LASSO

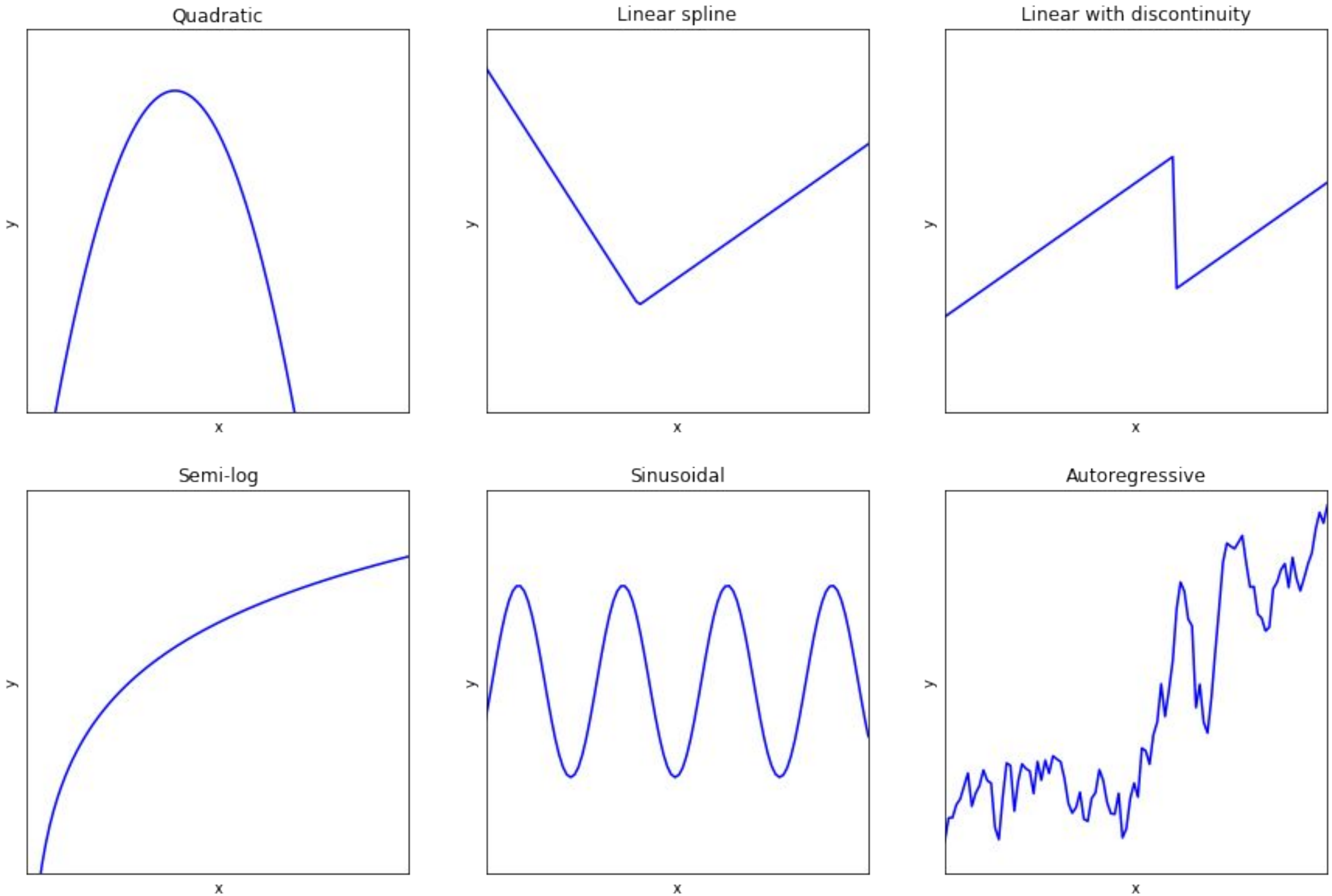
Regularized regression and model selection

Tim Padvitski

Linear regression and OLS


OL' RELIABLE
$y = X\beta + \varepsilon$

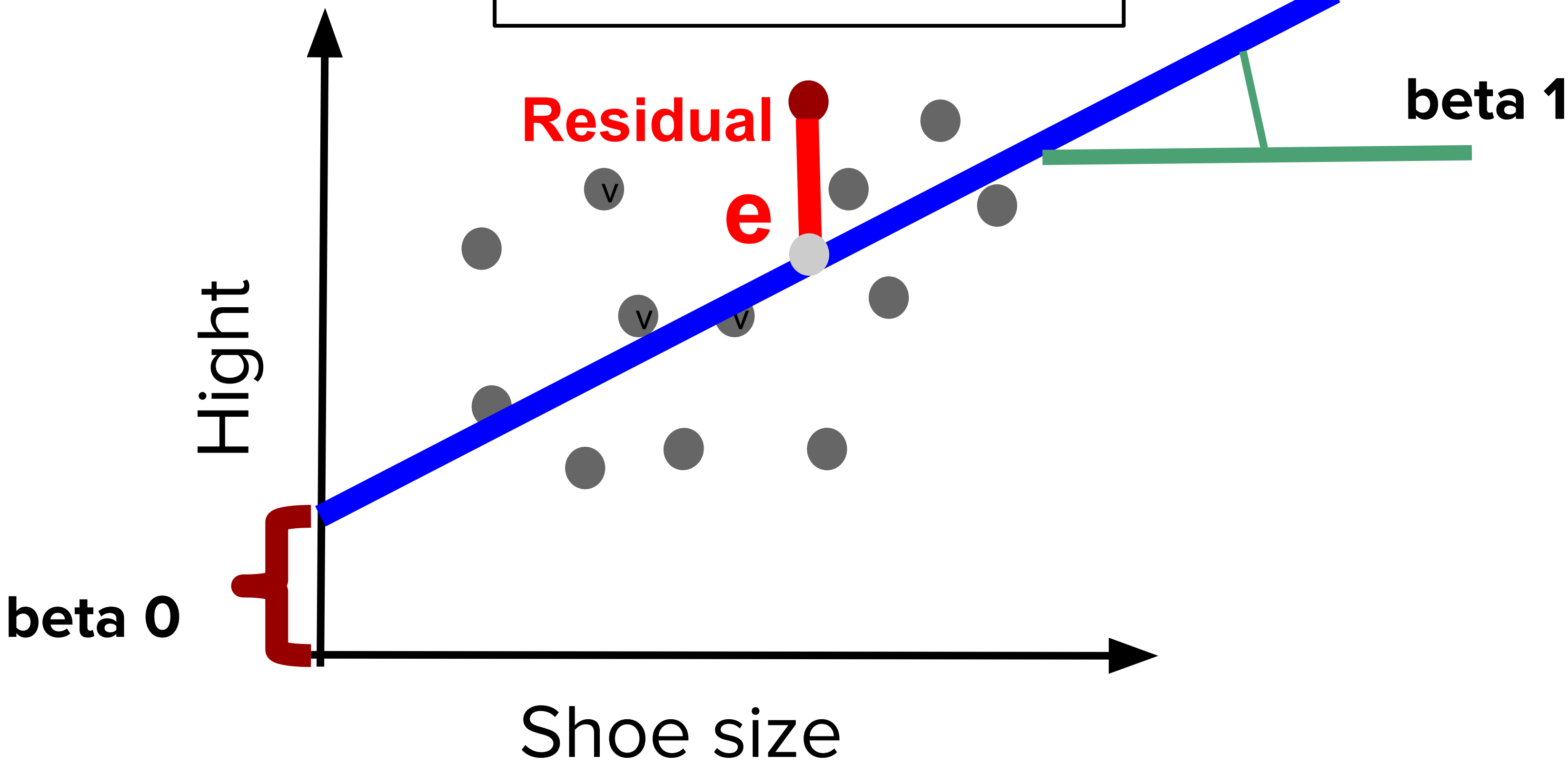
linear models do not require that all variables are explicitly linear



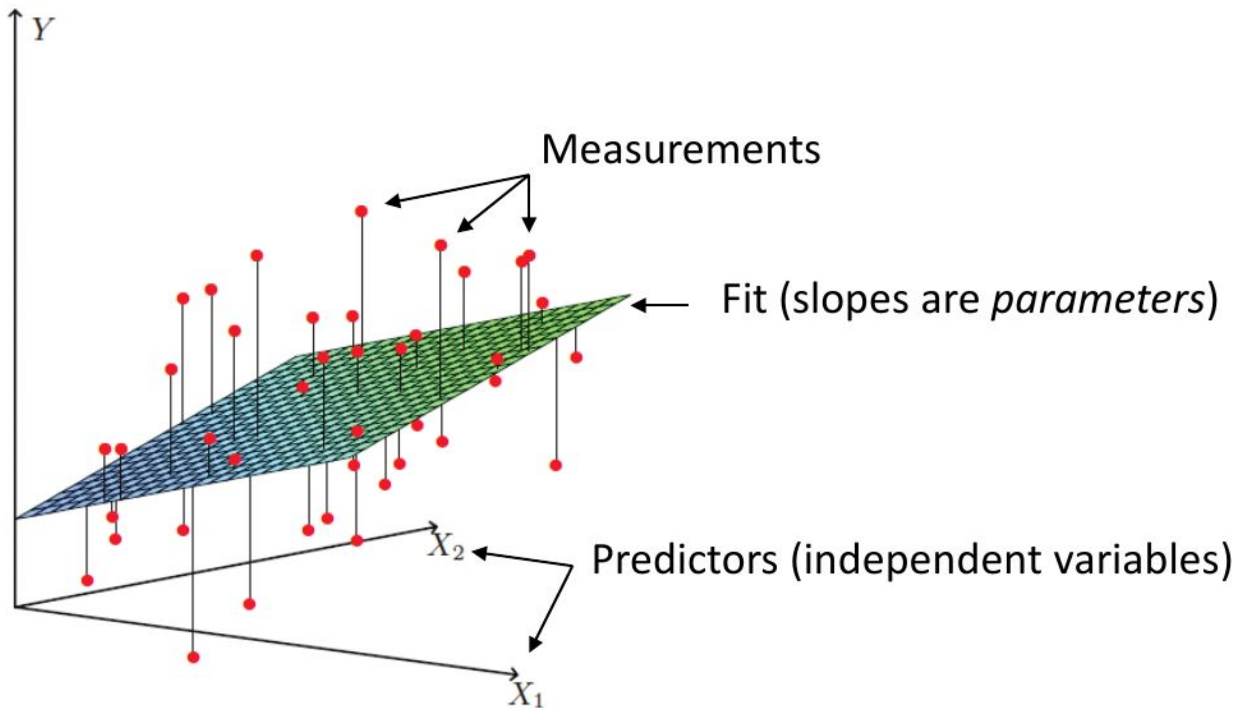
https://ryxcommar.com/2019/09/06/some-things-you-maybe-didnt-know-about-linear-regression/?fbclid=IwAR3v_ff9puRpQuLwxZdTmkQMi9ZBWnlWMivf7EZ91q11nXngNET_PzNH5ZE

Linear regression - 2 variables

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

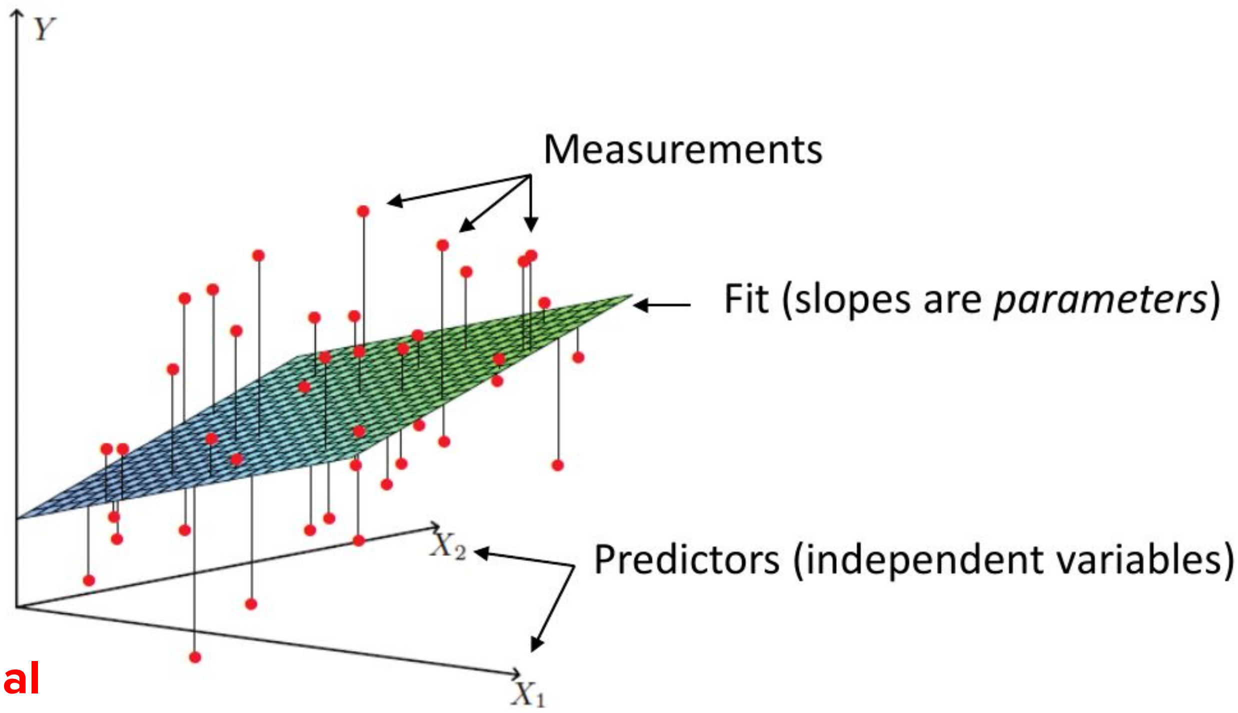


Linear regression : more than 2 variables



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots . + \beta_n X_n + \varepsilon$$

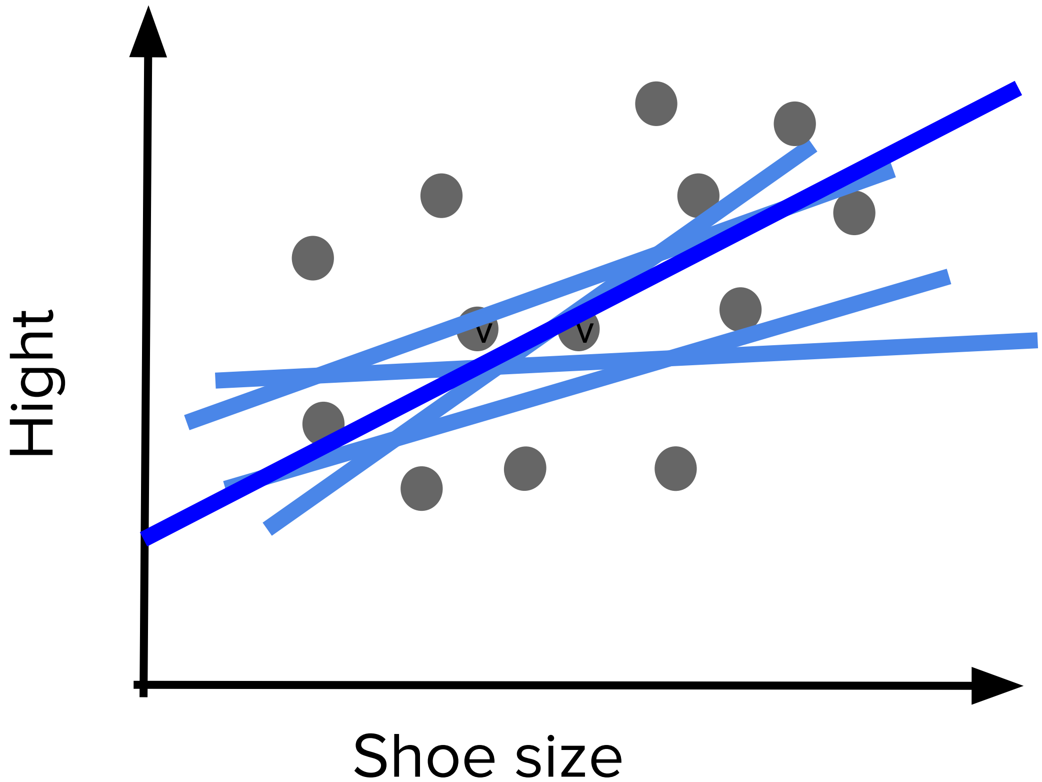
Linear regression : more than 2 variables



N-dimensional vector!

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear regression - fitting



Residual sum of squares : RSS

$$e_i = y_i - \hat{y}_i$$

Residual – difference between ith observed response value and ith predicted value from linear model

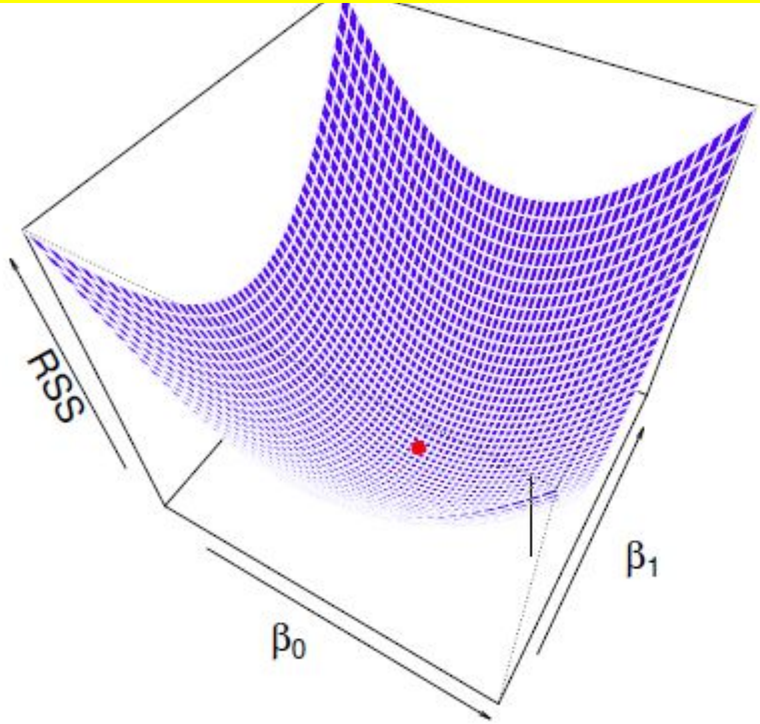
! minimize

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

$$\boxed{RSS(\beta) = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j}^{p}\beta_j x_{ij}\right)^2}$$

OLS estimator

OLS visually explained : http://setosa.io/ev/ordinary-least-squares-regression/



The value of *b* which minimizes RSS is called the **OLS (**ordinary least squares) **estimator for *β***

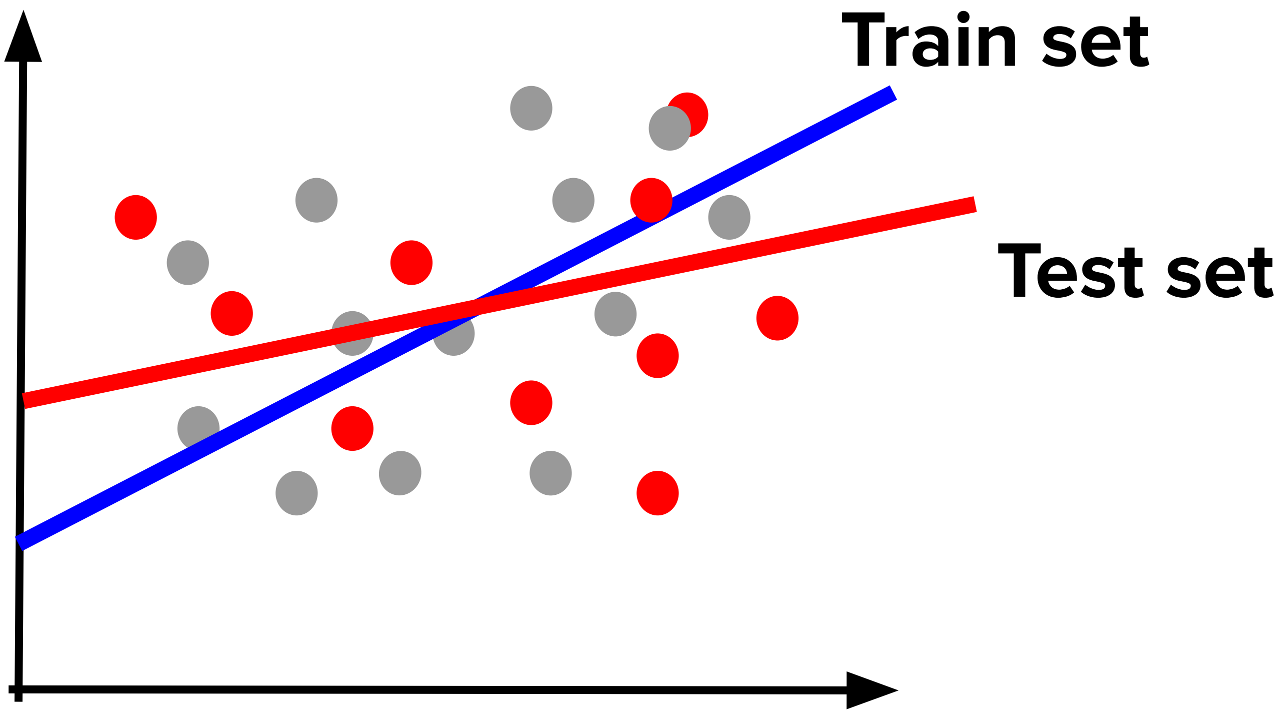
Solving linear models with OLS

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boxed{RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j}^{p} \beta_j x_{ij} \right)^2}$$

OLS visually explained : http://setosa.io/ev/ordinary-least-squares-regression/

New data points



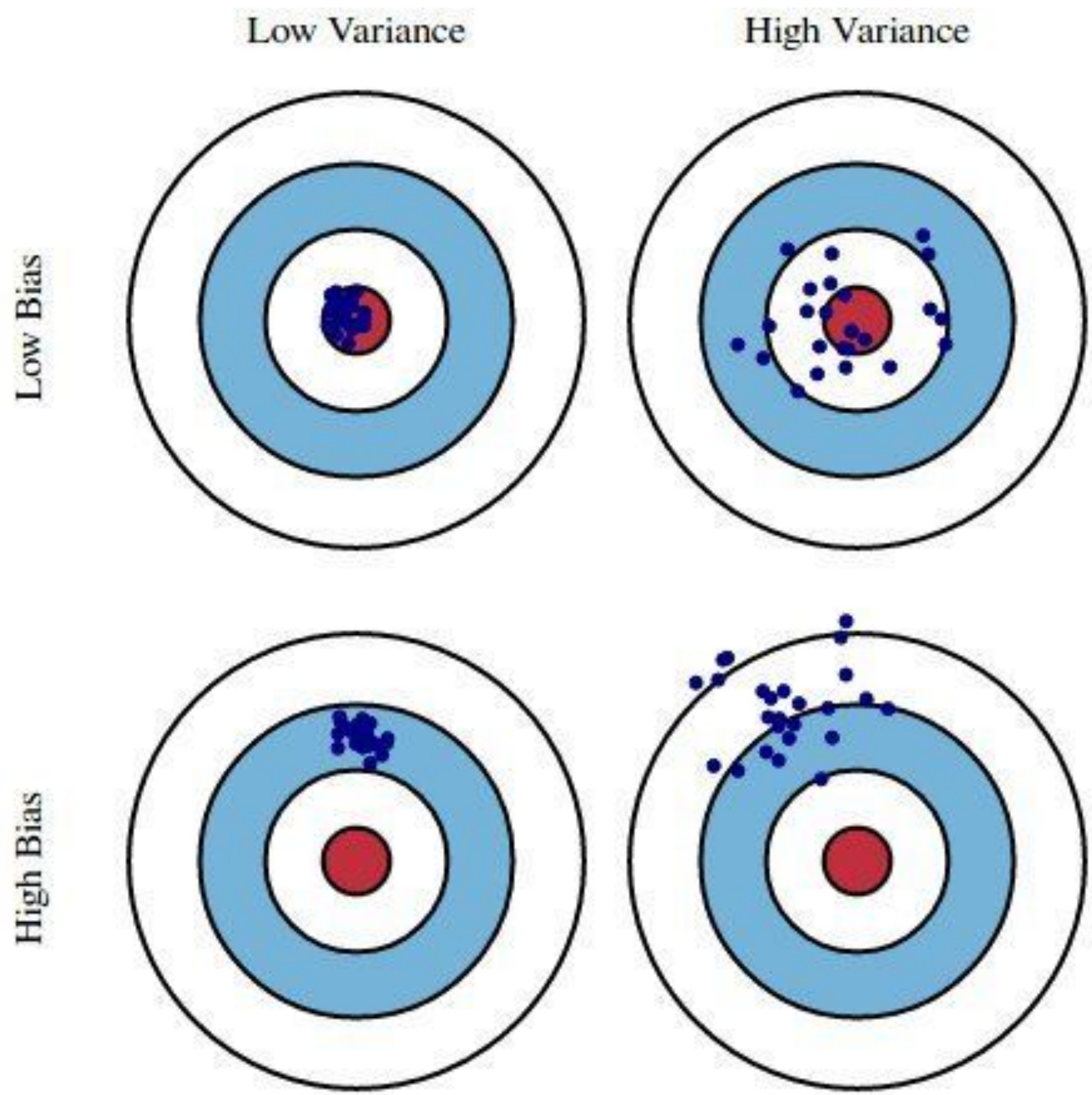
What is a good model ?

- As detailed as necessary
- As simple as possible

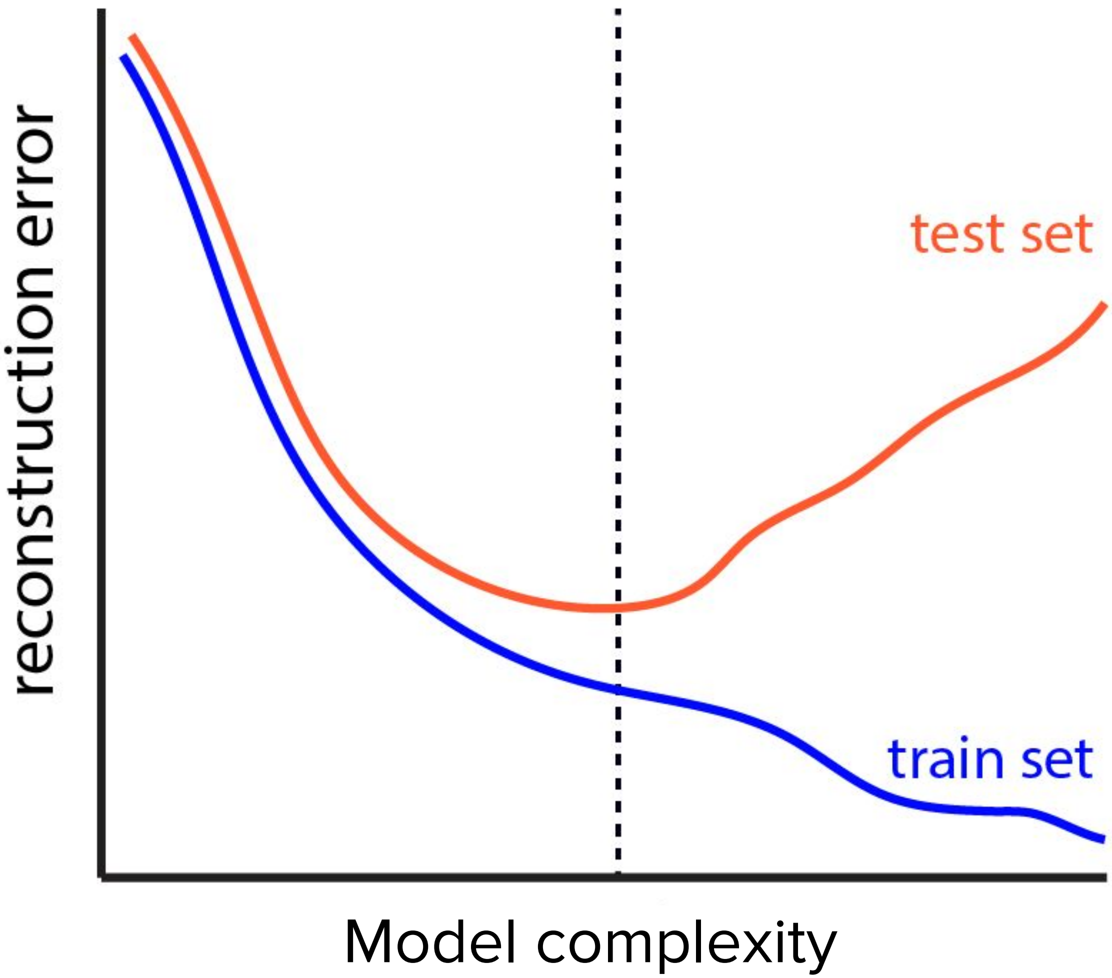
Bias-variance trade-off

Bias : deviation of average model from 'true' model

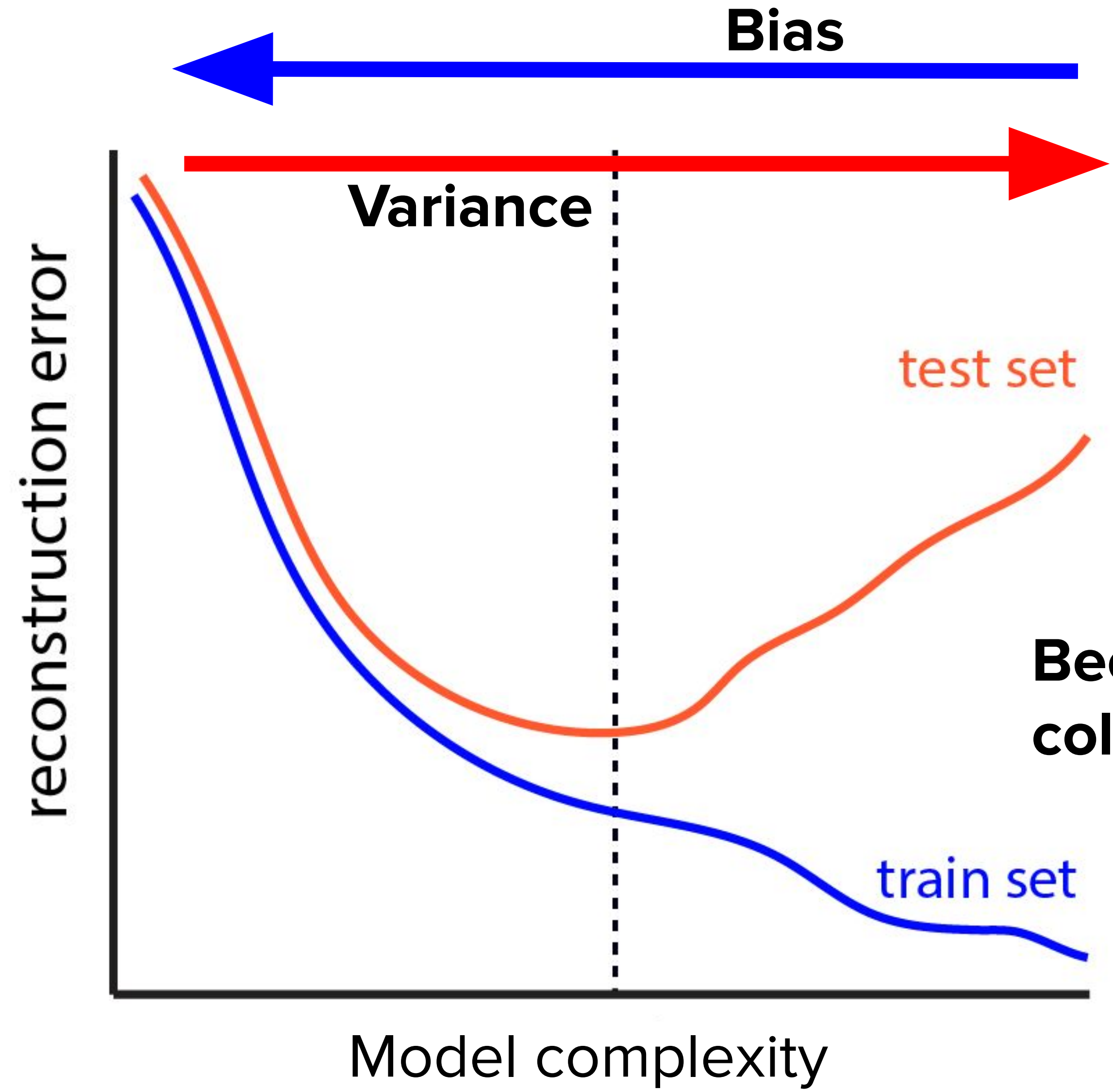
Variance : deviation between models learned from individual datasets



Bias-variance trade-off

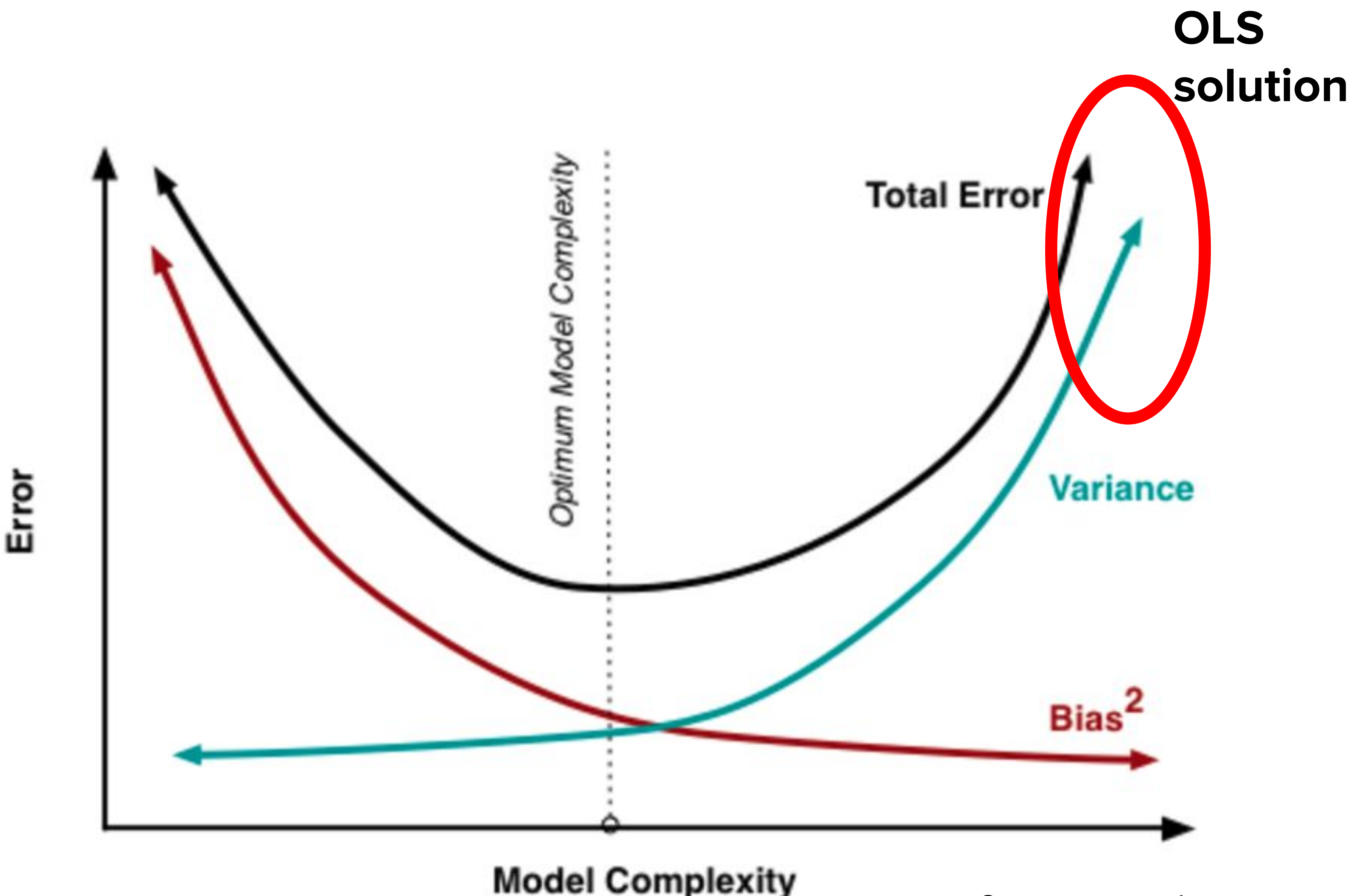


Bias-variance trade-off



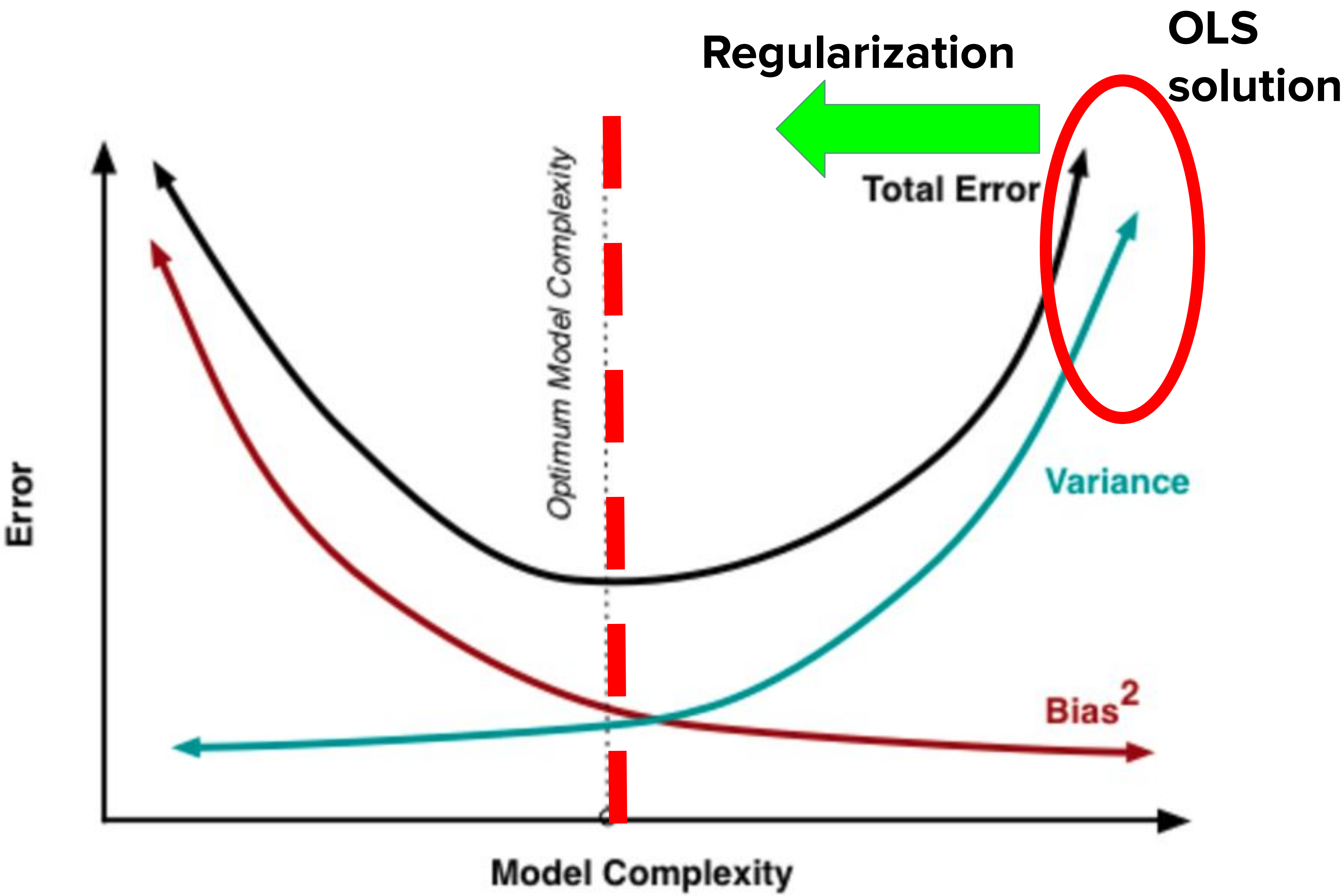
Because of collinearity

Bias-variance and total error



Source: researchgate.net

What is a good model ?



Issues with OLS solution

- OLS solution for regression has low bias but **high variance**:
    - Collinearity
    - Too many predictors.

- Impossible for $p > n$ problems (often: $p >> n$!).
  Poor solution for $p \sim n$

Regularization approaches

Stepwise regularized regression

Backward:

- start with full model
- remove least important parameters
- works only when N > p

Forward:

- start with intercept
- iteratively add most predictive parameters
- works even when p >> N

Stepwise approaches are **‘greedy’** and solutions are **not stable**

Penalized regression

RIDGE

LASSO

And many other :

- Partial least squares
- Principal component regression
- etc.

Ridge - L_2 penalty

$$f(X) = \beta_0 + \sum_{j}^{p} \beta_j X_j$$

Note! linear regression model stays the same!

$$\beta^{ridge} = argmin_\beta \left\{ \underbrace{\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} {\beta_j}^2}_{\text{penalty for coefficients}} \right\}$$

RSS

penalty for coefficients

λ = weighting factor

LASSO* - L_1 penalty

$$f(X) = \beta_0 + \sum_{j}^{p} \beta_j X_j$$

Note! linear regression model stays the same!

$$\beta^{lasso} = argmin_\beta \left\{ \underbrace{\frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{penalty for coefficients}} \right\}$$

λ = weighting factor

* Least Absolute Shrinkage and Selection Operator

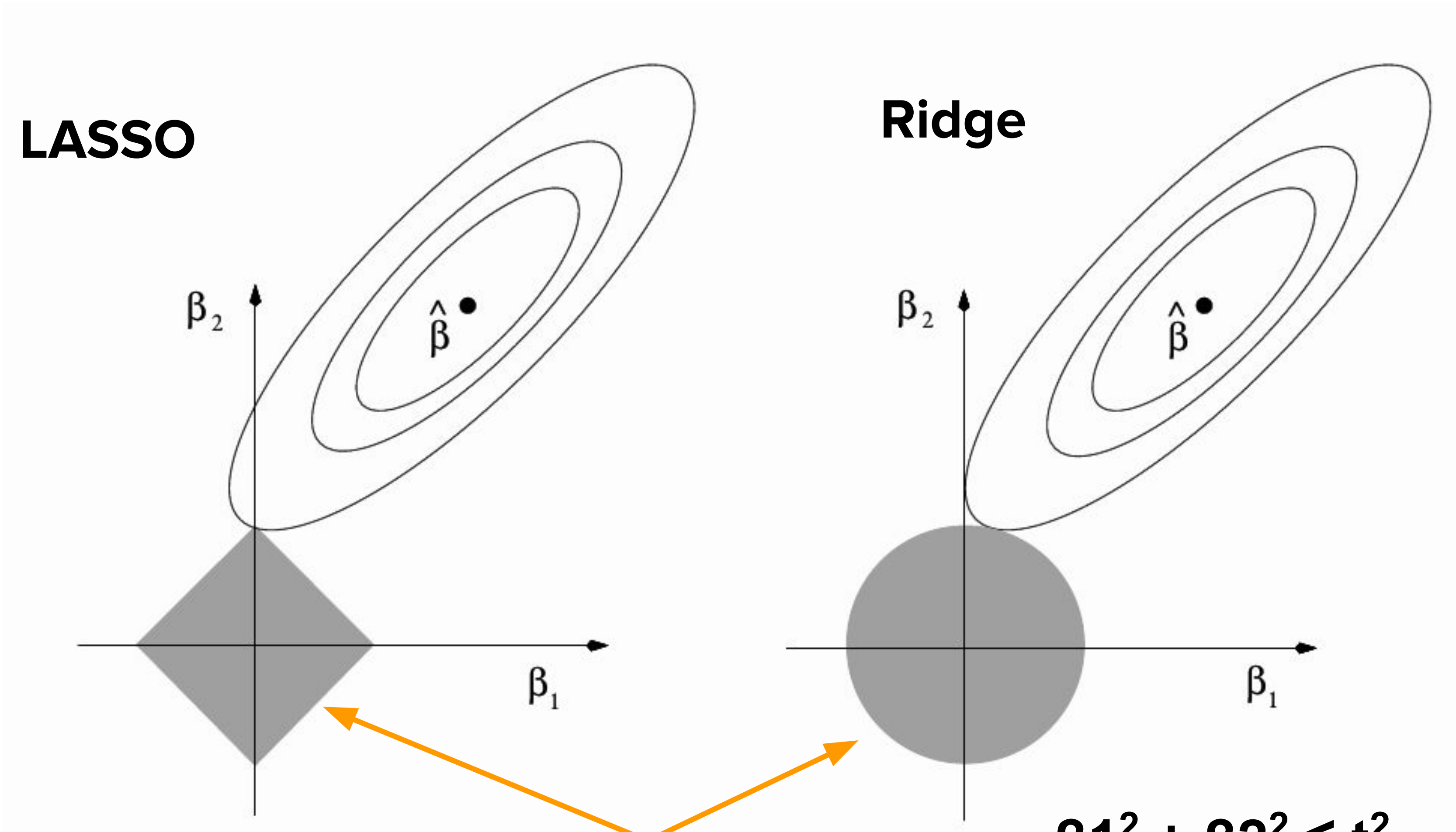
LASSO vs Ridge

$$\beta^{lasso} = argmin_{\beta}\left\{\underbrace{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{penalty for coefficients}}\right\}$$

$$\beta^{ridge} = argmin_{\beta}\left\{\underbrace{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda\sum_{j=1}^{p}\beta_j^{\,2}}_{\text{penalty for coefficients}}\right\}$$

λ = weighting factor (shrinkage coefficient)

LASSO sets many parameters to zero



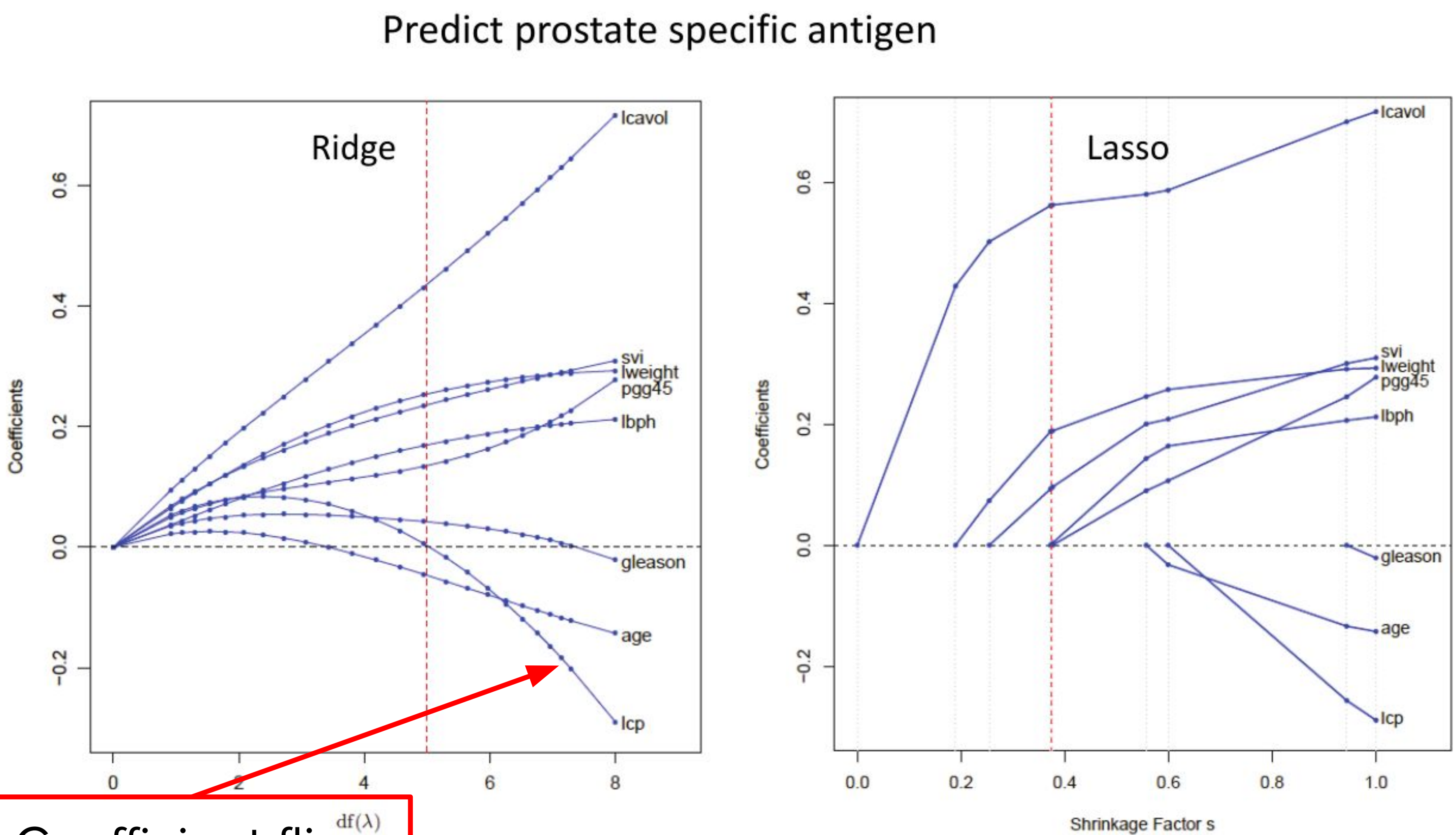
$|\beta 1| + |\beta 2| \leq t$

Regularizer

$\beta 1^2 + \beta 2^2 \leq t^2$

Friedman et.al. 2001

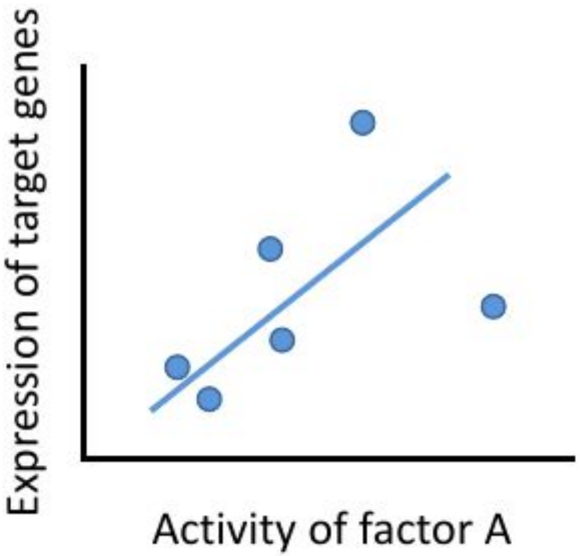
LASSO vs Ridge : prostate cancer

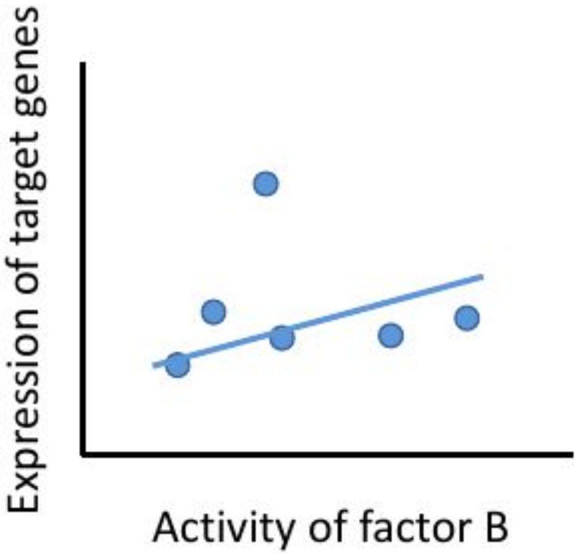


Coefficient flips the sign!

Friedman et.al. 2001

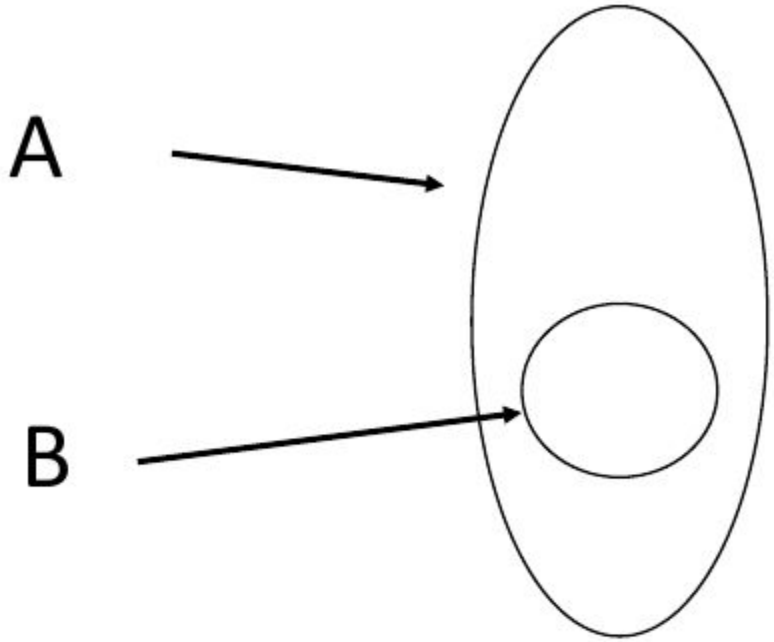
Example : two transcription factors





Positive correlation
in both cases.
Both are activators?

Example : two transcription factors

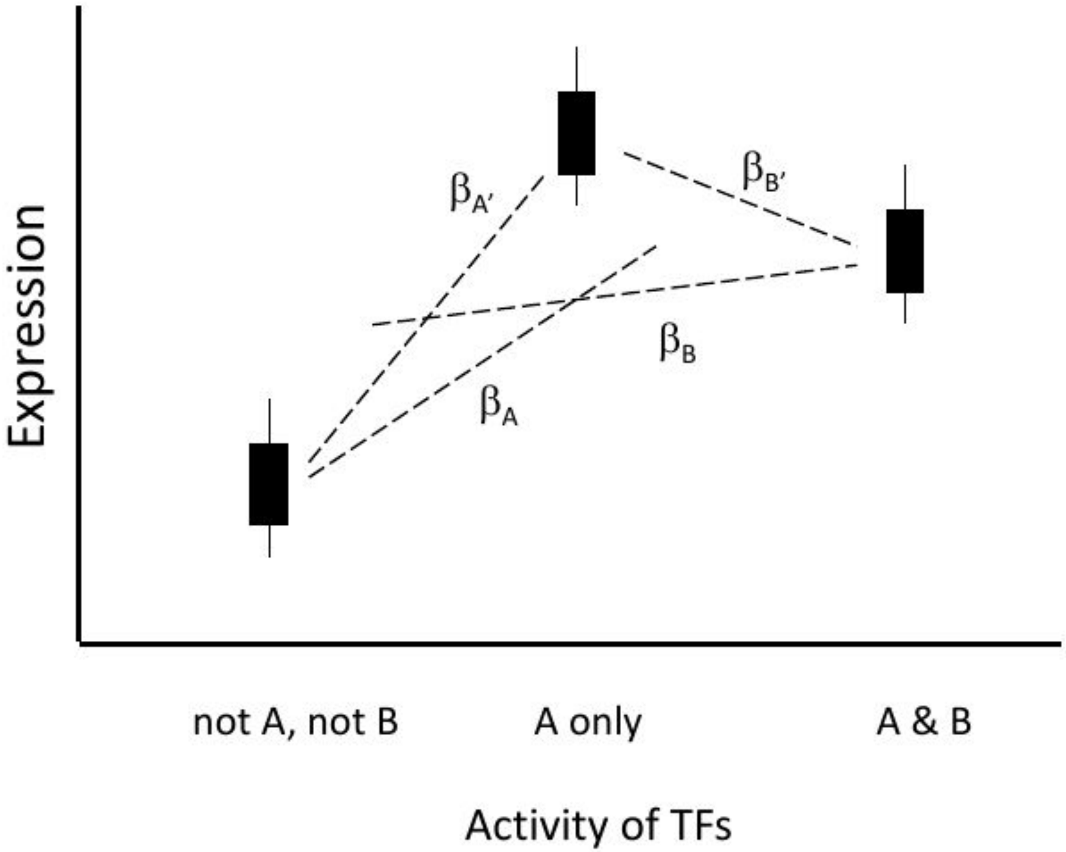


All target genes of B are also targets of A, but not the other way round!

A is an activator;
B is a repressor.

Friedman et.al. 2001

Example : two transcription factors



β_A, β_B: marginal effects
$\beta_{A'}$, $\beta_{B'}$: combined model

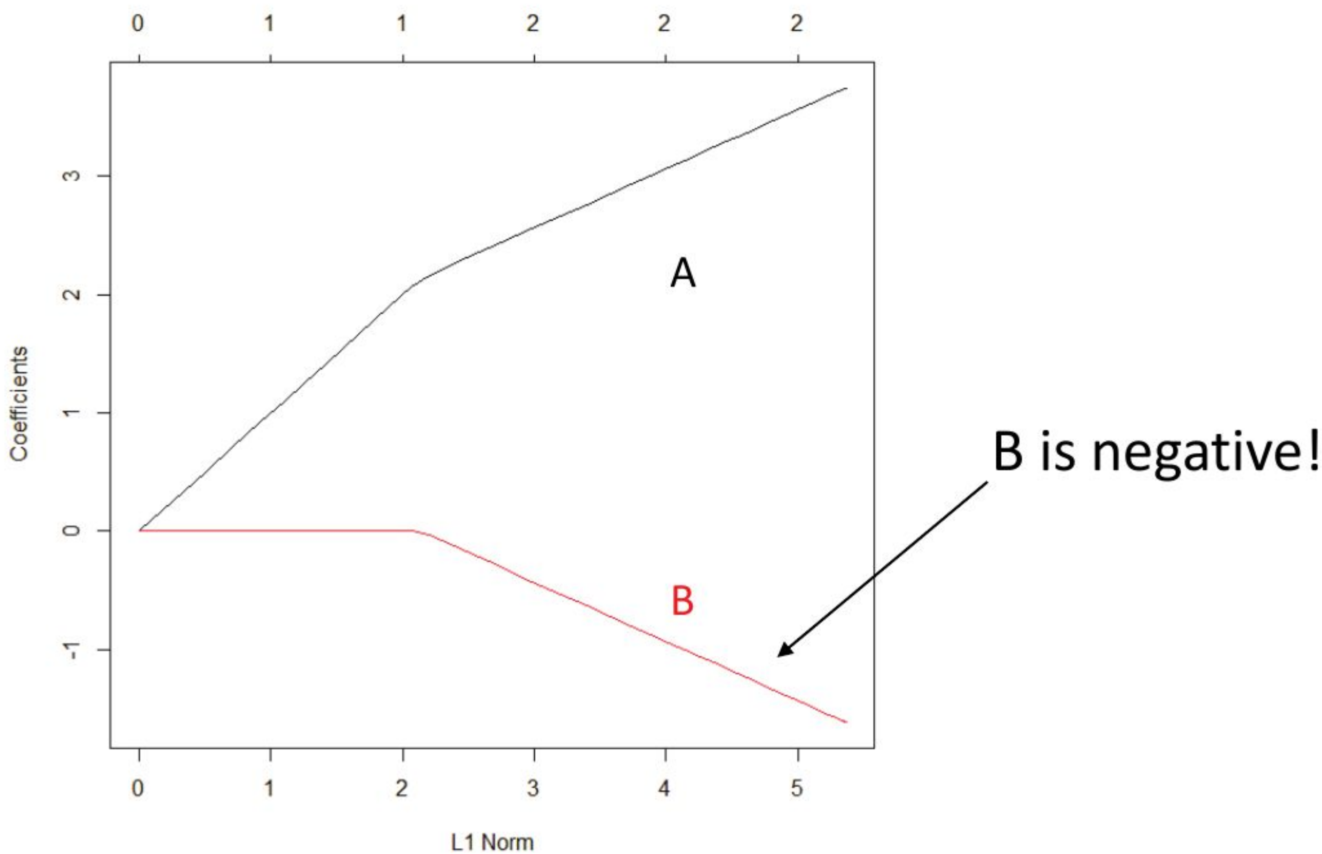
$$y = \beta_0 + \beta_A A$$

$$y = \beta_0 + \beta_B B$$

$$y = \beta_0 + \beta_{A'} A + \beta_{B'} B$$

Example : two transcription factors

Lasso results:



Friedman et.al. 2001

LASSO vs Ridge

$$\beta^{lasso} = argmin_\beta \left\{ \underbrace{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{penalty for coefficients}} \right\}$$

$$\beta^{ridge} = argmin_\beta \left\{ \underbrace{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} {\beta_j}^2}_{\text{penalty for coefficients}} \right\}$$

$\boxed{\lambda = \text{weighting factor (shrinkage coefficient)}}$

What is a good model ?

The new big problem:

? what is the optimal λ ?

Model selection approaches

Methods for performing model selection

AIC/BIC scores $AIC = -2logL + 2q$

A method for model selection that trades off goodness of fit with model complexity.

Methods for performing model selection

AIC/BIC scores

A method for model selection that trades off goodness of fit with model complexity.

Cross-validation

A method for choosing the value of λ that minimizes the generalization error on a held-out test validation set.

Methods for performing model selection

AIC/BIC scores

A method for model selection that trades off goodness of fit with model complexity.

Cross-validation

A method for choosing the value of λ that minimizes the generalization error on a held-out test validation set.

Stability methods

A class of methods for identifying the most “stable” model structure by using the idea that the same algorithm should yield similar results on similar datasets if the results are “stable”.

etc...

Methods for performing model selection

AIC/BIC scores **:(performs poorly in high dimensional setting**

A method for model selection that trades off goodness of fit with model complexity.

Cross-validation

A method for choosing the value of λ that minimizes the generalization error on a held-out test validation set.

Stability methods

A class of methods for identifying the most “stable” model structure by using the idea that the same algorithm should yield similar results on similar datasets if the results are “stable”.

etc...

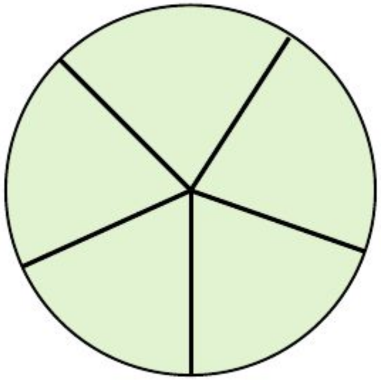
Cross validation

Algorithm:

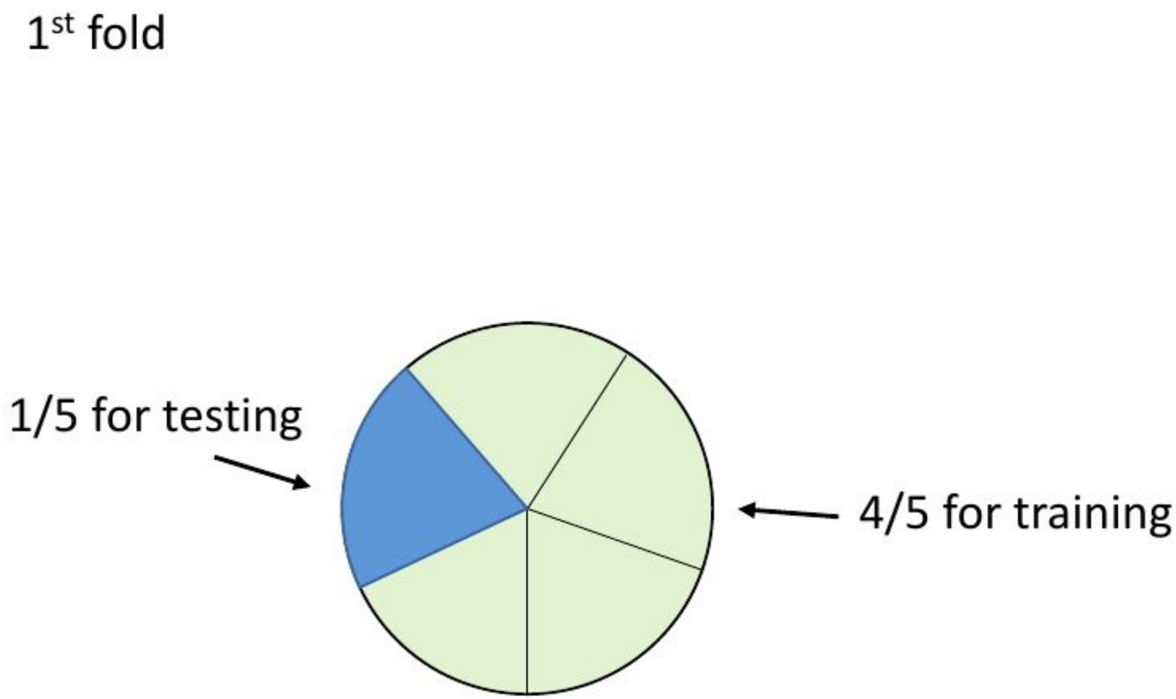
- Split data randomly into training and test set
- Fit based on training
- Test performance on test set
- Vary λ until optimal prediction

Example: 5-fold cross-validation

Split data into 5 random sub-sets

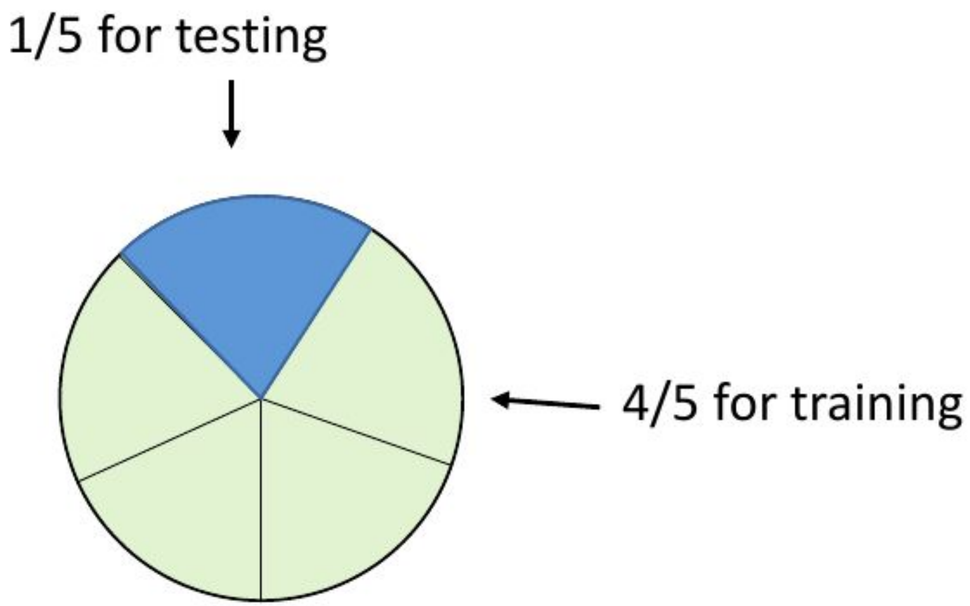


Example: 5-fold cross-validation



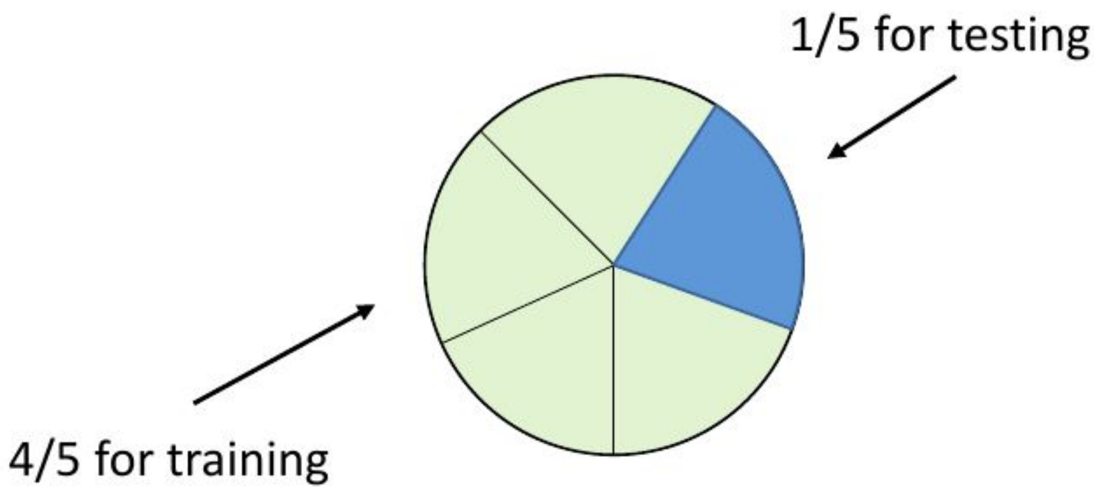
Example: 5-fold cross-validation

2nd fold



Example: 5-fold cross-validation

etc ...



Finally: average performance

Stability selection

Algorithm:

- Sample the data (take random subset)
- Train the model
- Repeat many times (e.g. 100) with different samples
- Get probability of inclusion (fraction of models) in which each parameter is included

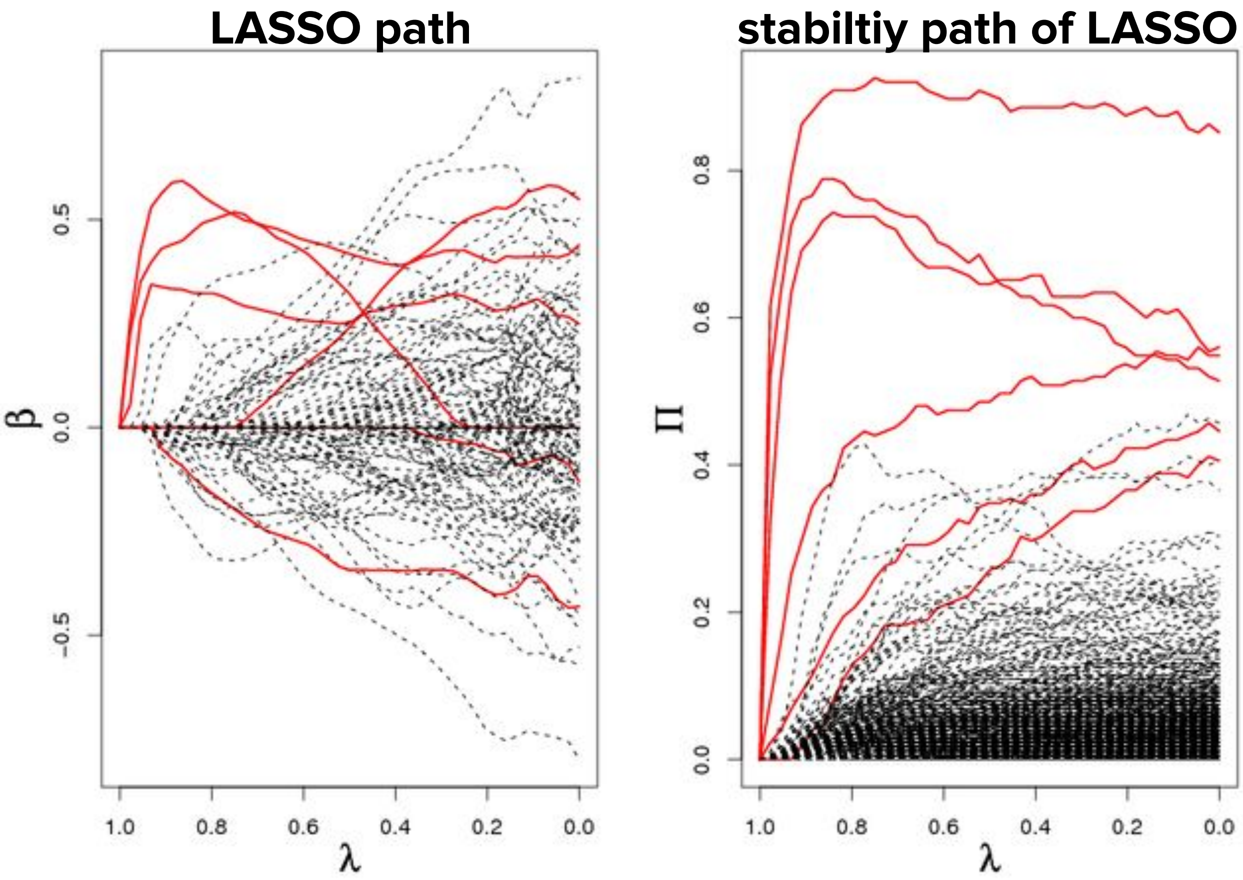
$$p_i = \frac{n_i}{N}$$

How many times was *i* included? (n_i)

Total number of permutations (N)

Bühlmann, Meinshausen 2009

Stability selection



Stability selection provides control over number of false positives

Take Home message

- Regularized regression can deal with p >> N problem and optimize bias-variance trade-off
- **Ridge** sets many coefficients *close* to zero ⇒ NO variable selection
- **Lasso** sets many coefficients to zero ⇒ variable selection
- Cross validation and stability selection help to select best model - amount of regularization λ