CODS COMAD

CGFLasso: Combating Multicollinearity using Domain Knowledge

SOUMYA SARKAR, NITIN SINGHAL, SHWETA JAIN, SHASHI SHEKHAR JHA

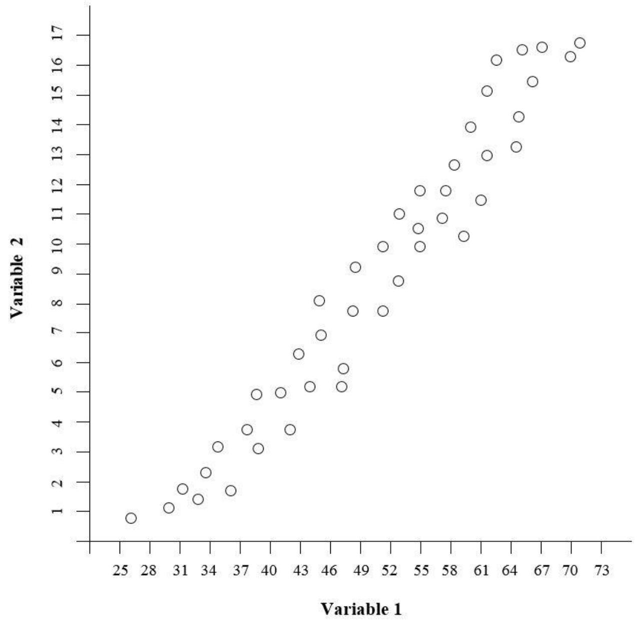
INDIAN INSTITUTE OF TECHNOLOGY ROPAR

(SHORT PAPER AND POSTER)

What is Multicollinearity?

- When all predictive variables are no longer independent in a linear regression model, multicollinearity problems exist.
- It can lead to inaccurate parameter estimates, low confidence on regressor coefficients and incorrect feature selection in domains such as biological sciences or finance.
- Can be detected using Variance Inflation Factor (VIF), but the effects are in terms of high standard deviation of the linear regressor coefficients

Multicollinearity



Note. The plot depicts the predictor variables, Variable 1 and Variable 2, as highly linearly related.

GFLasso and Associated Issues

- One way to tackle multicollinearity is using regularization techniques.
- Regularization techniques doesn’t drop variables and has strong theoretical backing.
- Ordinary regularization techniques don’t consider relations among features.
- Graph-guided Fused Lasso (GFLasso) uses an unweighted feature graph ***B*** but real life is not so binary, and also it depends on the correlation matrix solely.

$$penalty(w) = (1 - \alpha)\,\|w\|_1 + \alpha \sum_{ij} B_{ij}(w_i - sign(r_{ij})w_j)^2$$

Contextual GFLasso (CGFLasso)

- Prior domain knowledge can help offset a bad dataset.
- We propose the Contextual Graph-guided Fused Lasso (CGFLasso) which uses prior domain knowledge in the form of a prior matrix ***A*** and the real life data in the form of the correlation matrix **C** to determine the penalty.

$$penalty(w) = (1-\alpha)\,\|w\|_1 + \alpha \sum_{ij} B_{ij}(w_i - sign(r_{ij})w_j)^2$$

$$B = \begin{bmatrix} \psi^1_{11}A_{11} & \psi^1_{12}A_{12} & \psi^1_{13}A_{13} & \cdots \\ \psi^1_{21}A_{21} & \psi^1_{22}A_{22} & \cdots & \\ \cdots & & & \\ \cdots & & & \psi^1_{nn}A_{nn} \end{bmatrix} + \begin{bmatrix} \psi^2_{11}C_{11} & \psi^2_{12}C_{12} & \psi^2_{13}C_{13} & \cdots \\ \psi^2_{21}C_{21} & \psi^2_{22}C_{22} & \cdots & \\ \cdots & & & \\ \cdots & & & \psi^2_{nn}C_{nn} \end{bmatrix}$$

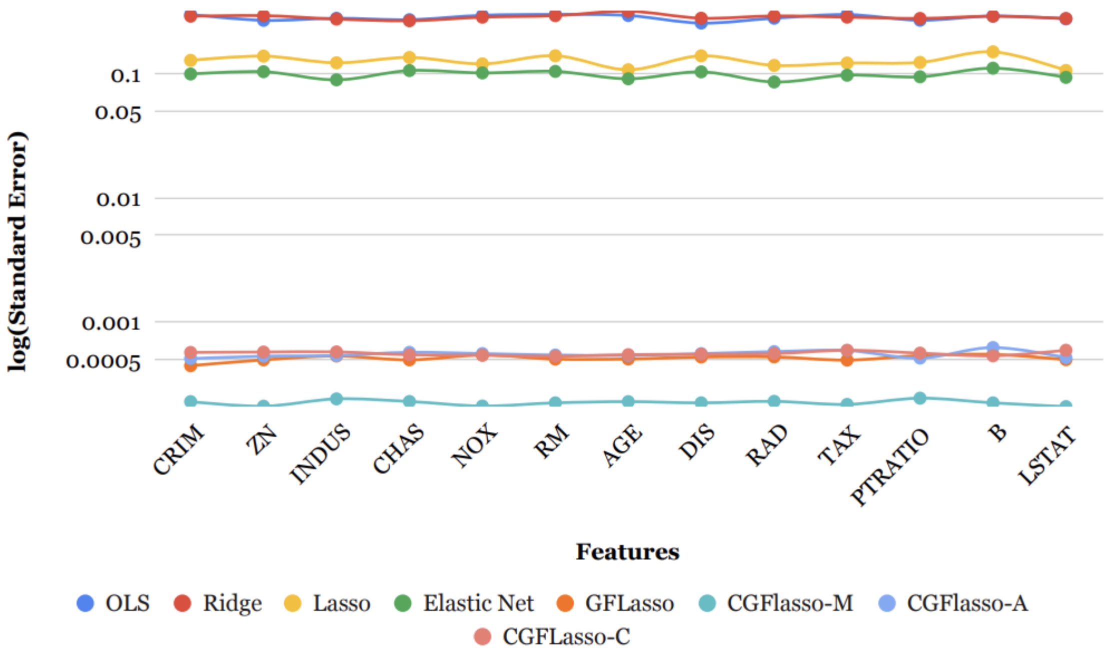
$$where\ \psi^1_{ij} = \frac{|A_{ij}|}{1 + |A_{ij} - C_{ij}|}\ and\ \psi^2_{ij} = 1 - \psi^1_{ij}$$

Choosing a Prior Matrix

- The method we provide is quite sensitive to the choice of prior matrix – however good data can possibly negate a bad prior too.
- We use three different ways to generate the prior matrix for three datasets:
  - CGFLasso with a prior created manually [**CGFLasso-M**]
  - CGFLasso with a prior generated correlation matrix by using a sample of data [**CGFLasso-C**]
  - CGFLasso with a prior created from the adjacency matrix used by GFLasso [**CGFLasso-A**].

The Boston House Pricing Dataset

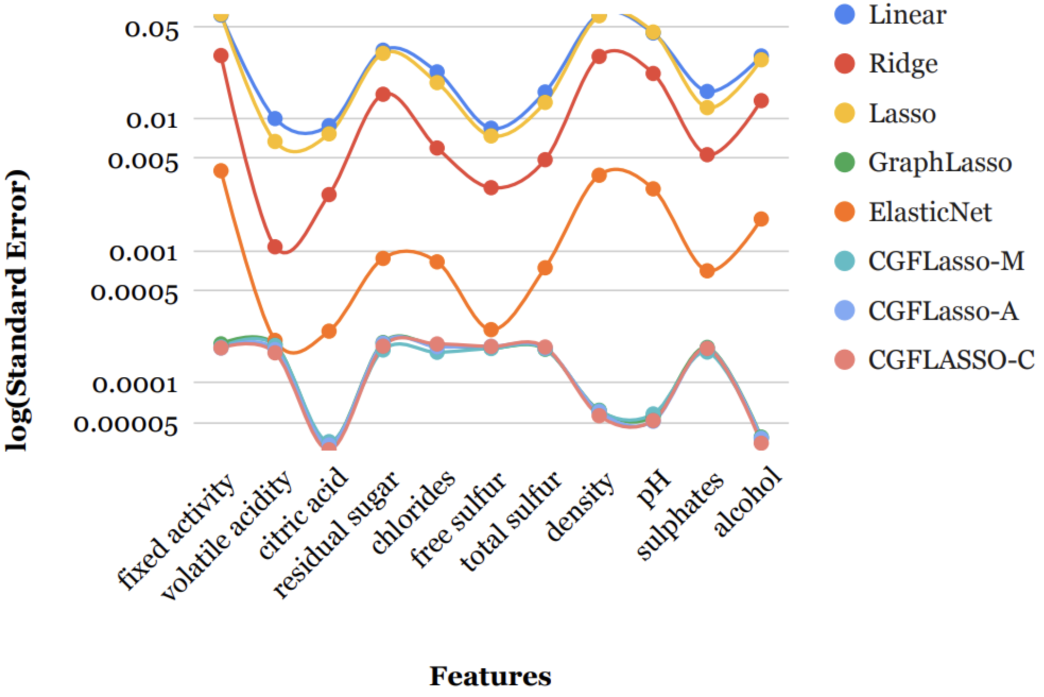
- The Boston Housing Dataset [2] consists of 13 continuous attributes (including the target variable MEDV) and 1 binary-valued attribute.
- We want to predict the Median value of owner-occupied homes in $1000’s given the other parameters.
- There are 506 data points.
- We found the range of VIF values to be within 2.1 and 84.97, with several greater than 10, and hence there is multicollinearity involved here.



(a) Standard Error Comparison of All Methods in BHP

The Red Wine Dataset

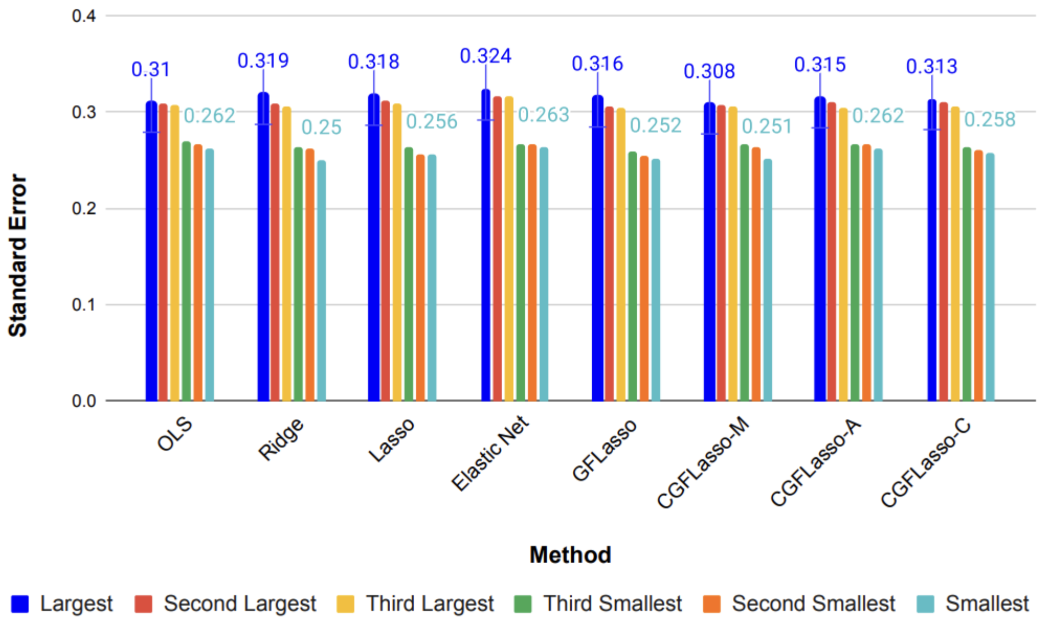
- The Red Wine Dataset [3] consists of 13 continuous attributes (including the target variable quality).
- We want to predict the quality score of the wine, given the other parameters.
- There are 1599 data points.
- We found the range of VIF values to be within 2.1 and 84.97, with several greater than 10, and hence there is multicollinearity involved here.



(b) Standard Error Comparison of All Methods in RW

The Stock Price Dataset

- The Stock Price Dataset [4] consists of 63 continuous attributes (including the target variable Close_forcast) and 1 nominal attribute.
- We want to predict the closing price of AppleTM stock, given the other parameters.
- As part of data pre-processing, we had to leave out the columns Date, CCI and Is_year_start due to them giving NaN entries in the correlation matrix calculations, leaving us with 60 columns and 1 target column.
- There are 3732 data points.
- We found the range of VIF values to be within 1.02 and 3.78 × 105 , with several greater than 10, and hence there is multicollinearity involved here.



(c) Standard Error Comparison of All Methods in SP for Largest and Smallest 3 values

Accuracy Comparison

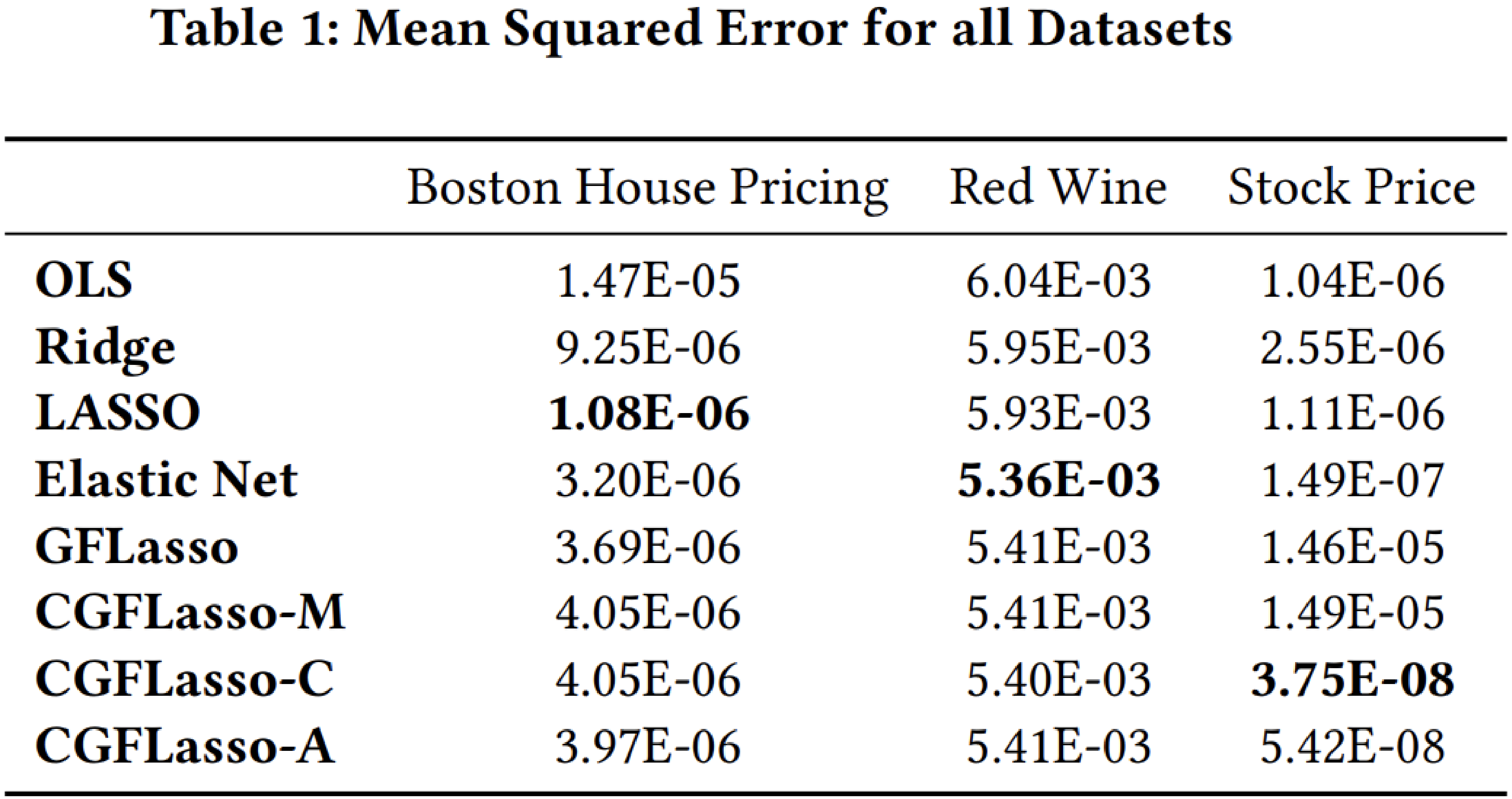


Table 1: Mean Squared Error for all Datasets

	Boston House Pricing	Red Wine	Stock Price
OLS	1.47E-05	6.04E-03	1.04E-06
Ridge	9.25E-06	5.95E-03	2.55E-06
LASSO	**1.08E-06**	5.93E-03	1.11E-06
Elastic Net	3.20E-06	**5.36E-03**	1.49E-07
GFLasso	3.69E-06	5.41E-03	1.46E-05
CGFLasso-M	4.05E-06	5.41E-03	1.49E-05
CGFLasso-C	4.05E-06	5.40E-03	**3.75E-08**
CGFLasso-A	3.97E-06	5.41E-03	5.42E-08

References

- [1] Seyoung Kim and Eric Xing. 2009. Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. PLoS genetics 5 (09 2009), e1000587. https://doi.org/10.1371/journal.pgen.1000587
- [2] David Harrison and Daniel Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management 5 (03 1978), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2
- [3] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems 47, 4 (2009), 547–553. https://doi.org/10.1016/j.dss.2009. 05.016 Smart Business Networks: Concepts and Empirical Evidence.
- [4] Nikhil Kohli. 2020. US Stock Market Data Technical Indicators (Version 1). https://www.kaggle.com/datasets/nikhilkohli/us-stock-market-data-60- extracted-features

GAME THEORY
AND MACHINE
LEARNING LAB

Visit us at
https://gtmliitrpr.github.io/

THANK YOU!

Feel free to reach out at

soumasarkar.19@gmail.com

Acknowledgments:

SERB
INDIA

भारतीय प्रौद्योगिकी संस्थान रोपड़
INDIAN INSTITUTE OF TECHNOLOGY ROPAR
धियो यो नः प्रचोदयात्