||\boldsymbol{w}||_2^2$$

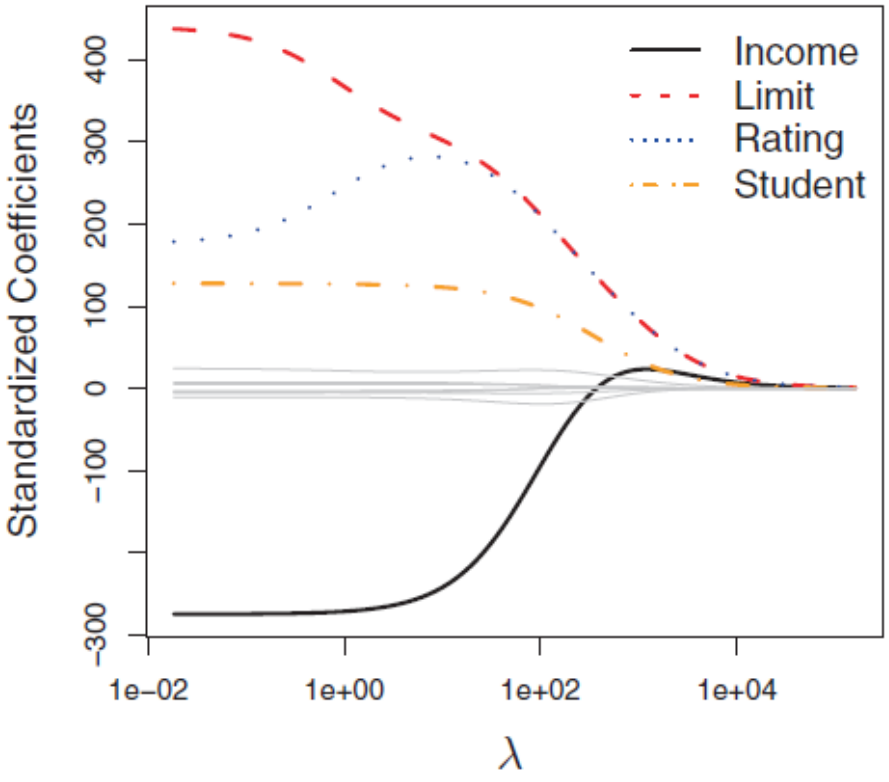
$$\frac{\partial}{\partial \boldsymbol{w}} RSS(\boldsymbol{w}) = -2X^T(\boldsymbol{y} - X\boldsymbol{w}) + 2\lambda\boldsymbol{w} = \boldsymbol{0}$$

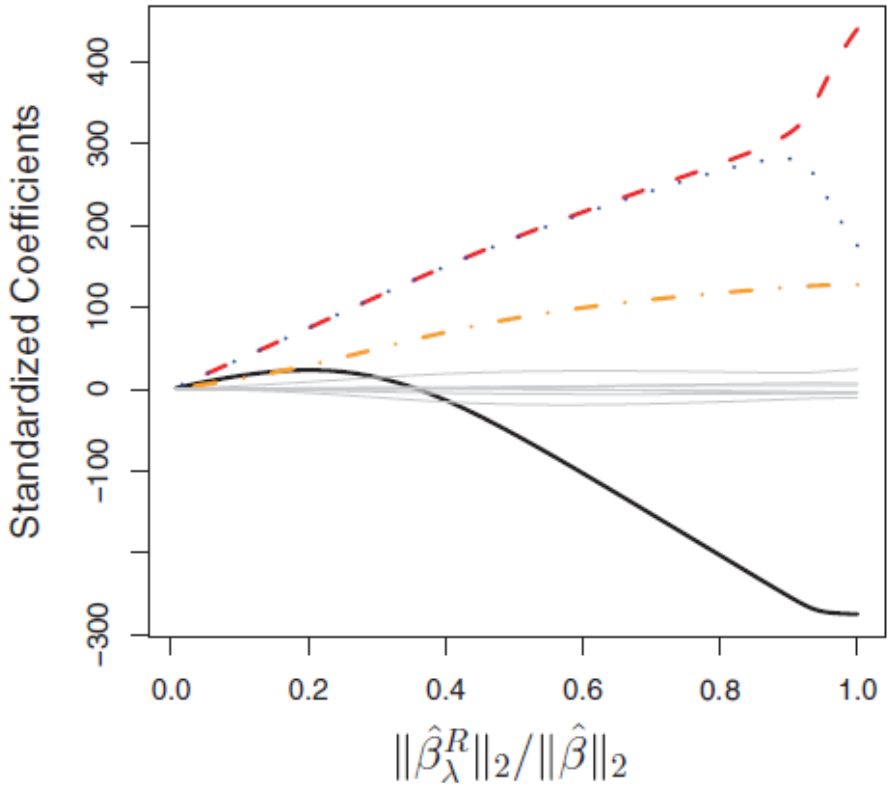
$$\Leftrightarrow \quad X^T X\boldsymbol{w} - X^T\boldsymbol{y} + \lambda\boldsymbol{w} = \boldsymbol{0}$$

$$\Leftrightarrow \quad (X^T X + \lambda I)\boldsymbol{w} = X^T\boldsymbol{y}$$

$$\Leftrightarrow \quad \boldsymbol{w} = (X^T X + \lambda I)^{-1} X^T\boldsymbol{y}$$

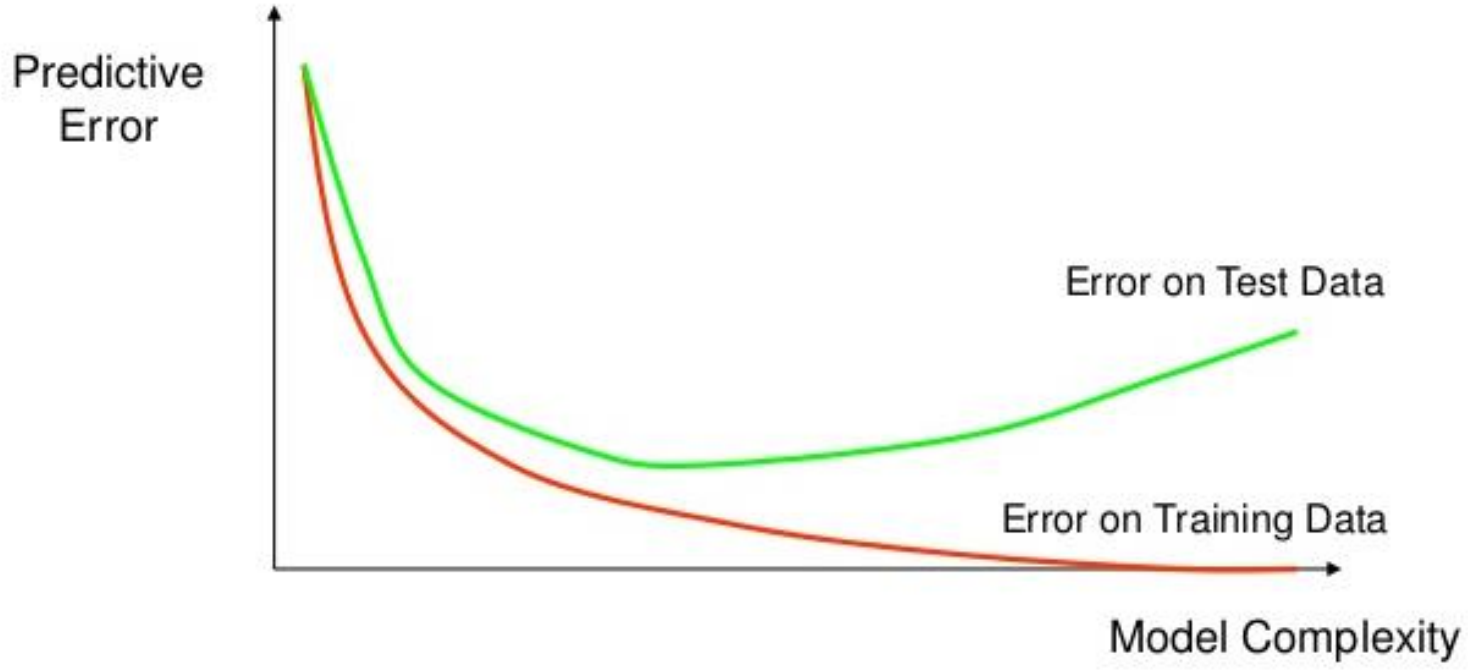
Ridge regression in action on the credit dataset



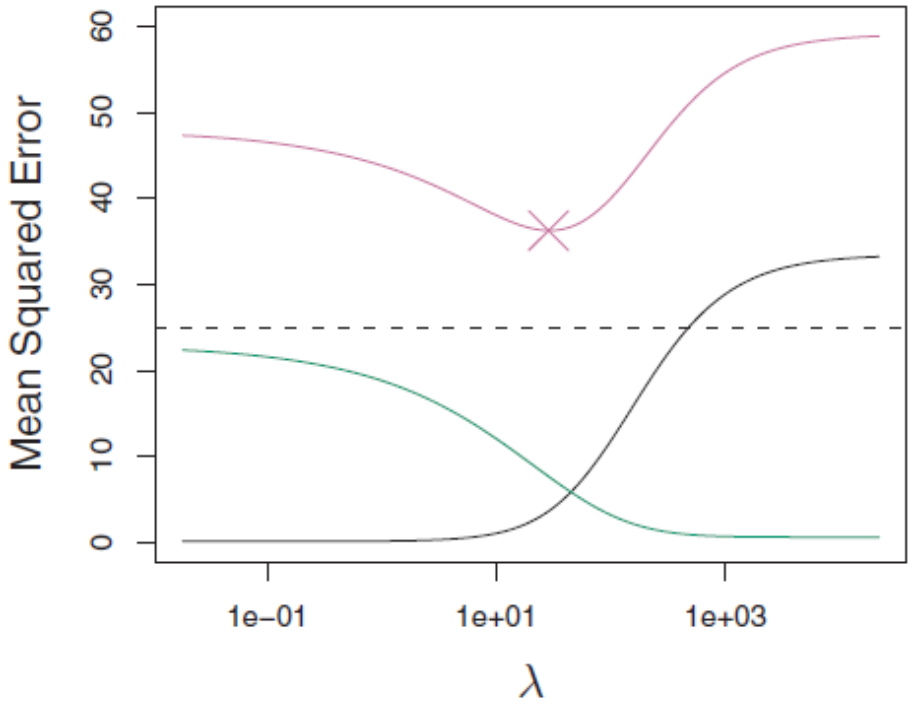


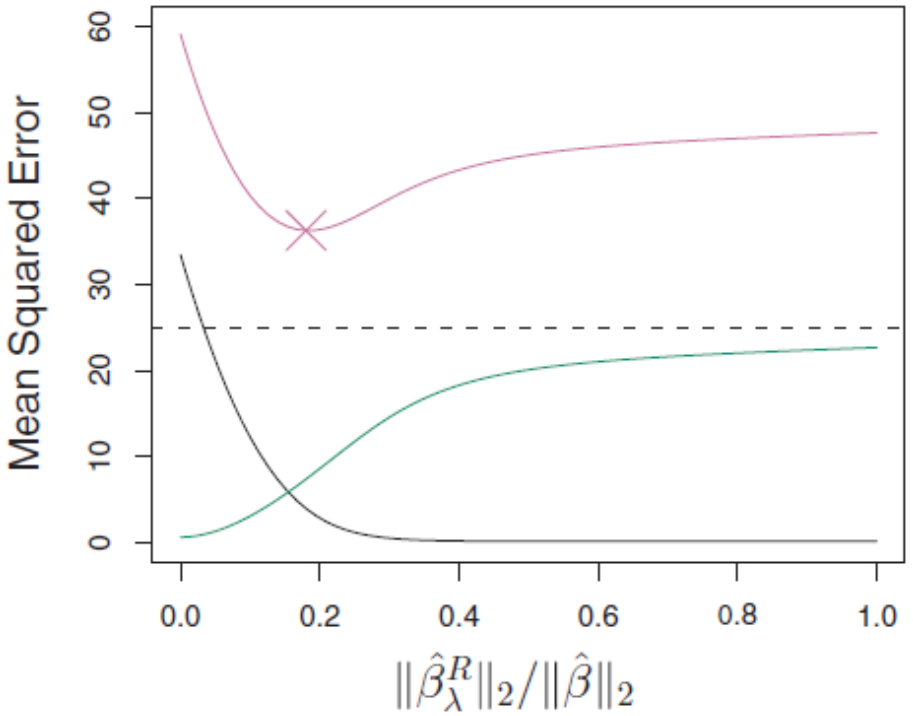
Ridge regression prevents overfitting

$$\hat{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} w_j^2$$



Ridge regression improves the prediction error (simulated data)



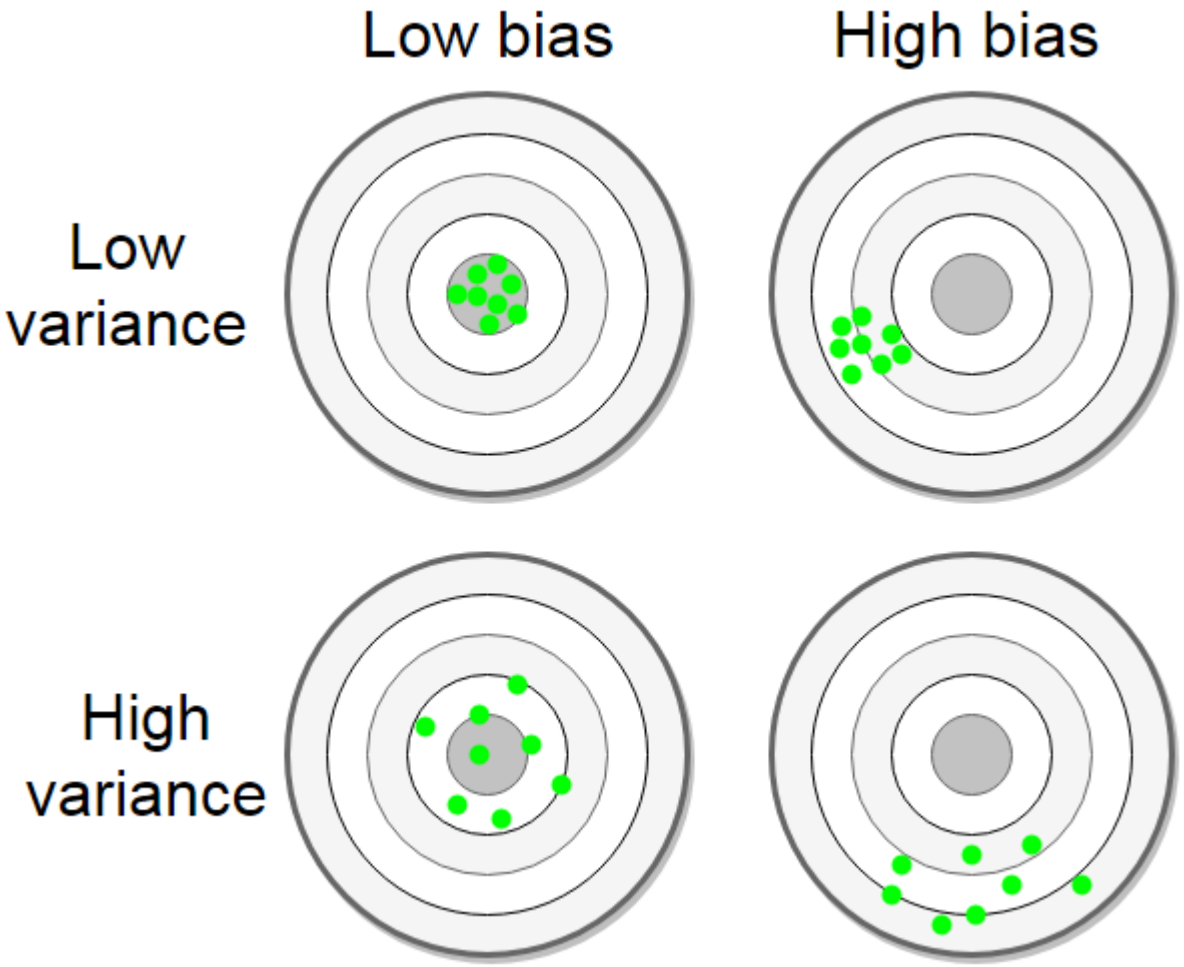


Red curve = prediction error

Green curve = variance

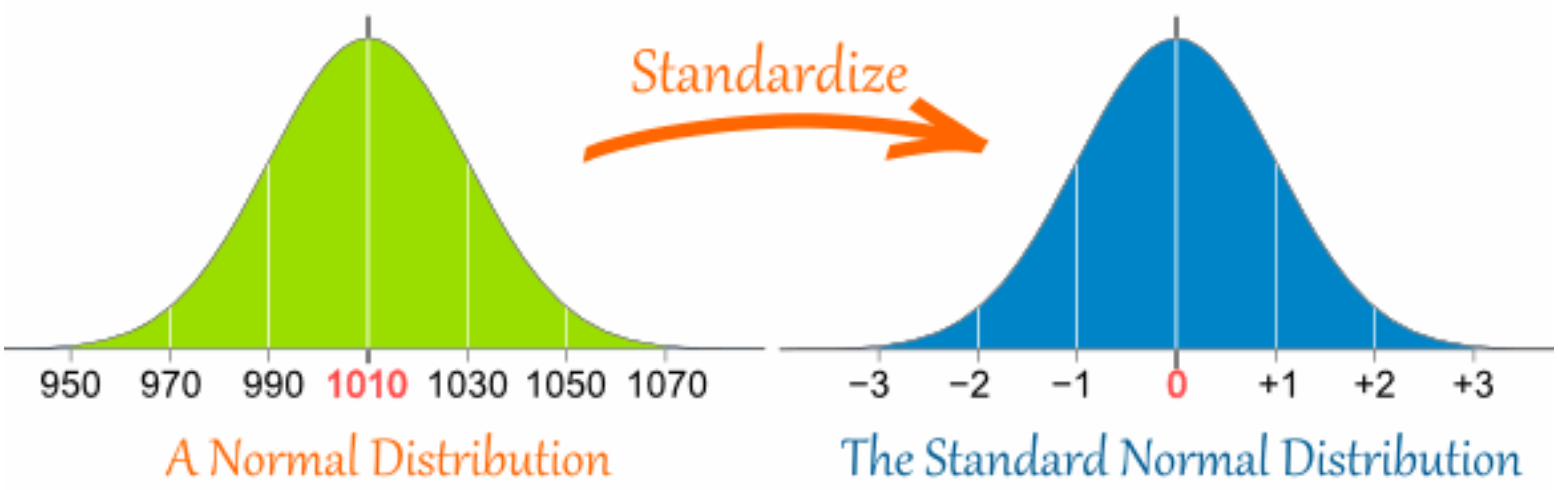
Black curve = squared bias

An intuitive interpretation of bias and variance



Ridge regression: why is it important to standardize the data?

$$\hat{\boldsymbol{w}} = \operatorname{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} w_j^2$$



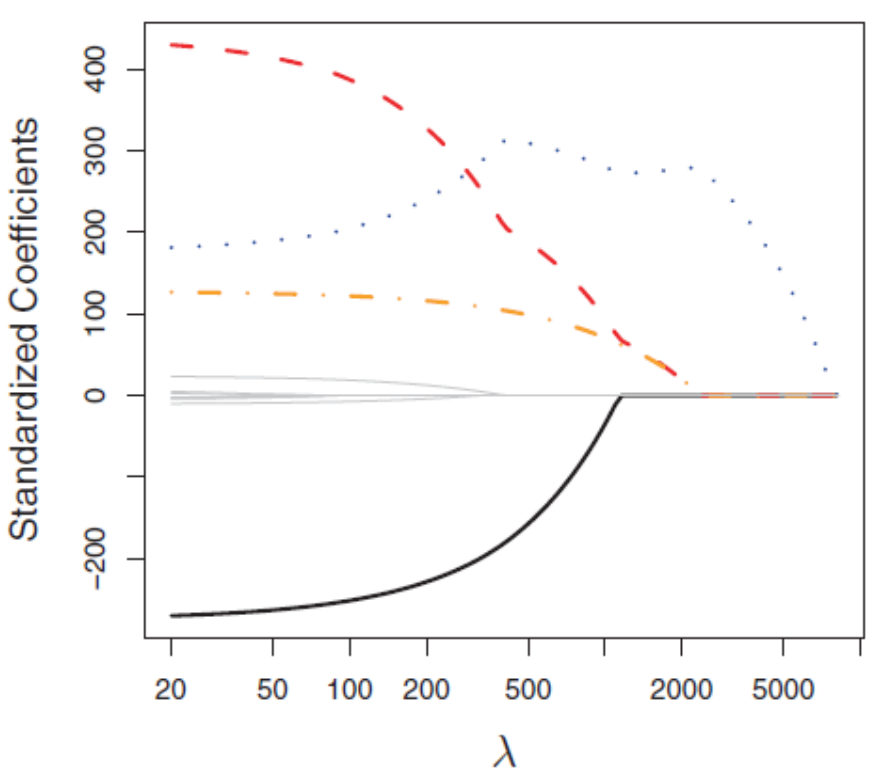
Lasso: L1-regularization of the parameter vector

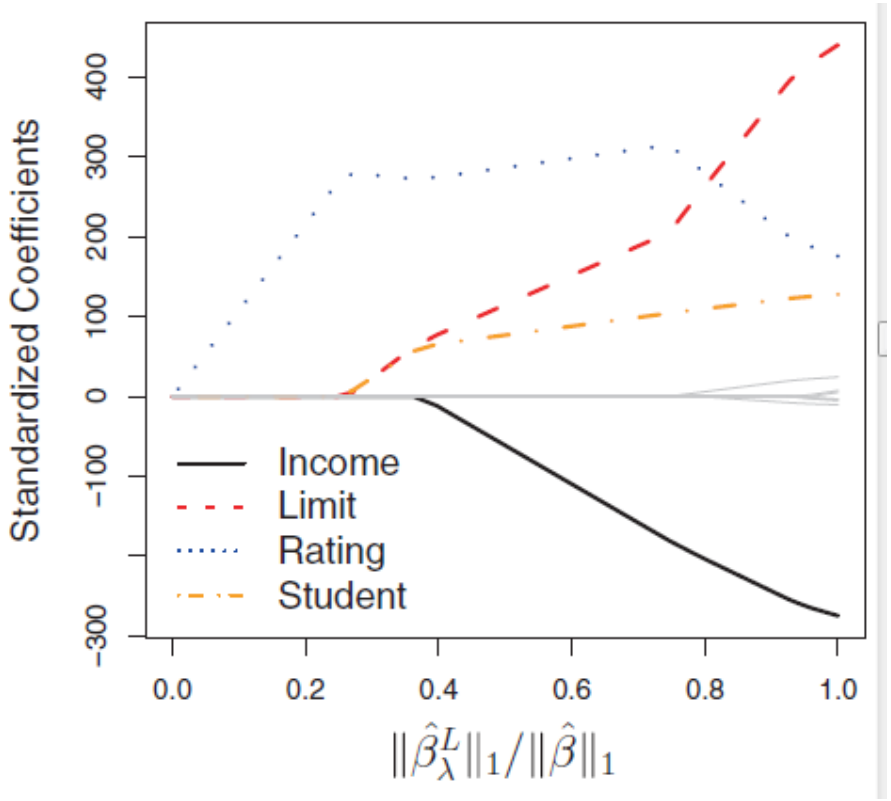
- Slightly different penalty term:

$$\hat{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |w_j|$$

- L1-norm as penalty term instead of L2-norm
- Has a very similar effect as L2-regularization
- No closed-form solution, but efficient algorithms exist

LASSO on the credit dataset





What is the advantage of Lasso compared to ridge regression?

Another mathematical formulation of ridge regression and LASSO

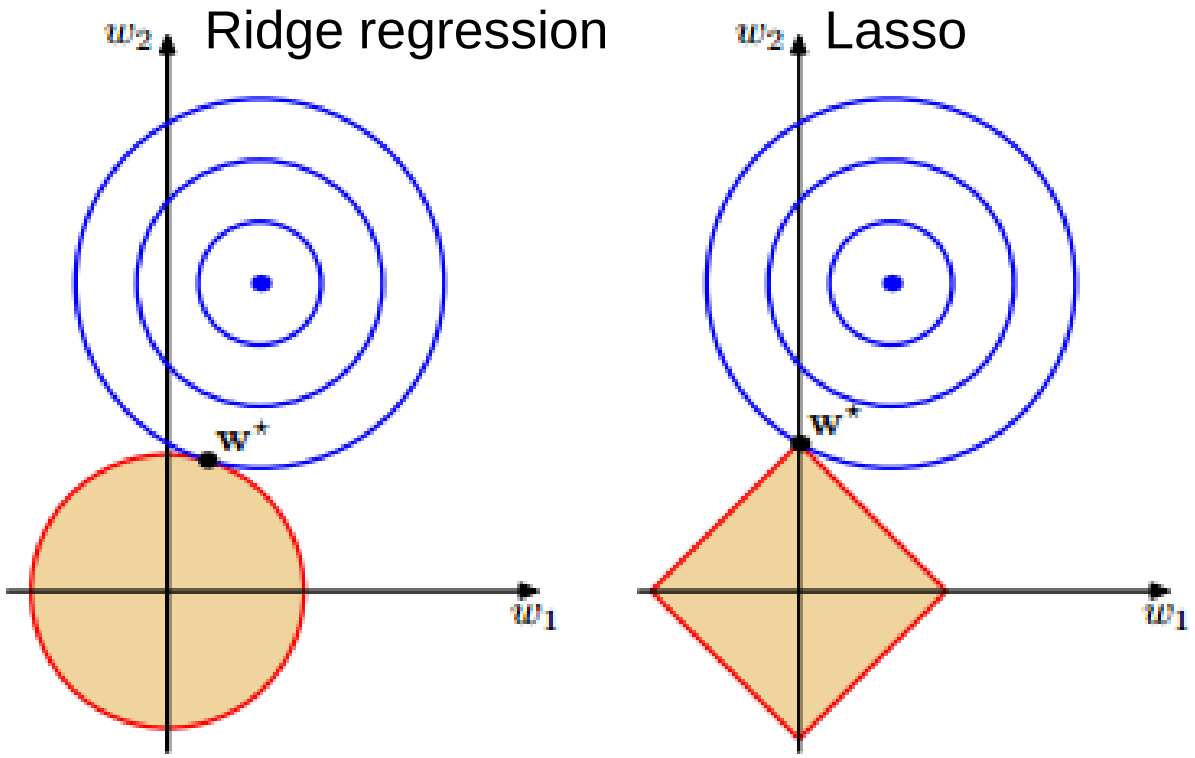
Lasso:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |w_j| = s$$

Ridge regression:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} w_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} w_j^2 = s$$

The small difference between ridge regression and lasso: a parameter space perspective



Both methods produce values for the parameters that are closer to zero

The set of allowed values slightly differs for both methods

=> will lead to different models

Regularization in other models: the example of logistic regression

$$\min_{\boldsymbol{w}} -\sum_{i=1}^{n} \left(y_i \log p_{\boldsymbol{w}}(\boldsymbol{x}_i) + (1 - y_i) \log(1 - p_{\boldsymbol{w}}(\boldsymbol{x}_i)) \right) + \lambda \sum_{j=1}^{p} w_j^2$$