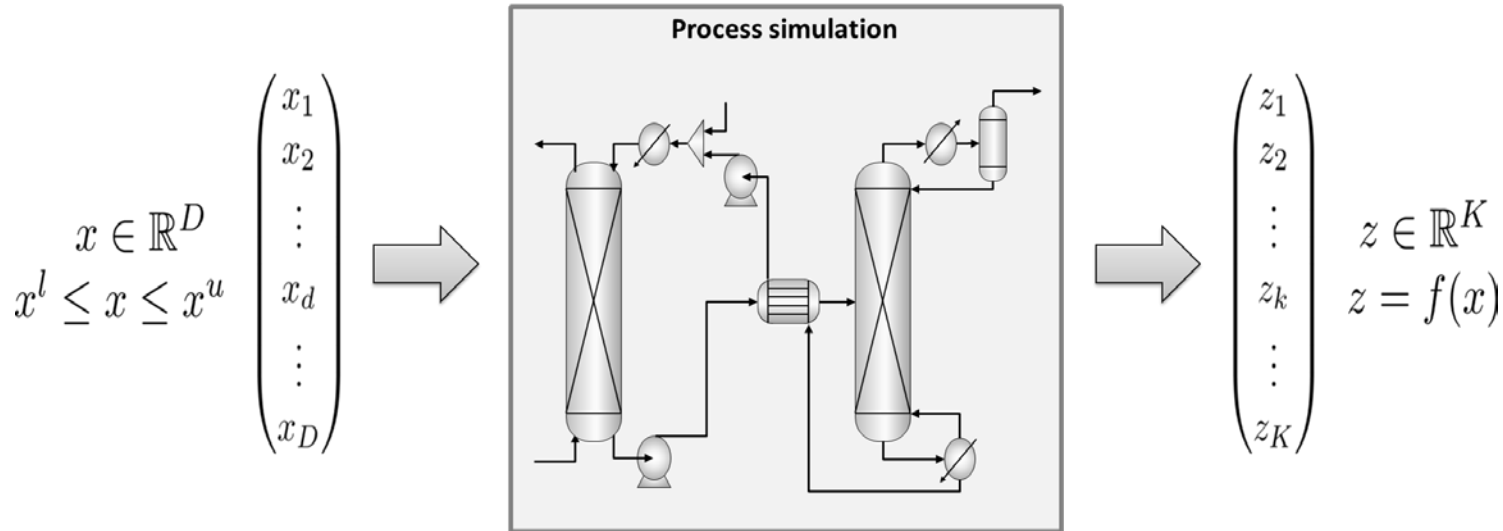


Subset Selection in Multiple Linear Regression



Zachary Wilson

ztw@andrew.cmu.edu

Nick Sahinidis

Sahinidis@cmu.edu



This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

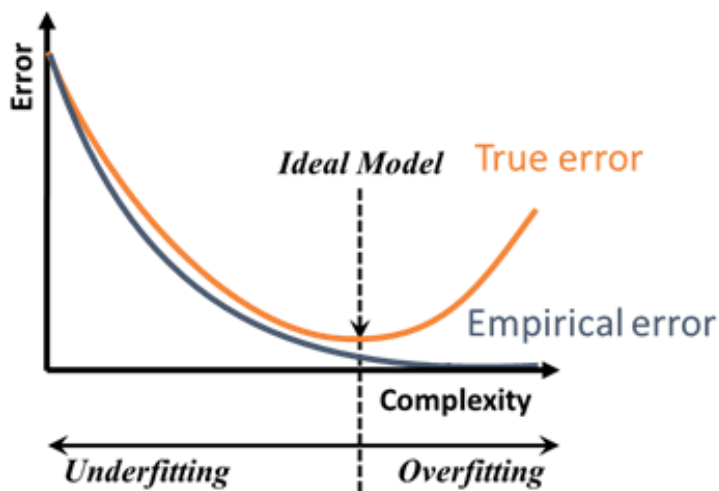
Subset Selection in Multiple Linear Regression

Step 1: Define a large set of potential basis functions

$$\hat{z}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 \frac{x_1}{x_2} + \beta_5 \frac{x_2}{x_1}$$

Step 2: Model reduction

$$\hat{z}(x) = \beta_0 + \beta_2 x_2 + \beta_5 \frac{x_2}{x_1}$$



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

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

$$\begin{aligned} \min_{K \in \{1, \dots, K^u\}} \quad & [\phi_{\beta, y}(\beta, y)|_K] + \phi_K(K) \\ \text{s.t.} \quad & \end{aligned}$$

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

MIQP formulations

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

Alternative Model Selection Techniques

- Regularization – LASSO, Ridge Regression
- Stepwise Heuristics