MATH 4280

Lecture Notes 4: Model selection

Nonlinear regression

- Consider a set of n data points

$$(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3),\ \cdots\ , (x_n, y_n)$$

- We fit the data to a general nonlinear function of the form

Generic regression

$$f(x) = f(x, \boldsymbol{\beta})$$

where $\boldsymbol{\beta} \in \mathbb{R}^m$ and $m < n$ underdetermined

- The parameter is obtained by minimizing

curve fitting regression

$$E_2(\boldsymbol{\beta}) = \sum_{k=1}^{n} (f(x_k, \boldsymbol{\beta}) - y_k)^2$$

- We can find $\boldsymbol{\beta} \in \mathbb{R}^m$ by solving

classical linear regression

$$\frac{\partial E_2}{\partial \beta_j} = \sum_{k=1}^{n} (f(x_k, \boldsymbol{\beta}) - y_k) \frac{\partial f}{\partial \beta_j} = 0 \quad j = 1, 2, 3, \cdots, m$$

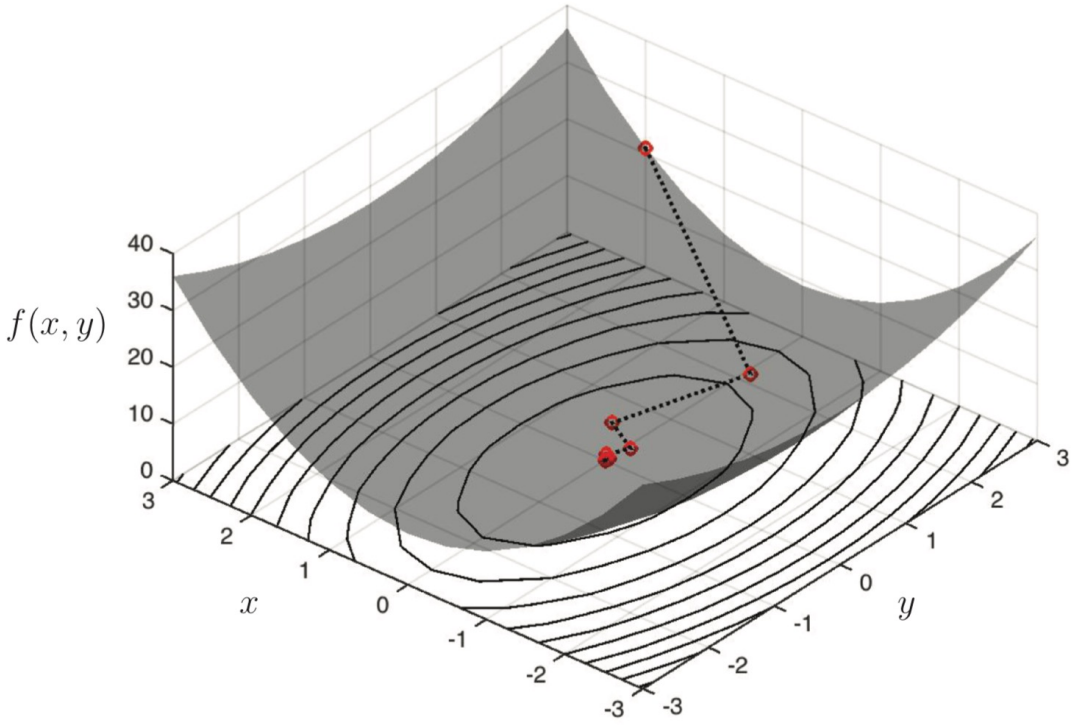
Gradient descent method

- To minimize the function $f(x)$, we solve $\nabla f(x) = 0$
- Perform the iteration

$$\mathbf{x}_{k+1}(\delta) = \mathbf{x}_k - \delta \nabla f(\mathbf{x}_k)$$

- To find the best δ, we minimize the function

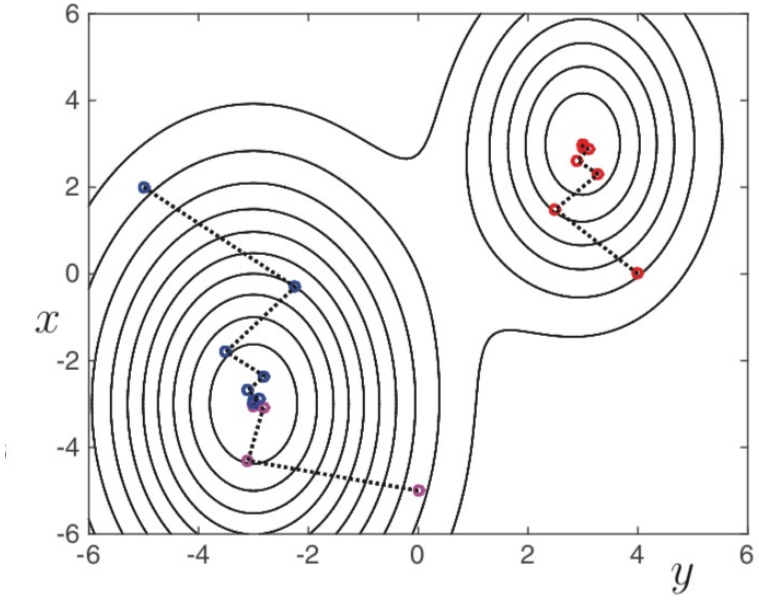
$$F(\delta) = f(\mathbf{x}_{k+1}(\delta))$$



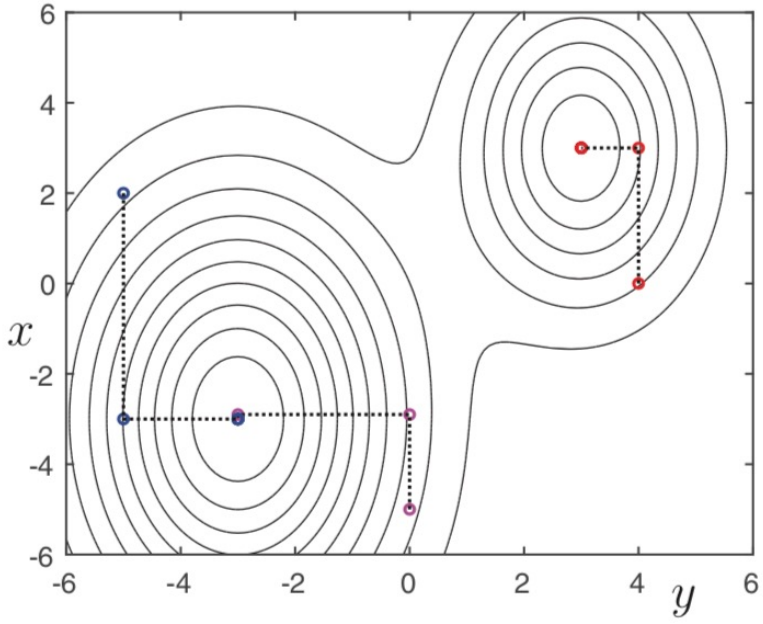
Alternating descent

- Idea: minimizing one variable at a time, keeping other variables fixed

Note that the choice of initial guess is important



Gradient descent



Alternating descent

Model selection

- The error metric may not be a good indicator, as more parameters will give smaller errors
- These additional parameters may have no meaning Over-fitting
- The use of sparsity can give a good model selection strategy

Example

$$\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\beta}) \text{ of } (4.4)$$

- We consider the following linear model fitting problem

$$\begin{bmatrix} | & | & | & \cdots & | \\ 1 & x_j & x_j^2 & \cdots & x_j^{p-1} \\ | & | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{100}) \end{bmatrix}$$

- This gives an overdetermined linear system $Ax = b$
- Assume that the data is noisy and has the following form

$$f(x) = x^2 + \mathcal{N}(0, \sigma)$$

- We consider the above as a model selection problem

- We use two ways to solve the problem
- The first way is the least-squares fit

For overdetermined systems $\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$

Error alone as a metric is potentially problematic since almost any method can produce a reliable, low-error model

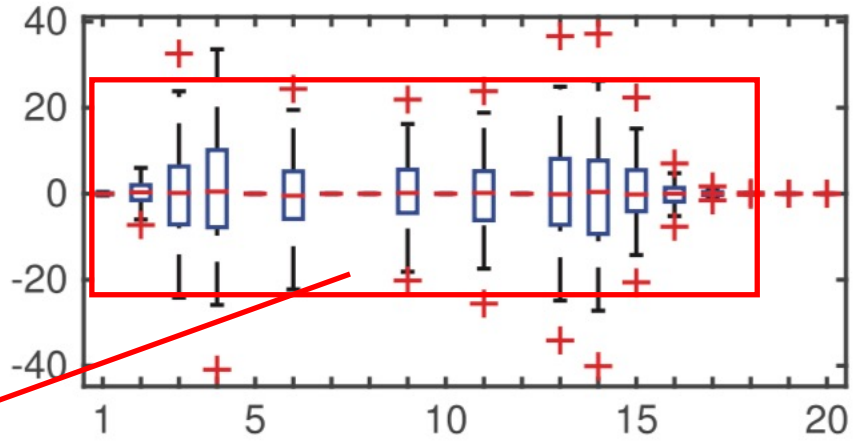
- The second way is

For underdetermined systems $\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 + \lambda_1 \|\mathbf{x}\|_1$

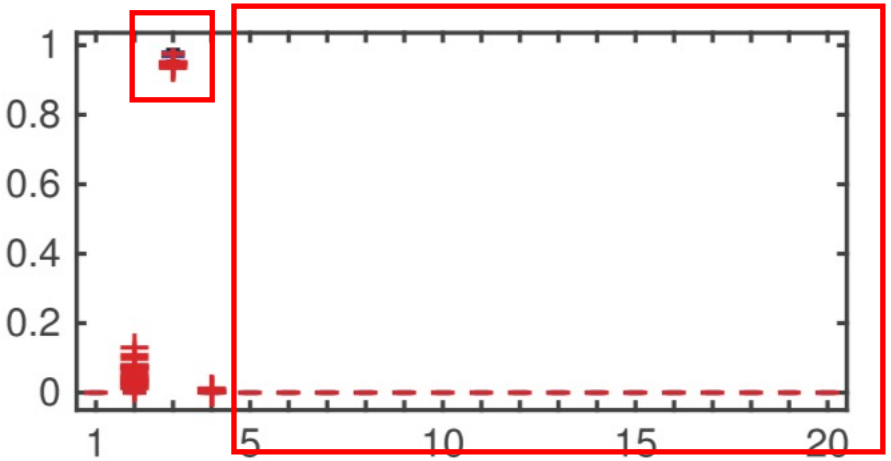
The additional term $\lambda_1 \|\mathbf{x}\|_1$ is called regularization

It is also called LASSO (least absolute shrinkage and selection operator)

Boxplots for errors on f using 100 realizations



20 degress of polynomial



Significant variability in the loading values for the strictly l2 based method, and low variability for l1 weighted methods.

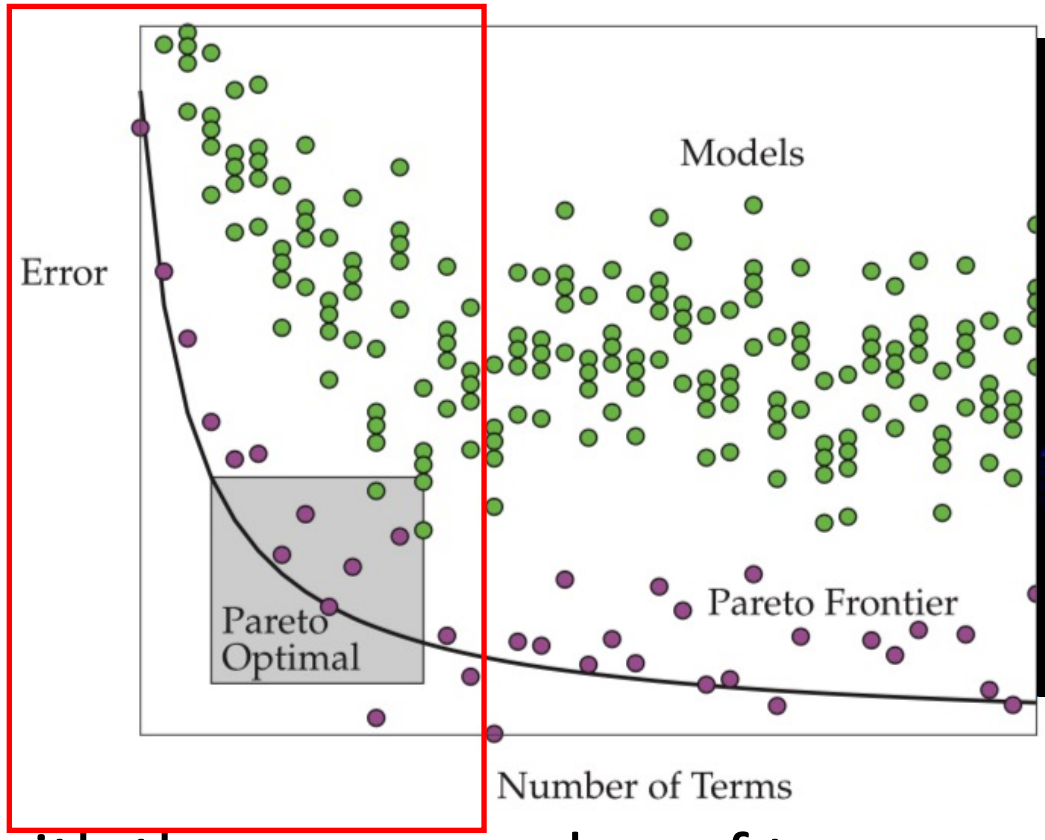
Least-squares

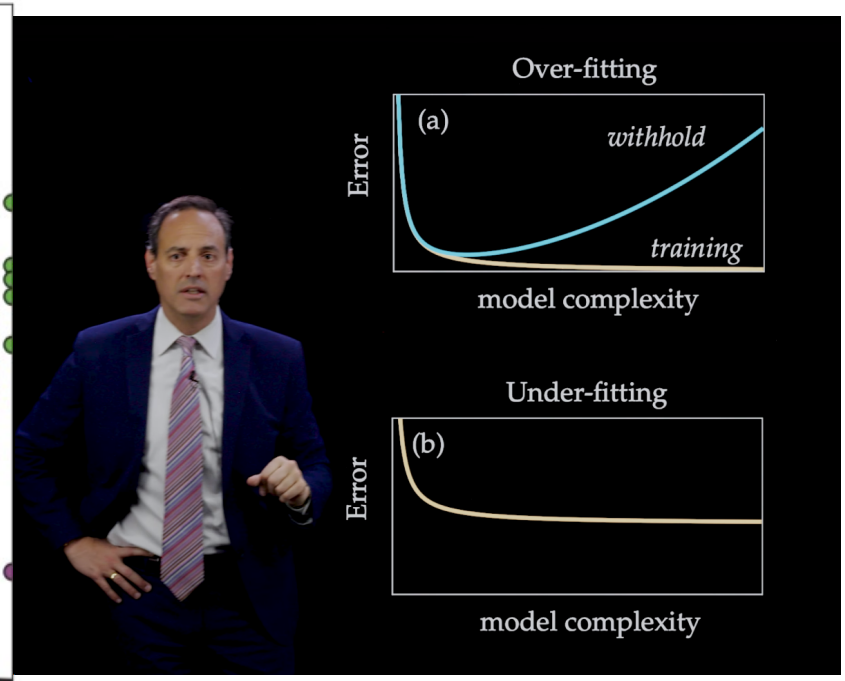
LASSO

higher degree terms are penalized

Pareto optimality

- Occam's razor: when you have two theories making the same predictions, the simpler one is the more likely
- Pareto's 80/20 rule: 80% of sales come from 20% of clients in business
- Model selection is not simply about reducing error, it is about producing a model that has a high degree of interpretability, generalization and predicative capabilities

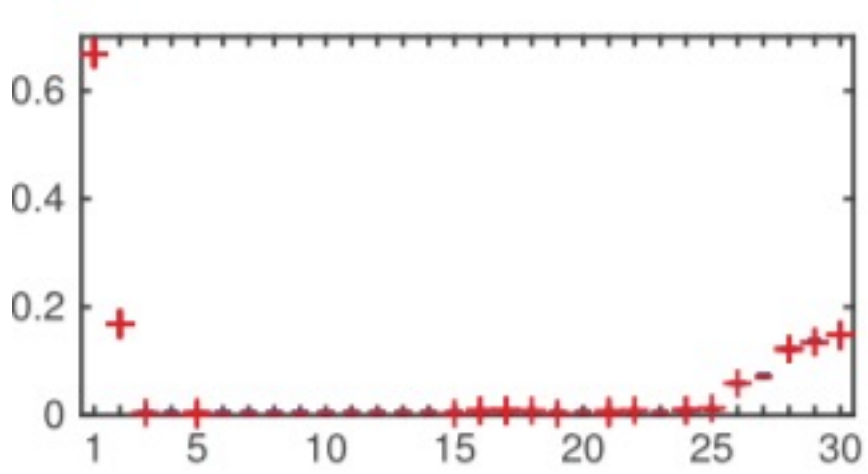




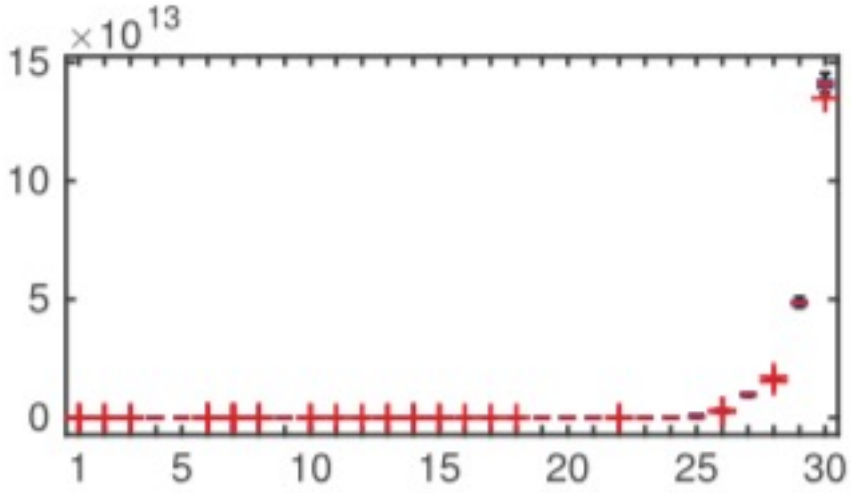
- Many models with the same number of terms
- Pareto Frontier is defined by the models that produce the lowest error for a given number of terms
- The solid line is an approximation of the Pareto Frontier
- Pareto Optimal solutions are models that produce accurate models while remaining simple

Overfitting

- Consider the same example using parabolic data with noise
- 100 training data obtained within the region $[0,4]$
- 100 testing data from the region $[4,8]$
- More terms give overfitted model, resulting in worse predictive power



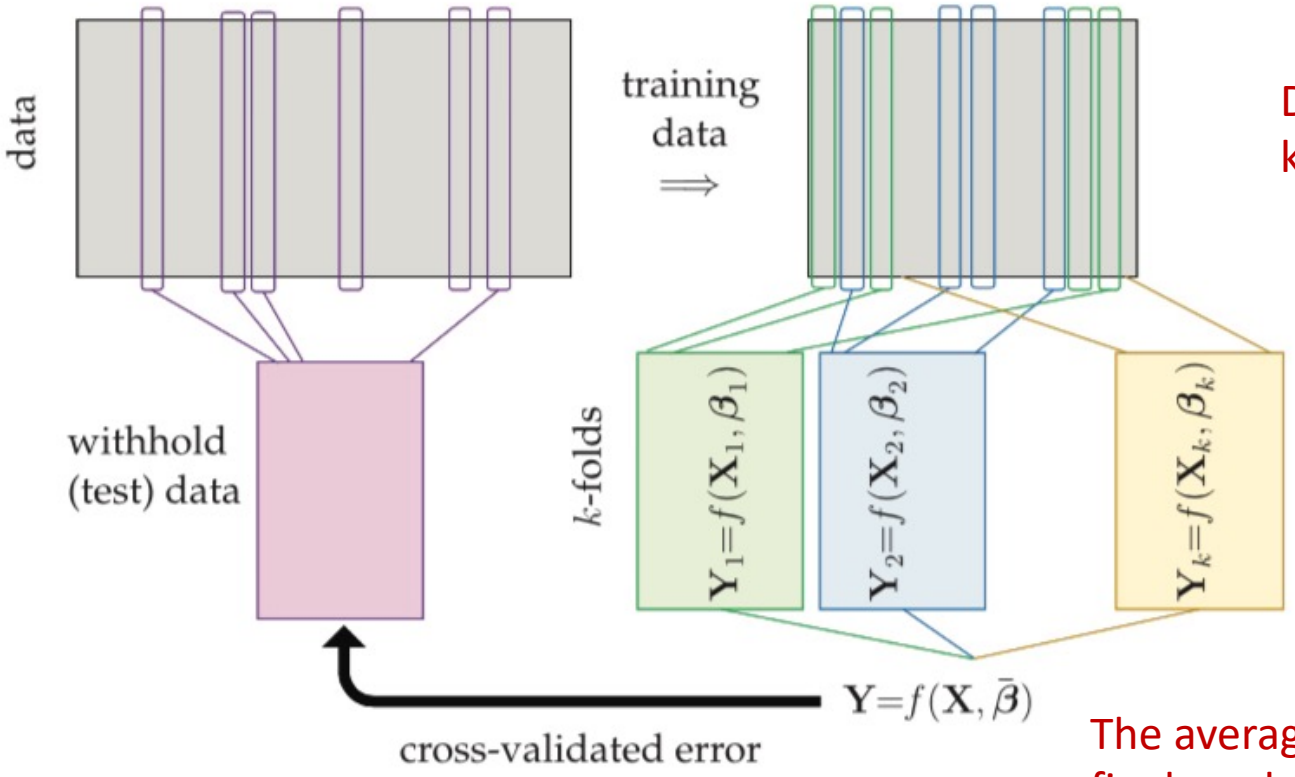
Training error



Testing error

Cross-validation

- To overcome the consequences of overfitting
- Often use the k-fold cross-validation



Part of the data are withheld for testing

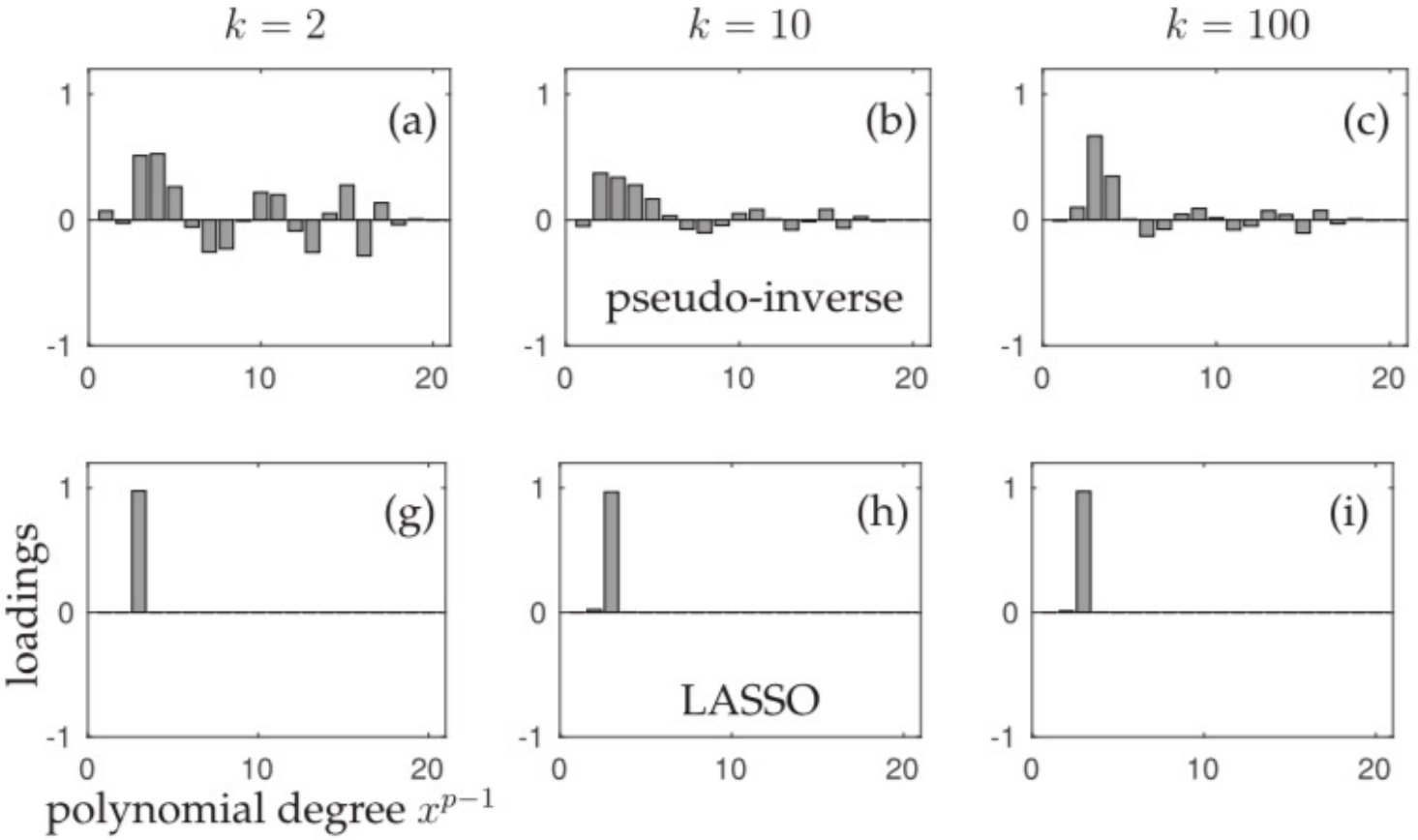
Data is randomly divided into k groups

A model is obtained for each group

The average parameter is used as the final model

Example

- Consider the same example using 100 parabolic data with noise



$$\bar{\boldsymbol{\beta}} = (1/k)\sum_{j=1}^{k} \boldsymbol{\beta}_j$$

Thresholding can be applied to remove small terms