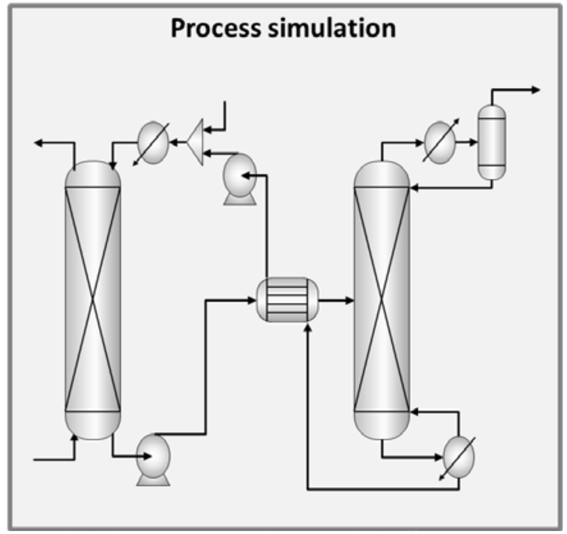
Subset Selection in Multiple Linear Regression

$$x \in \mathbb{R}^D \qquad x^l \leq x \leq x^u \qquad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \\ \vdots \\ x_D \end{pmatrix}$$

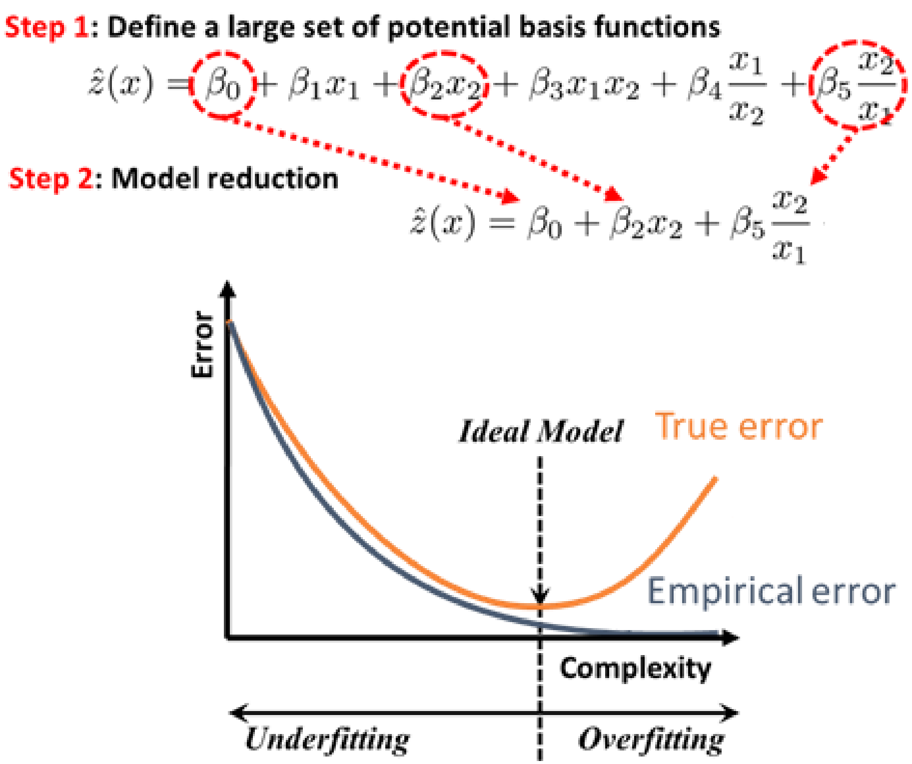


$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \\ \vdots \\ z_K \end{pmatrix} \qquad z \in \mathbb{R}^K \qquad z = f(x)$$

Zachary Wilson
ztw@andrew.cmu.edu

Nick Sahinidis
Sahinidis@cmu.edu

Subset Selection in Multiple Linear Regression



Subset Selection is used to build surrogate models that are

- **Accurate** representations of higher order functions or black-box simulations
- **Simple** in functional form, tailored for algebraic optimization

Fitness Criterion

- **Balances** model complexity with reduction in empirical error
- **Penalize directly** for the number of explanatory variables in the regression model

IP Formulations of Fitness Criterion

$$\begin{aligned} \min \quad & \frac{1}{2} x^T Q x + c^T x \\ \text{s.t.} \quad & -Mz_j \leq \beta_j \leq Mz_j \quad (j = 1, 2..., k) \\ & z_j \in \{0, 1\} \end{aligned}$$

$$\begin{aligned} \min_{K \in \{1,\dots,K^u\}} \quad & [\phi_{\beta,y}(\beta, y)|_K] + \phi_K(K) \\ \text{s.t.} \quad & \\ & \min_{\beta,y} \quad [\phi_{\beta,y}(\beta, y)|_K] \\ & \text{s.t.} \quad -Mz_j \leq \beta_j \leq Mz_j \\ & \qquad \sum_{j \in J} z_j \leq K \\ & \qquad z_j \in \{0, 1\} \end{aligned}$$

MIQP formulations

- Solved **directly** (Cp, BIC)
- Solved in **nested optimization problem** (AIC,MSE)

Alternative Model Selection Techniques

- Regularization – LASSO, Ridge Regression
- Stepwise Heuristics