project4_Annapureddy_Githika

1. Scientific or Statistical Question

I want to identify if there are significant differences between lasso, ridge, and linear regression for data that is not correlated, mildly correlated, highly correlated, and almost perfectly correlated.

---

2. Data

15 x variables are generated and 1 y variable is generated. The x variables are randomly generated from a multivariate normal distribution with equal pairwise correlation values, 0.10 for no correlation, 0.50 for mild correlation, 0.95 for high correlation, or 0.99 for almost perfect correlation.

y = 1 * x2 + 2 *x3 + 3 * x4 + 4 * x5 + noise

noise is a random multivariate number.

---

3. Estimates

I am estimating if y can be predicted from all 15x's by linear, ridge, and lasso regression. It is replicated 100 times for each correlation and each model. Each model's performance is measured by MSE. Models are trained on 70% of each dataset and tested on 30%. The data is split into train_x, train_y, test_x, and test_y data. The models attempt to learn the relationship between train_x and train_y. Then, they attempt to predict test_y values for given test_x data. The difference between the predicted test_y values and the true y values is calculated by MSE.

---

4. Methods

I evaluate Linear Regression, Ridge Regression (alpha = 0), and Lasso Regression (alpha = 1).
All models are fit using the same train/test split per dataset.
I hypothesize that Lasso will perform better in high correlation settings due to its variable selection property.
Ridge is expected to perform better than Linear in highly correlated settings.

---

5. Performance Criteria

I will use Mean Squared Error to estimate the differences. This is calculated as a summary of all repetitions. I will compare the models by model type and by correlation.

---

6. Simulation Plan

- 100 simulations per correlation level (4 total)
- Each simulation generates 100 observations
- 3 models run per simulation
- Train/test split = 70/30
- MSE is recorded for each run
- No major changes from Project III, but data and correlation structure expanded

---

7. Anticipated Challenges or Limitations

- Runtime increases due to nested simulations
- Model assumptions (e.g., linearity, normality) may not hold in practice
- Random variability in small sample sizes
- Extensions: try larger nnn, tune lambda via cross-validation, include variable selection metrics like number of non-zero coefficients in Lasso