1 Introduction

This exercise involves using Monte Carlo simulations to analyze the performance of Lasso and Ridge and Principal Component regression models under conditions of sparsity and collinearity. Students will generate datasets with varying degrees of sparsity and collinearity, apply regression techniques, and use Monte Carlo methods to estimate model performance.

2 Objective

Understand the impact of sparsity and collinearity on regression models and demonstrate the ability to apply Lasso, Ridge and PC regression to datasets with these characteristics.

2.1 Dataset Generation

Given:

- n: number of samples, i.e. $n = 150$
- p: number of features, i.e. $p = 150$
- k: number of informative features (with $k \ll p$)

Procedure:

1. Generate an $n \times p$ matrix X where each element $x_{ij} \sim N(0, 1)$.
2. Define a coefficient vector $\beta \in \mathbb{R}^p$ such that:

$$\beta_j = \begin{cases} 0 & \text{if } j = [1, 5, 32], \\ U_{[0,1]} & \text{otherwise.} \end{cases}$$

   HINT: When you generate the β's using uniform random distribution, exclude zeros in the simulation.

3. Generate the response vector y as:

$$y = X\beta + \epsilon,$$

   where $\epsilon \sim N(0, \sigma^2 I)$ is the noise vector.

2.2 Collinear Dataset Generation

To simulate datasets with collinearity among the features, we induce correlation among the predictors.

2.2.1 High Collinearity

Given a correlation factor c close to 1 (e.g., 0.9):

1. Generate an initial feature vector $x_1 \sim N(0, 1)$ of size n.
2. For each subsequent feature x_j (for $j = 2$ to p):

$$x_j = c \cdot x_1 + \sqrt{1 - c^2} \cdot \epsilon_j,$$

where $\epsilon_j \sim N(0, 1)$ is independent noise, ensuring $\text{corr}(x_1, x_j) = c$.

2.2.2 Low Collinearity

For a lower correlation factor c (e.g., 0.1), follow the same procedure as above but with the respective c.

2.3 Monte Carlo Simulation

Each regression model will be run multiple times (e.g., 100 simulations) to estimate the performance metrics such as Mean Squared Error (MSE) and model coefficients stability. The Monte Carlo estimate of the MSE is given by:

$$\text{MSE}_{\text{MC}} = \frac{1}{M} \sum_{m=1}^{M} \text{MSE}_m, \tag{1}$$

where M is the number of simulations, and MSE_m is the MSE of the m-th simulation.

3 Tasks

1. Generate the datasets as specified.
2. Apply Lasso and Ridge and Principal Component regression to each dataset.
3. Use cross-validation to determine the optimal regularization parameter for each model.
4. Perform Monte Carlo simulations to estimate the average MSE and analyze the variability of the regression coefficients.