CGFLasso: Combating Multicollinearity using Domain Knowledge

Soumya Sarkar, Nitin Singhal, Shweta Jain, Shashi Shekhar Jha

Department of Computer Science & Engineering, IIT Ropar

About Multicollinearity

- Multicollinearity is a statistical phenomenon wherein multiple predictor variables are severely correlated in a multiple regression model.
- May lead to inaccurate parameter estimates [1]
- Leads to low confidence on regressor coefficients [1]
- May cause incorrect feature selection [1]
- Affects several domains such as biological sciences or finance.

Checking for Multicollinearity

- Standard way to detect multicollinearity in a dataset is to use the Variance Inflation Factor:

$$VIF = \frac{1}{1 - R^2}$$

- Effect of Multicollinearity - High standard deviation of the coefficients

Regularisation and Feature Relationship Structures

- Regularization is a computationally efficient method to incorporate constraints in any loss function without the loss of features.
- Regularisation can help deal with multicollinearity
- Ordinary l_1 and l_2 norms do not consider the relationship between the features.

GFLasso and Associated Issues

- Most general approach to model relations as a dependency graph with feature nodes, and edges representing relationships.
- GFLasso regularization [2] defines such a setting:
  - $G = (V, E)$ is a given graph
  - $V = \{1, ..., n\}$ is the set of features as nodes
  - E is the set of edges denoting their relation
  - $B \in R_{n*n}$ is the adjacency matrix of G ($B_{i,j} \in \{0, 1\}$)
- **Penalty Term** $(penalty(w))$:

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

where r_{ij} is the correlation between two features.

- **Issue**: The features have hard relations amongst each other - either a relation exists (1), or not (0).
- A partial relationship could be more appropriate in scenarios where such clear-cut relations are not defined.

Contextual GFLasso

- Similar assumptions to GFLasso and the same penalty term
- We define the matrix B as:

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

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

where A is the prior domain knowledge matrix ($A_{ij} \in [-1, 1]$), and C is the correlation matrix ($C_{ij} \in [-1, 1]$).

Contextual GFLasso (cont.)

- ψ is a critical parameter that determines value of prior knowledge over current data.
- Based on quality of dataset we can reduce the effect of it by prioritizing the domain knowledge.

Choosing a Prior Matrix

- We show the results of three methods -
  - 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**]
- For the Boston Housing Price Dataset [3], we generate CGFLasso-M based on our ideas of how the variables are correlated to each other due to its simplicity.
- For the Red Wine Dataset [4], we generate CGFLasso-M using Google Bard.
- For the Stock Price Dataset [5] we use a manually perturbed version of the correlation matrix due to our lack of familiarity with its many features.

Experimental Results

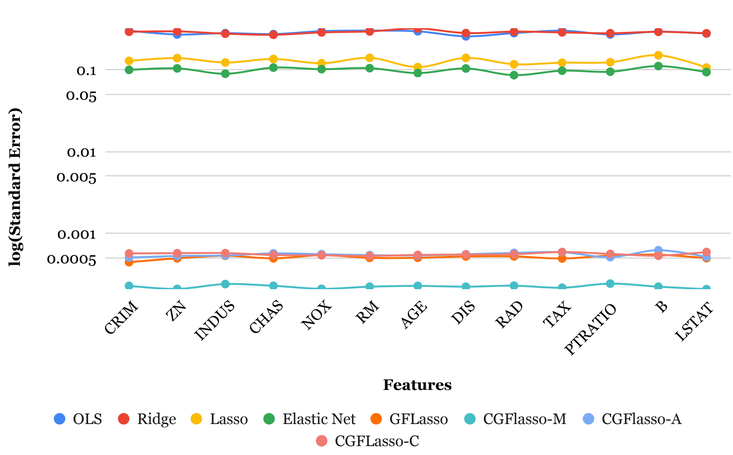

Figure 1. Standard Error Comparison of All Methods in BHP

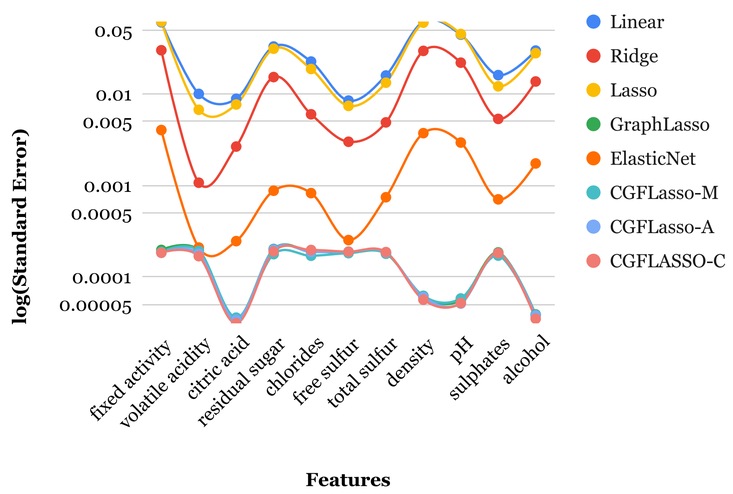

Figure 2. Standard Error Comparison of All Methods in RW

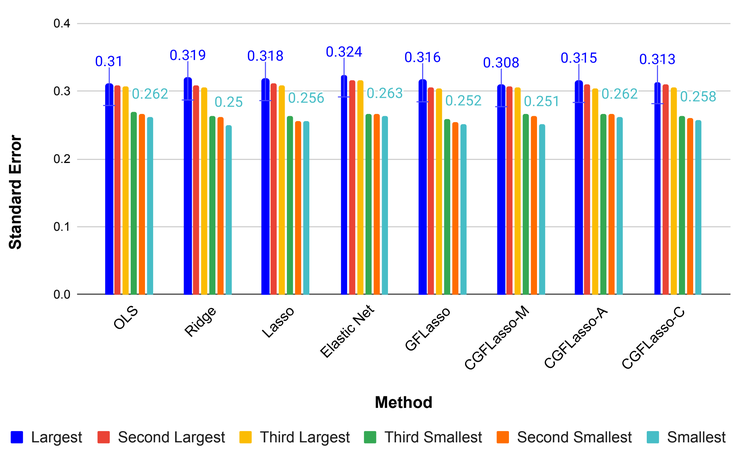

Figure 3. Standard Error Comparison of All Methods in SP for Largest and Smallest 3 values

Experimental Results (contd.)

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

Table 2. Range of Standard Error (SE) of standard error value σ_k for feature k in all Datasets after perturbing on CFLasso-C Prior

Dataset	Minimum SE of σ_k	Maximum SE of σ_k
BHP	2.03E-02	2.64E-02
RW	6.31E-02	9.25E-02
SP	6.41E-03	1.78E-02

- We manage to reduce standard error of coefficients by several factors as compared to standard regularisation techniques and improve upon GFLasso as shown in Figures 1, 2 and 3.
- We show that accuracy is comparable or better than other methods in Table 1.
- We show that having a good domain knowledge is necessary in Table 2, since it can be sensitive.

Conclusions and Future Work

- We show our method reduces standard error of coefficients significantly with correct domain knowledge.
- Our work enables one to incorporate domain knowledge as a regularisation technique to provide better domain generalisability for any regressor method.
- By incorporating notions of causality during the graph preparation we could create even better models in the future.

Acknowledgements

This work has been further supported by SERB, India (CRG/2022/007927).

References

[1] Emine Ozgur Bayman and Franklin Dexter. 2021. Multicollinearity in logistic regression models. Anesth. Analg. 133, 2 (Aug. 2021), 362–365.

[2] 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

[3] 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

[4] 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.

[5] Nikhil Kohli. 2020. US Stock Market Data Technical Indicators (Version 1). https://www.kaggle.com/datasets/nikhilkohli/us-stock-market-data-60-extracted-features

2022csm1013@iitrpr.ac.in, 2022csm1011@iitrpr.ac.in, shwetajain@iitrpr.ac.in, shashi@iitrpr.ac.in