Machine learning

David Zarzoso
Aix Marseille Univ, CNRS, Centrale Med, M2P2 UMR 7340, Marseille

Beyond univariate linear regression

Multivariate linear regression
The Bias-Variance decomposition
Shrinkage methods
- Ridge regression
- Lasso
- Elastic net

Some widely accepted notations for data

Input variable $\rightarrow$ X

Ex: time, distance to obstacle, city, colour, …

Output variable

- Quantitative Y. Ex: temperature, pressure, speed, angle, price
- Categorical G. Ex: animal, flower

p-dimensional variable: $X^T = (X_1, X_2, \dots, X_p)$, $Y^T = (Y_1, Y_2, \dots, Y_p)$

Ex: position, force, velocity, …

Observation $\rightarrow$ x, y or g

Ex: time=2s, distance to obstacle=10m, city=Houston, colour=red, …

For N observations, the i-th observation is x_i, y_i, g_i, with $i = 1, 2, \dots, N$. An observation can be a scalar or a vector

We can write in matrix form N observations of a p-dimensional variable X $\rightarrow$

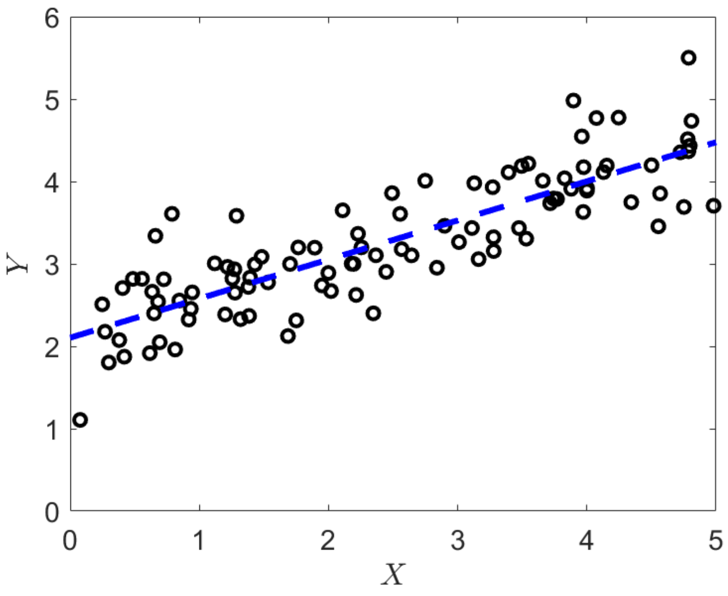
$$\begin{array}{cc} \text{Dimension} \longrightarrow & \\ \boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} & \Big\downarrow \text{Observation} \end{array}$$

Multivariate linear regression as a prediction model

Let us assume without any loss of generality that we want to predict a real-valued output $Y \in \mathbb{R}$

The most common choice for the loss function is the *squared error loss* $\mathcal{L}(Y, \hat{Y}) = \left\| Y - \hat{Y} \right\|^2$

We try to find the function f making the assumption that it is linear, i.e. we can write $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$



$$\mathcal{L}(Y, \hat{Y}) = \mathcal{L}(Y, X; \beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 = (y - \boldsymbol{X}\beta)^T (y - \boldsymbol{X}\beta)$$

$$y^T = (y_1, y_2, \ldots, y_N) \qquad \beta^T = (\beta_0, \beta_1, \ldots, \beta_p) \qquad x_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{iN})$$

$$\frac{\partial \mathcal{L}(Y, X; \beta)}{\partial \beta} = -2\boldsymbol{X}^T (y - \boldsymbol{X}\beta) = 0 \Rightarrow \boldsymbol{X}^T (y - \boldsymbol{X}\beta) = 0 \Rightarrow \hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T y$$

Recovering the *standard* linear regression

Taking again the expression for the expected prediction error and assuming that $y \in \mathbb{R}$

$$EPE(f) = \mathbb{E}\left[\mathcal{L}\left(Y, \hat{Y}\right)\right] = \mathbb{E}\left\|Y - \hat{Y}\right\|^2 = \iint \left(y - f(x)\right)^2 \mathcal{P}(x, y)dxdy$$

We can replace $f(x)$ by $x^T\beta$

$$EPE(x^T\beta) = \iint (y - x^T\beta)^2 \mathcal{P}(x, y)dxdy \xrightarrow{d/d\beta} \iint x(y - x^T\beta)\mathcal{P}(x, y)dxdy$$

Finally, we get

$$\hat{\beta} = \left(\mathbb{E}(XX^T)\right)^{-1}\mathbb{E}(XY)$$

The expression we found earlier $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$ is the same as $\hat{\beta} = \left(\mathbb{E}(XX^T)\right)^{-1}\mathbb{E}(XY)$ when the expectations are estimated using the average over the training data.

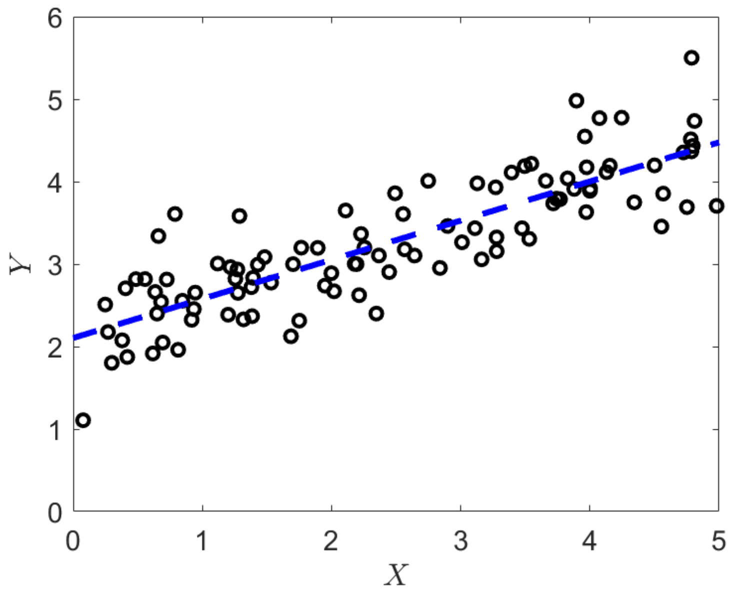
Multivariate linear regression as a prediction model

Let us assume now that we want to predict a real-valued vector $Y \in \mathbb{R}^m$

We keep as loss function the *squared error loss* $\mathcal{L}(Y, \hat{Y}) = \left\| Y - \hat{Y} \right\|^2$

And we try to find the function f making the assumption that it is linear, i.e. we can write $f_k(X) = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk}$

$$\mathcal{L}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \mathcal{L}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{B}) = \sum_{k=1}^{m} \sum_{i=1}^{N} \left(y_{ik} - \beta_{0k} - \sum_{j=1}^{p} x_{ij} \beta_{jk} \right)^2$$



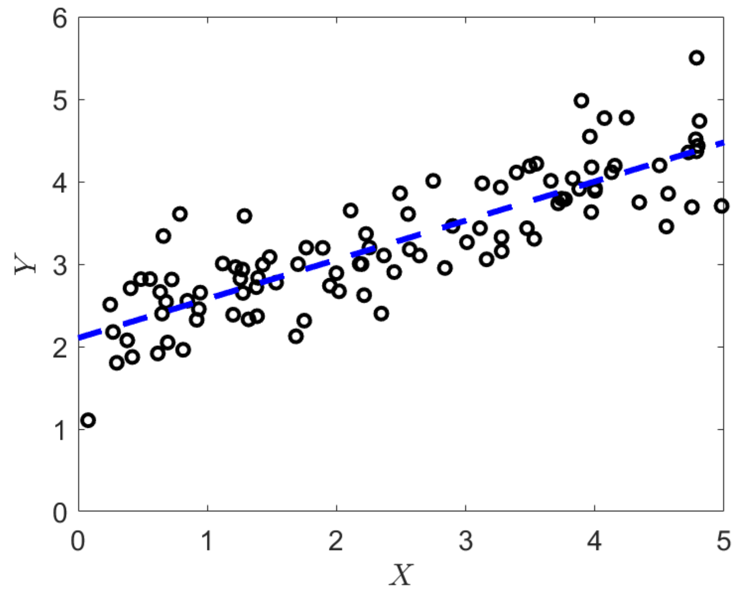
Multivariate linear regression as a prediction model

Let us assume now that we want to predict a real-valued vector $Y \in \mathbb{R}^m$

We keep as loss function the *squared error loss* $\mathcal{L}(Y, \hat{Y}) = \left\|Y - \hat{Y}\right\|^2$

And we try to find the function f making the assumption that it is linear, i.e. we can write $f_k(X) = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk}$

$$\mathcal{L}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \mathcal{L}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{B}) = \sum_{k=1}^{m} \sum_{i=1}^{N} \left(y_{ik} - \beta_{0k} - \sum_{j=1}^{p} x_{ij} \beta_{jk}\right)^2 = tr[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})]$$



$$y^T = (y_1, y_2, \dots, y_N) \qquad \beta^T = (\beta_0, \beta_1, \dots, \beta_p) \qquad x_i^T = (1, x_{i1}, x_{i2}, \dots, x_{iN})$$

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{Nm} \end{pmatrix} \quad \boldsymbol{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1N} \\ \vdots & \ddots & \vdots \\ \beta_{m1} & \cdots & \beta_{Nm} \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

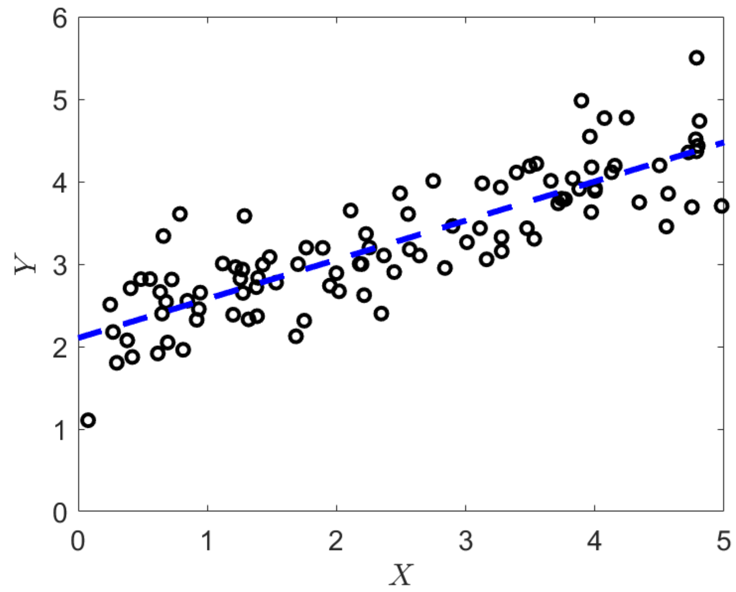
Multivariate linear regression as a prediction model

Let us assume now that we want to predict a real-valued vector $Y \in \mathbb{R}^m$

We keep as loss function the *squared error loss* $\mathcal{L}(Y, \hat{Y}) = \|Y - \hat{Y}\|^2$

And we try to find the function f making the assumption that it is linear, i.e. we can write $f_k(X) = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk}$

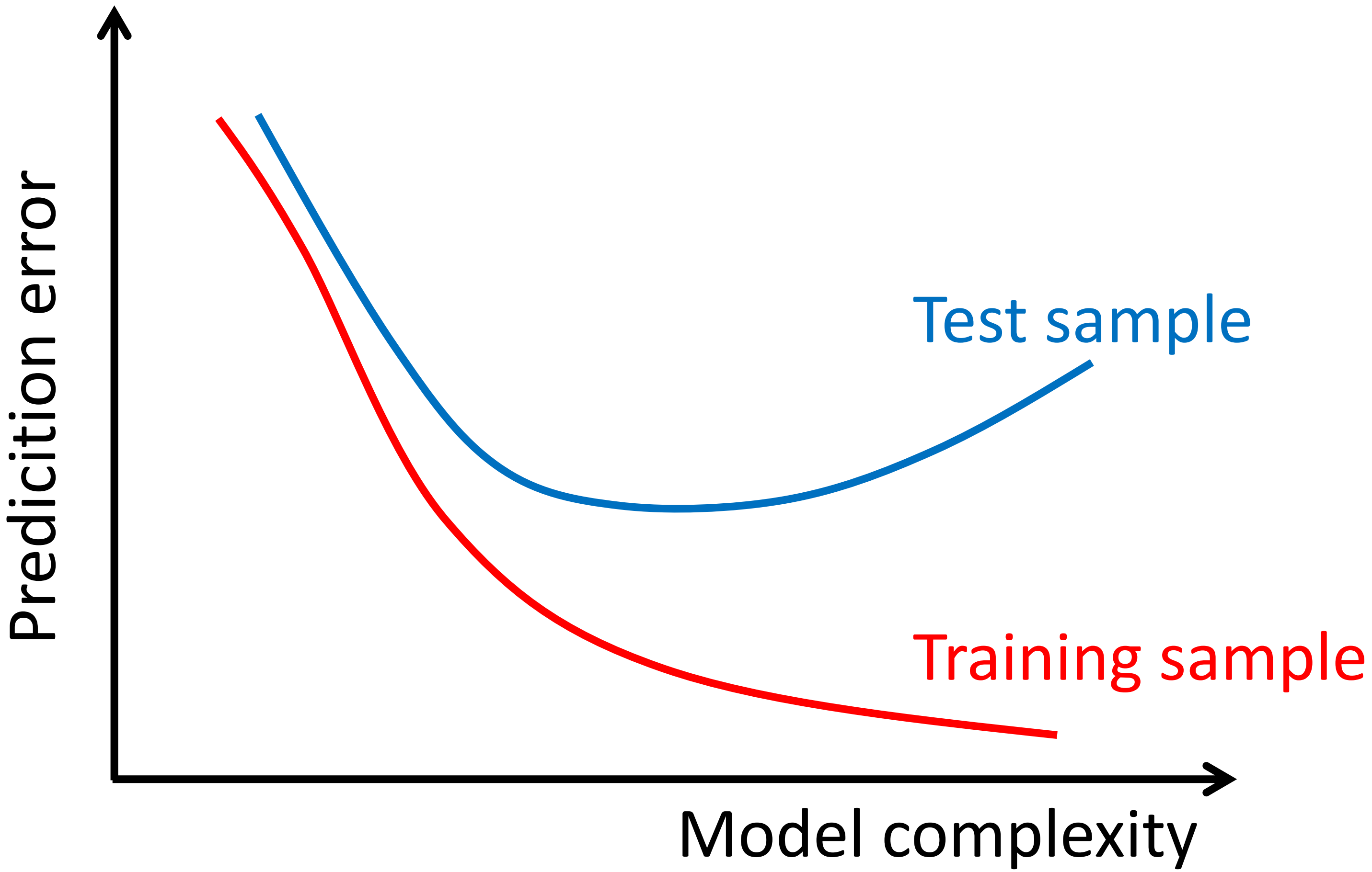
$$\mathcal{L}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \mathcal{L}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{B}) = \sum_{k=1}^{m} \sum_{i=1}^{N} \left(y_{ik} - \beta_{0k} - \sum_{j=1}^{p} x_{ij}\beta_{jk}\right)^2 = tr[(\boldsymbol{Y} - \boldsymbol{XB})^T(\boldsymbol{Y} - \boldsymbol{XB})]$$



$$y^T = (y_1, y_2, \dots, y_N) \qquad \beta^T = (\beta_0, \beta_1, \dots, \beta_p) \qquad x_i^T = (1, x_{i1}, x_{i2}, \dots, x_{iN})$$

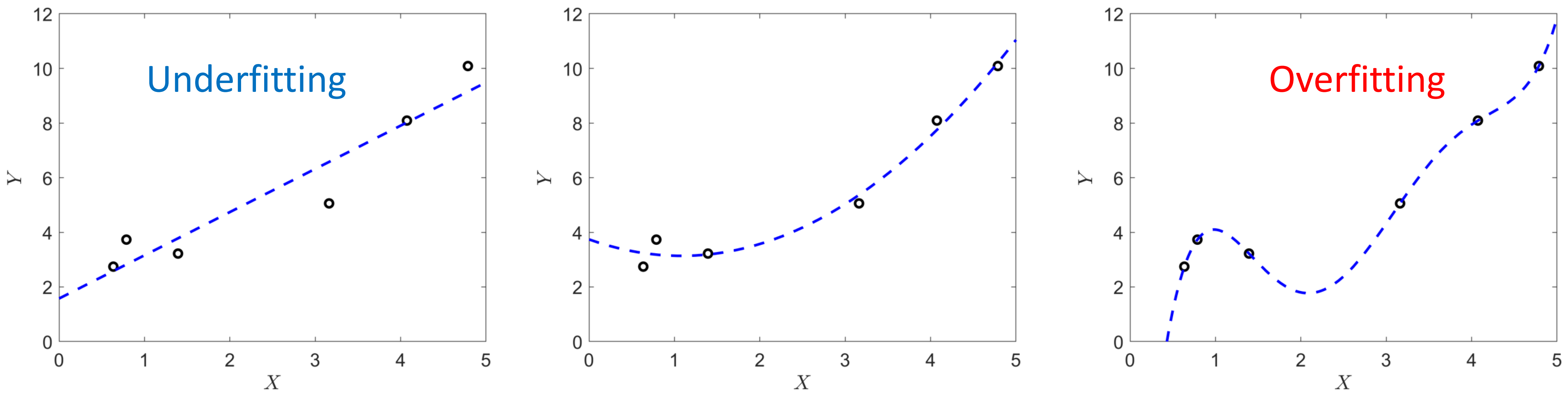
$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{Nm} \end{pmatrix} \quad \boldsymbol{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1N} \\ \vdots & \ddots & \vdots \\ \beta_{m1} & \cdots & \beta_{Nm} \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{B})}{\partial \boldsymbol{B}} = -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{XB}) = 0 \Rightarrow \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{XB}) = 0 \Rightarrow \widehat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$


Prediciition error
Test sample
Training sample
Model complexity

The bias-variance tradeoff

Complexity (variance)

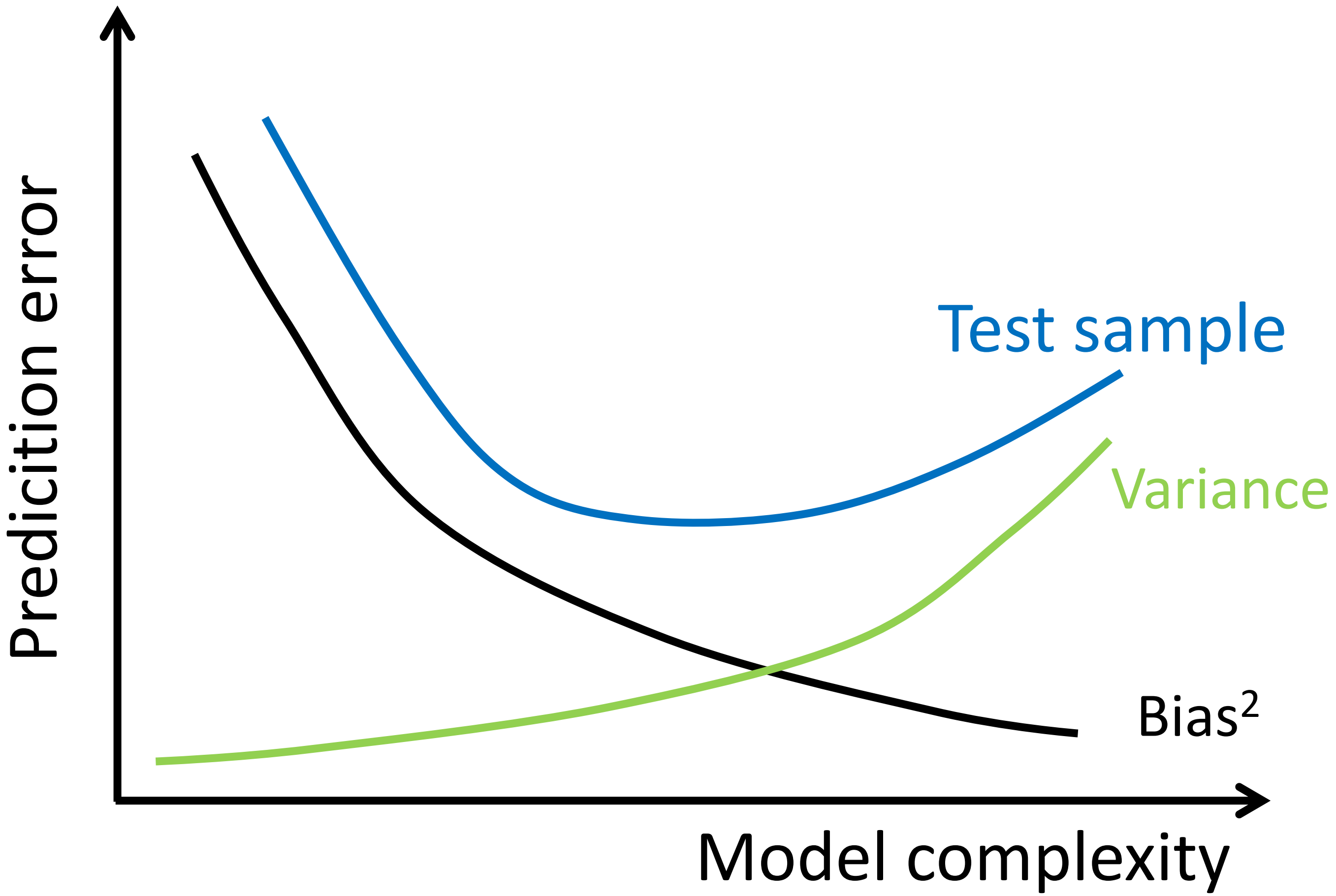


$Y = \beta_0 + \beta_1 X$

$Y = \beta_0 + \beta_1 X + \beta_2 X^2$

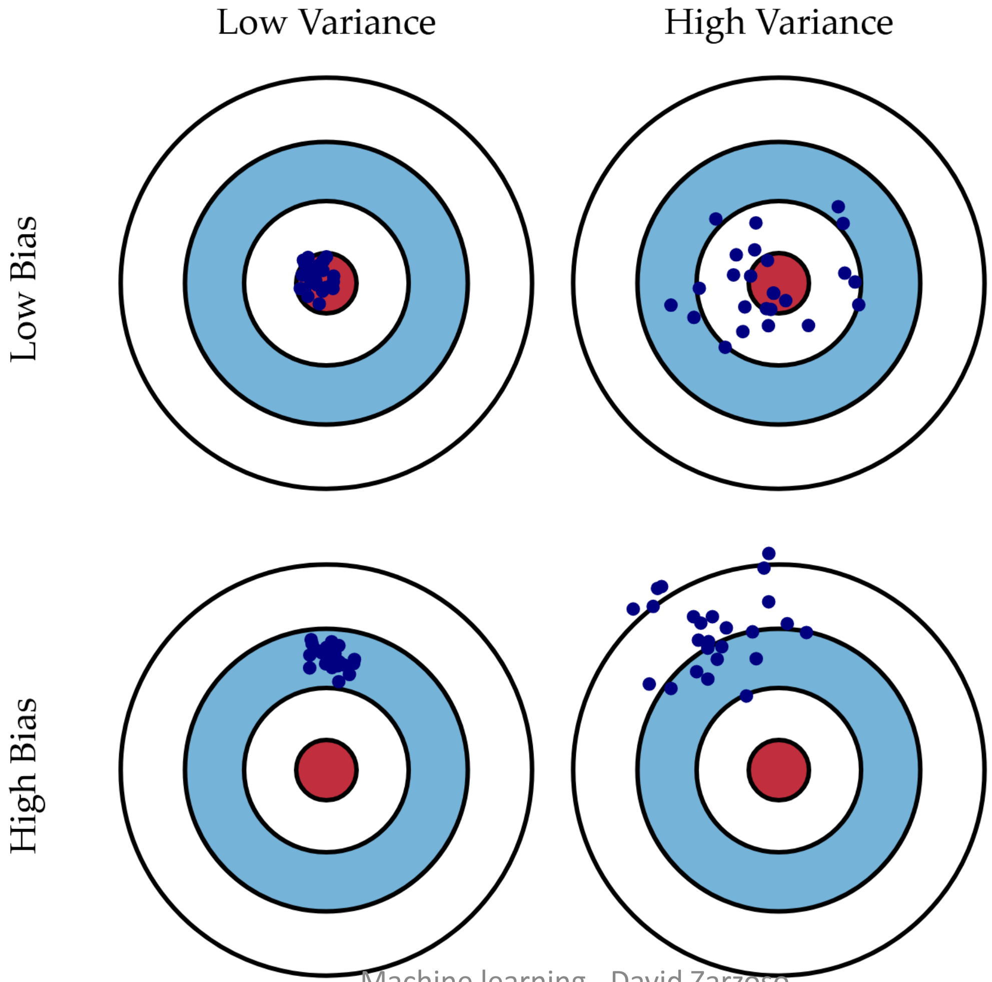
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5$

Bias



Test sample
Variance
Bias^2
Predicition error
Model complexity

Schematic illustration of bias-variance

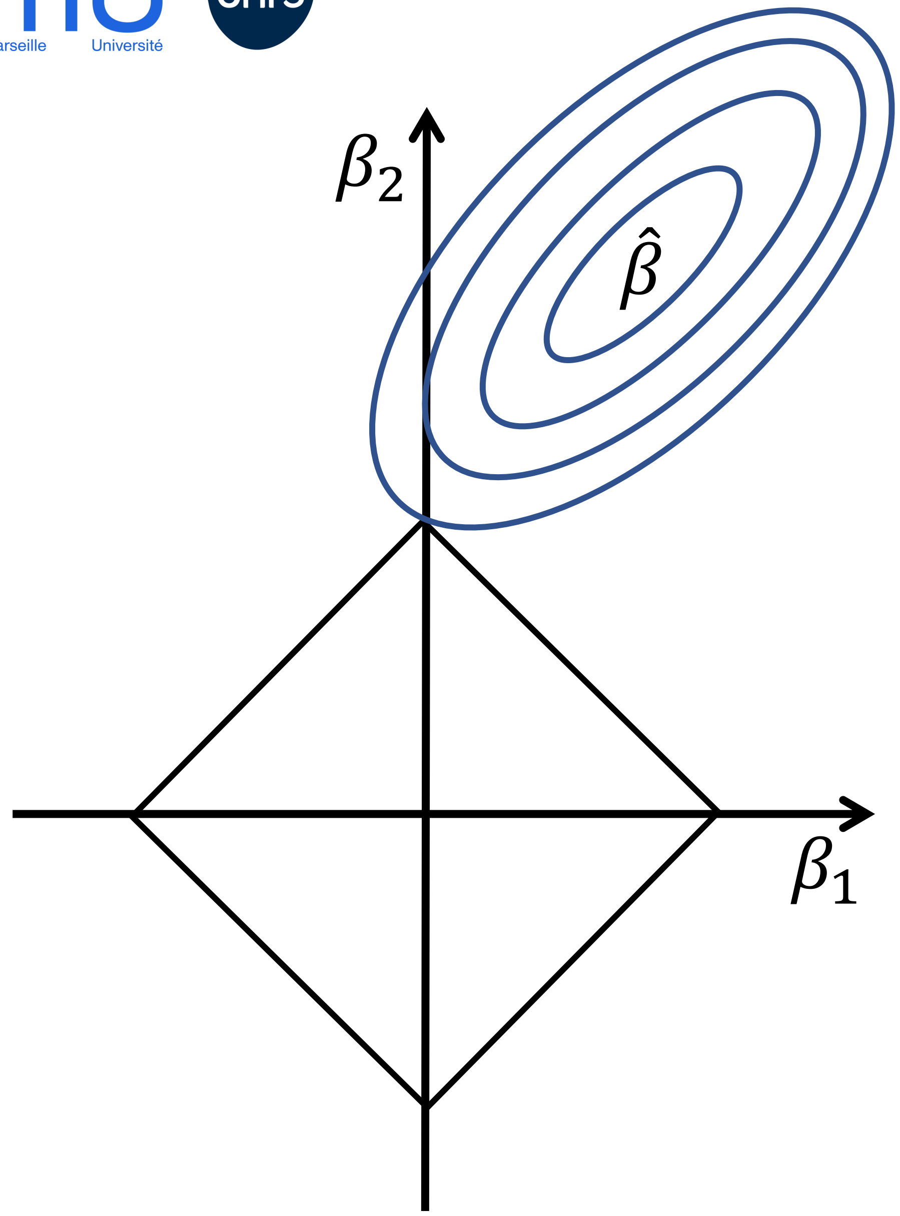


Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

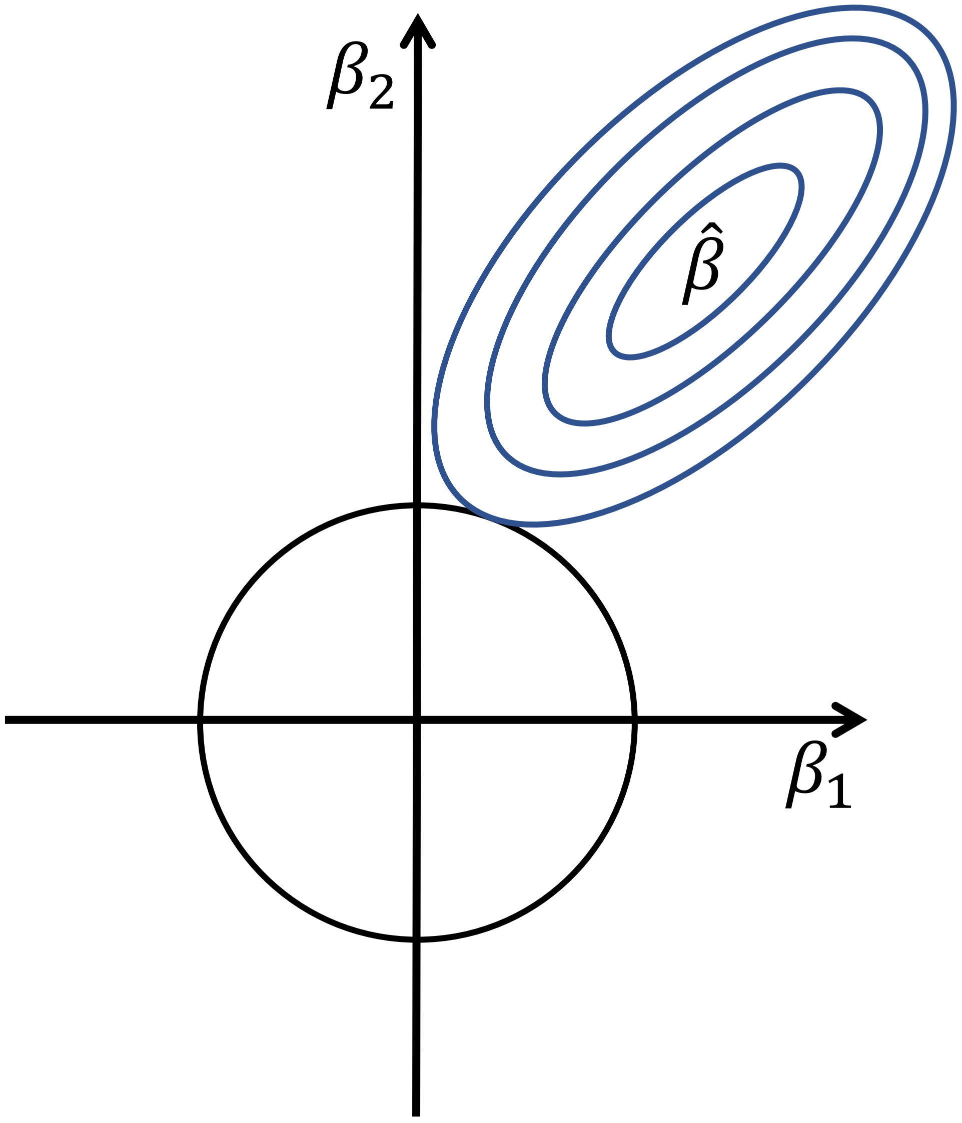
Need for regularization

Interpretation → Smaller subset of predictors which exhibit the strongest effects

Prediction accuracy → Shrink some parameters (reduce the variance and increase the bias)



Lasso



Ridge