Asignment 4

Ujjwal Chowdhury

Assignmemt on regularised regression

Select a dataset from the UCI dataset on regression (exclude categorical ovaribales for now). First inspect the predictors for multicollinearity and select a subset which is linearly independent. Next find solution of linear regression model using normal equations and sequential gradient descent (Widrow Hoff). Next fit a LASSO and Ridge model using scipy.optimize library. Plot the solution path for different values of α using "lassopath" and "ridgepath" libraries of scipy. Compare the three models in terms of the solution obtained, prediction accuracy etc. Interpret the Lasso and Ridge solutions in terms of selection and shrinkage. Upload two files - one code and one pdf with the brief report.

Report

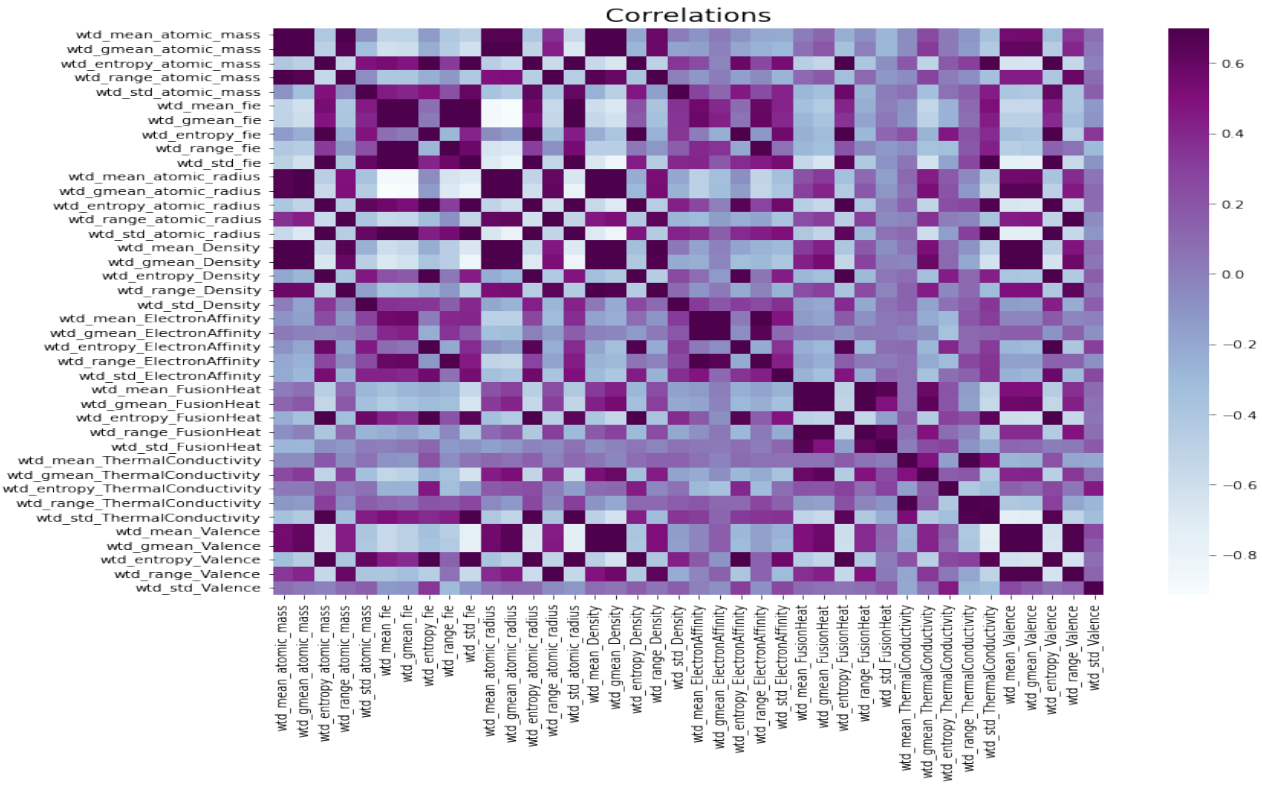
Data Set :- I am using the Superconductivity Data Set from UCI.

Summary of the data :- I get the summary of the data using the .info function.

Filtering the data :- In this step I remove all other columns except the weighted columns.

Counting NA values:- In this step I check whether any NA values are inside any columns. In this case there is no NA value.

Multicollinearity :- Checking the correlation between each columns in the data frame using Heat map correlation matrix.



Selecting Features:- Here I have created a function which remove all highly correlated columns (threshold value = 0.6).

```
correlation_select(df_c,0.6)
```

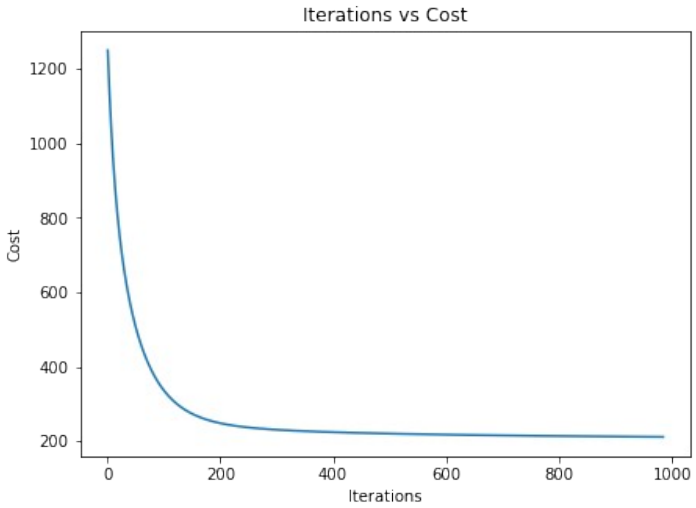
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21263 entries, 0 to 21262
Data columns (total 14 columns):
 #   Column                              Non-Null Count  Dtype  
---  ------                              --------------  -----  
 0   wtd_mean_atomic_mass                21263 non-null  float64
 1   wtd_entropy_atomic_mass             21263 non-null  float64
 2   wtd_std_atomic_mass                 21263 non-null  float64
 3   wtd_mean_fie                        21263 non-null  float64
 4   wtd_range_atomic_radius             21263 non-null  float64
 5   wtd_mean_ElectronAffinity           21263 non-null  float64
 6   wtd_entropy_ElectronAffinity        21263 non-null  float64
 7   wtd_std_ElectronAffinity            21263 non-null  float64
 8   wtd_mean_FusionHeat                 21263 non-null  float64
 9   wtd_mean_ThermalConductivity        21263 non-null  float64
 10  wtd_gmean_ThermalConductivity       21263 non-null  float64
 11  wtd_entropy_ThermalConductivity     21263 non-null  float64
 12  wtd_mean_Valence                    21263 non-null  float64
 13  wtd_std_Valence                     21263 non-null  float64
```

OLS:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          critical_temp   R-squared (uncentered):                   0.826
Model:                            OLS   Adj. R-squared (uncentered):              0.826
Method:                 Least Squares   F-statistic:                              7220.
Date:                Mon, 04 Apr 2022   Prob (F-statistic):                        0.00
Time:                        14:04:07   Log-Likelihood:                         -94122.
No. Observations:               21263   AIC:                                  1.883e+05
Df Residuals:                   21249   BIC:                                  1.884e+05
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
wtd_mean_atomic_mass            0.0154      0.006      2.613      0.009       0.004       0.027
wtd_entropy_atomic_mass        16.4392      0.796     20.650      0.000      14.879      18.000
wtd_std_atomic_mass             0.1584      0.010     16.666      0.000       0.140       0.177
wtd_mean_fie                    0.0184      0.001     18.083      0.000       0.016       0.020
wtd_range_atomic_radius         0.0068      0.005      1.350      0.177      -0.003       0.017
wtd_mean_ElectronAffinity      -0.1462      0.006    -23.018      0.000      -0.159      -0.134
wtd_entropy_ElectronAffinity  -23.8004      0.741    -32.127      0.000     -25.252     -22.348
wtd_std_ElectronAffinity        0.0805      0.010      8.259      0.000       0.061       0.100
wtd_mean_FusionHeat             0.0198      0.015      1.342      0.180      -0.009       0.049
wtd_mean_ThermalConductivity    0.4737      0.006     79.897      0.000       0.462       0.485
wtd_gmean_ThermalConductivity  -0.4951      0.008    -61.739      0.000      -0.511      -0.479
wtd_entropy_ThermalConductivity 17.7504     0.712     24.938      0.000      16.355      19.146
wtd_mean_Valence               -1.9158      0.215     -8.920      0.000      -2.337      -1.495
wtd_std_Valence               -11.4171      0.406    -28.126      0.000     -12.213     -10.621
==============================================================================
Omnibus:                      351.542   Durbin-Watson:                   0.828
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              699.383
Skew:                          -0.035   Prob(JB):                    1.35e-152
Kurtosis:                       3.886   Cond. No.                     6.11e+03
```

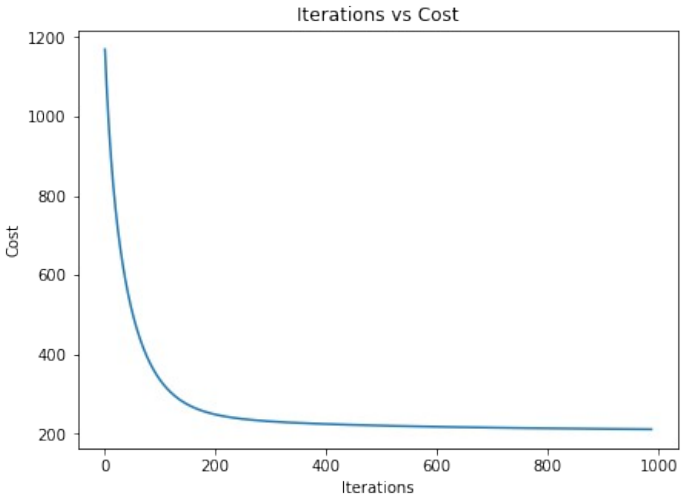
Batch Gradient Descent :-

Total number of iterations = 984
Train Score: 0.41684103478349477
Test Score: 0.4286460967327166



Sequential Gradient Descent :-

Total number of iterations = 987
Train Score: 0.6430747438781501
Test Score: 0.6432614848770073



SGD coefficients

Columns	Coefficient Estimate
wtd_mean_atomic_mass	1.273164e+11
wtd_entropy_atomic_mass	3.825199e+10
wtd_std_atomic_mass	-1.119290e+11
wtd_mean_fie	-4.729064e+10
wtd_range_atomic_radius	4.023643e+09
wtd_mean_ElectronAffinity	1.938587e+11
wtd_entropy_ElectronAffinity	-3.526704e+10
wtd_std_ElectronAffinity	8.467363e+10
wtd_mean_FusionHeat	5.015692e+10
wtd_mean_ThermalConductivity	-8.937411e+10
wtd_gmean_ThermalConductivity	1.037190e+11
wtd_entropy_ThermalConductivity	1.681675e+11
wtd_std_Valence	-3.921789e+10

LASSO :-

MSE training set 448.64022669539935
MSE test set 466.01888024885574

Accuracy of training set 0.6170671926145267
Accuracy of test set 0.6079409994390365

LASSO coefficients

Columns	Coefficient Estimate
wtd_mean_atomic_mass	-0.000000
wtd_entropy_atomic_mass	0.000000
wtd_std_atomic_mass	0.000000
wtd_mean_fie	0.073033
wtd_range_atomic_radius	-0.000000
wtd_mean_ElectronAffinity	-0.000000
wtd_entropy_ElectronAffinity	0.000000
wtd_std_ElectronAffinity	0.000000
wtd_mean_FusionHeat	-0.000000
wtd_mean_ThermalConductivity	0.068023
wtd_gmean_ThermalConductivity	-0.000000
wtd_entropy_ThermalConductivity	-0.000000
wtd_mean_Valence	-0.000000
wtd_std_Valence	-0.000000

Ridge Regression :-

MSE of training set = 419.2101073342187
Accuracy of training set 0.6507528338821016
Accuracy of test set 0.6473209505616128

Coefficients estimates

MSE = 412.8128055949876

Columns	Coefficient Estimate
wtd_mean_atomic_mass	0.024337
wtd_entropy_atomic_mass	17.026830
wtd_std_atomic_mass	0.159140
wtd_mean_fie	0.023474
wtd_range_atomic_radius	0.013268
wtd_mean_ElectronAffinity	-0.151132
wtd_entropy_ElectronAffinity	-23.373274
wtd_std_ElectronAffinity	0.077065
wtd_mean_FusionHeat	0.022391
wtd_mean_ThermalConductivity	0.480450
wtd_gmean_ThermalConductivity	-0.493430
wtd_entropy_ThermalConductivity	18.731995
wtd_mean_Valence	-1.589036
wtd_std_Valence	-11.633624

Elastic Net:-

Accuracy of training set 0.6063905407734923
Accuracy of test set 0.5975760671689727

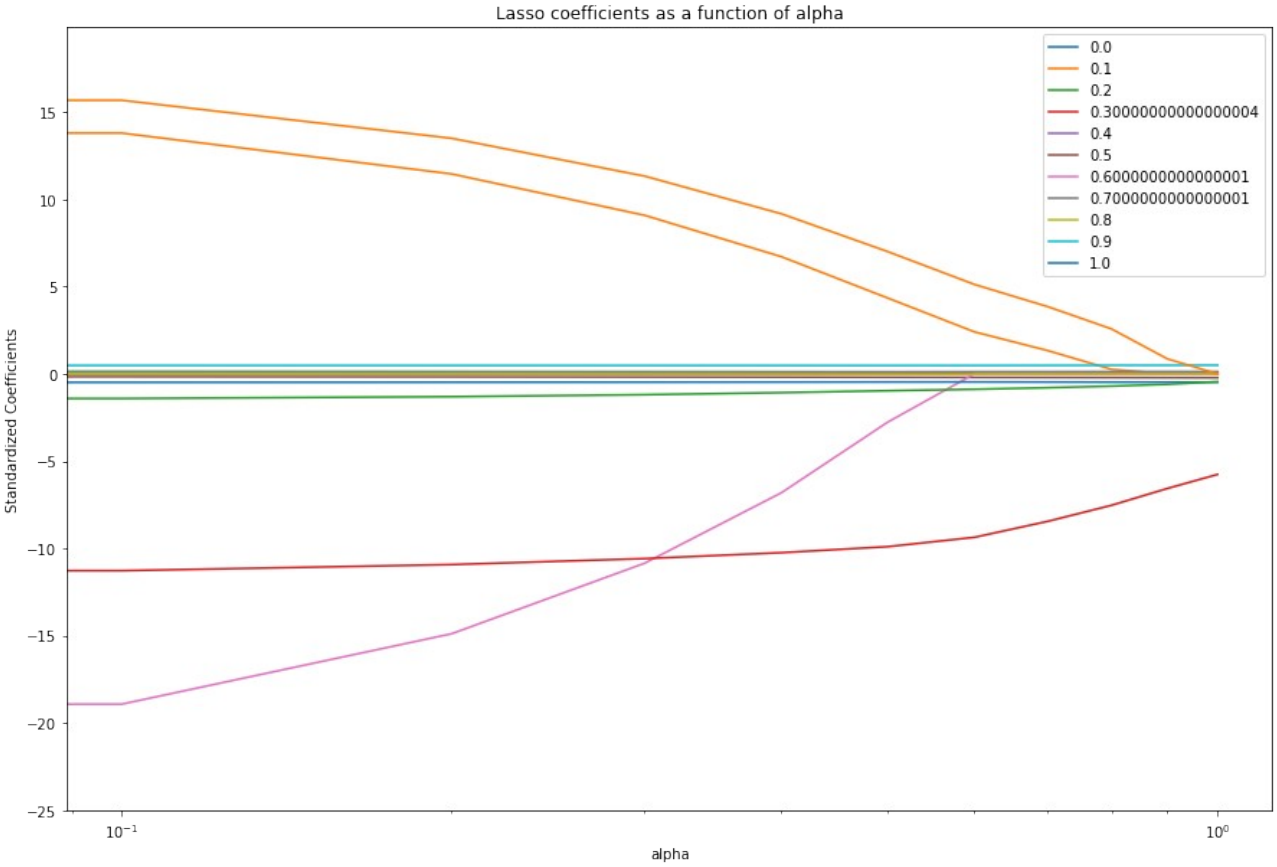
Coefficients estimates

Columns	
wtd_mean_atomic_mass	-0.002080
wtd_entropy_atomic_mass	1.139267
wtd_std_atomic_mass	0.091971
wtd_mean_fie	0.041641
wtd_range_atomic_radius	-0.026168
wtd_mean_ElectronAffinity	-0.210998
wtd_entropy_ElectronAffinity	-0.317390
wtd_std_ElectronAffinity	0.080545
wtd_mean_FusionHeat	-0.006153
wtd_mean_ThermalConductivity	0.503547
wtd_gmean_ThermalConductivity	-0.491537
wtd_entropy_ThermalConductivity	1.123867
wtd_mean_Valence	-1.106421
wtd_std_Valence	-2.506409

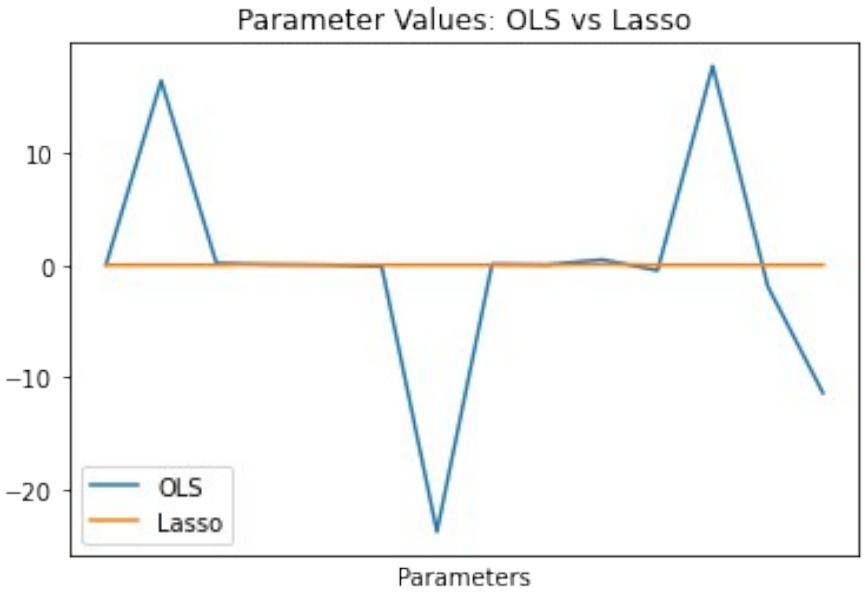
Observations :-

In LASSO regularization some coefficient values are becoming zero.

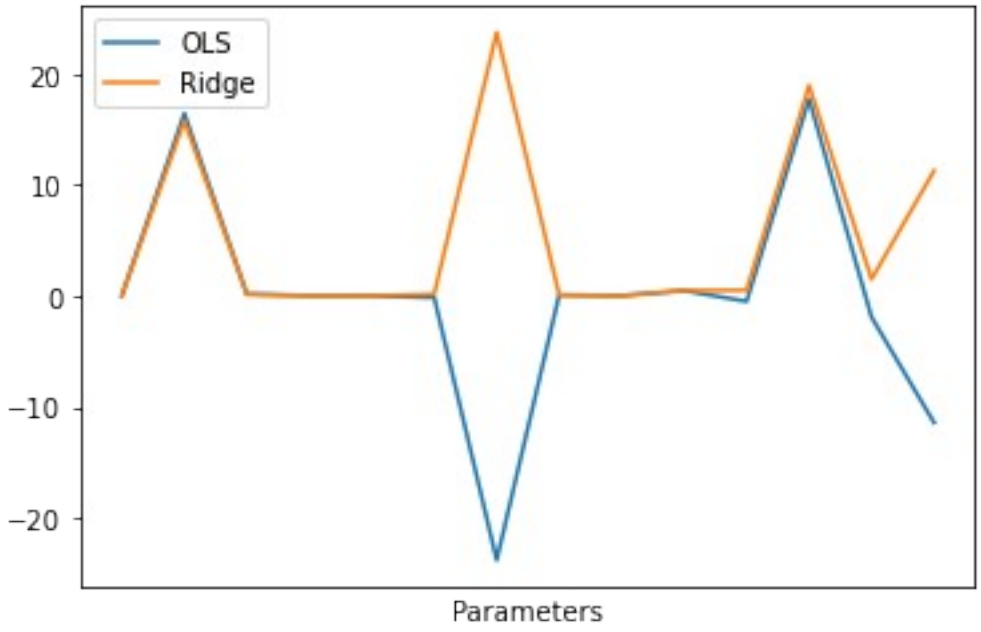
Plotting LASSO coefficients as a function of alpha



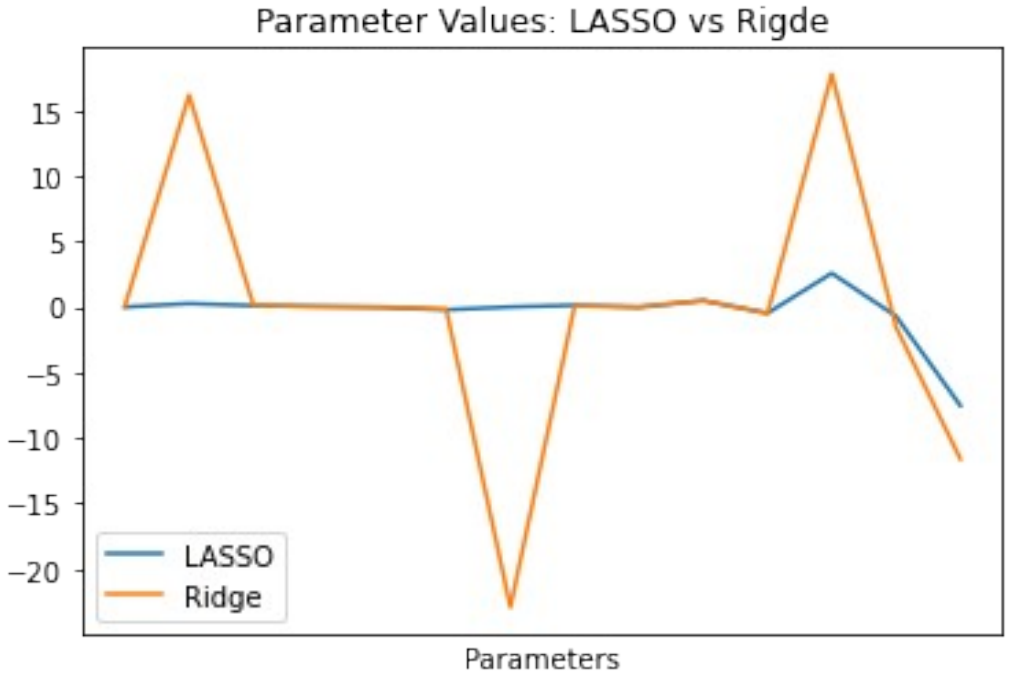
Comparing OLS & LASSO Coefficients:



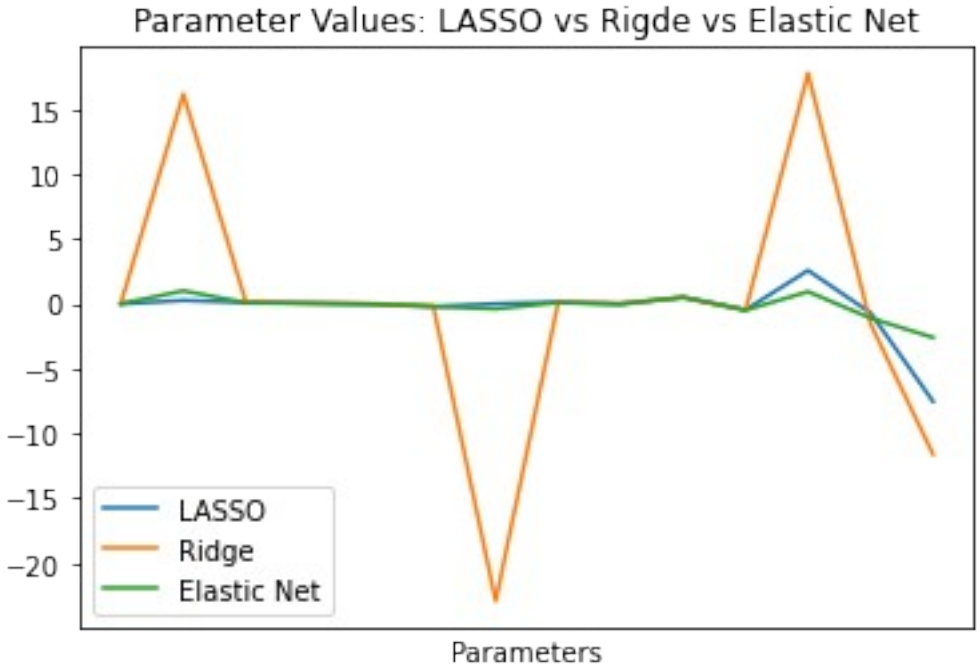
Comparing OLS & Ridge Coefficients:



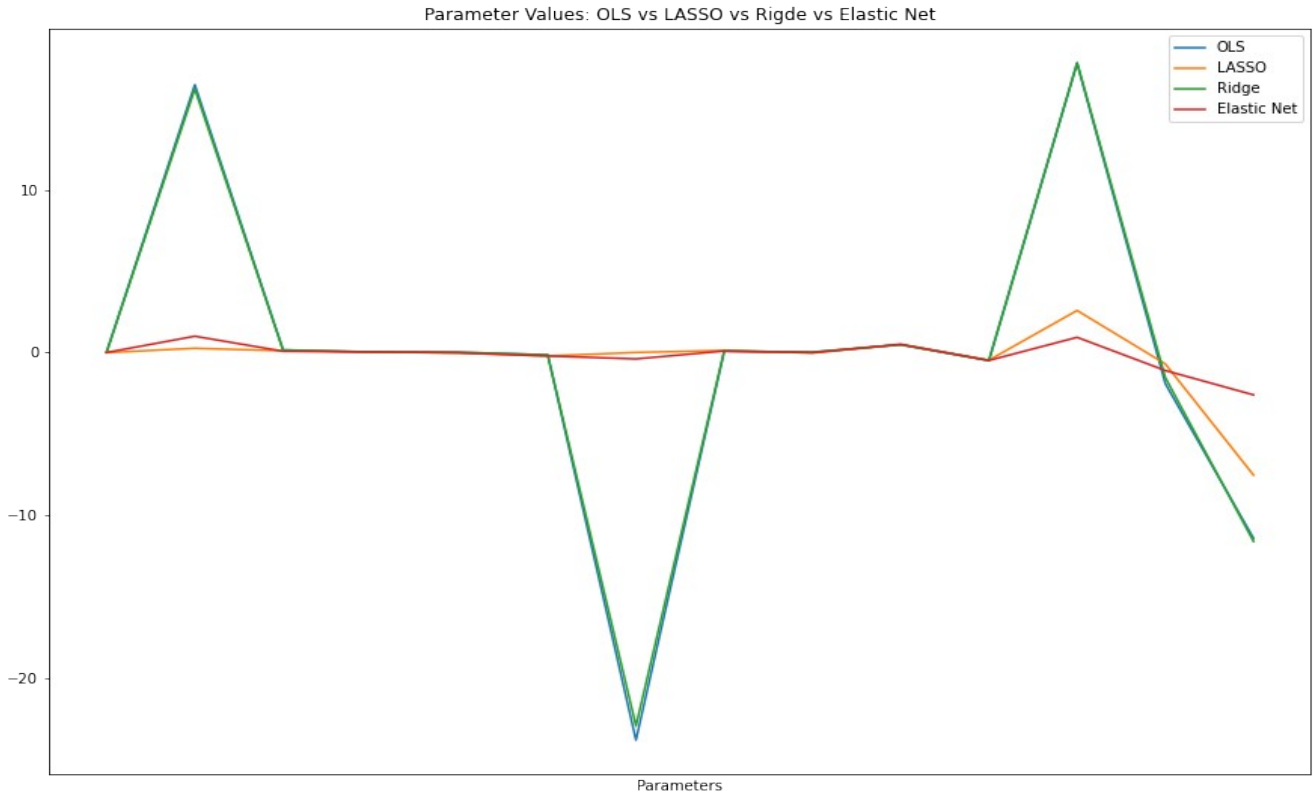
Comparing LASSO & Ridge Coefficients:



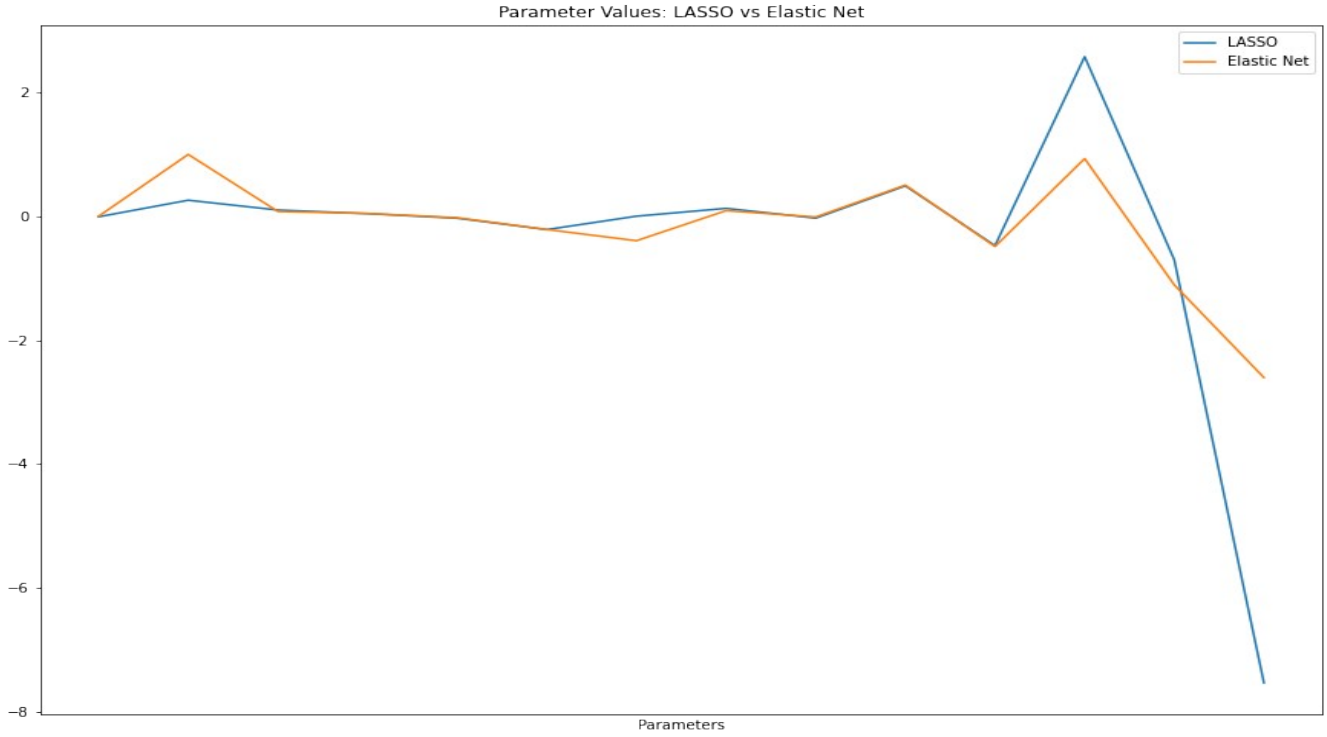
Comparing LASSO , Ridge and Elastic Net Coefficients:



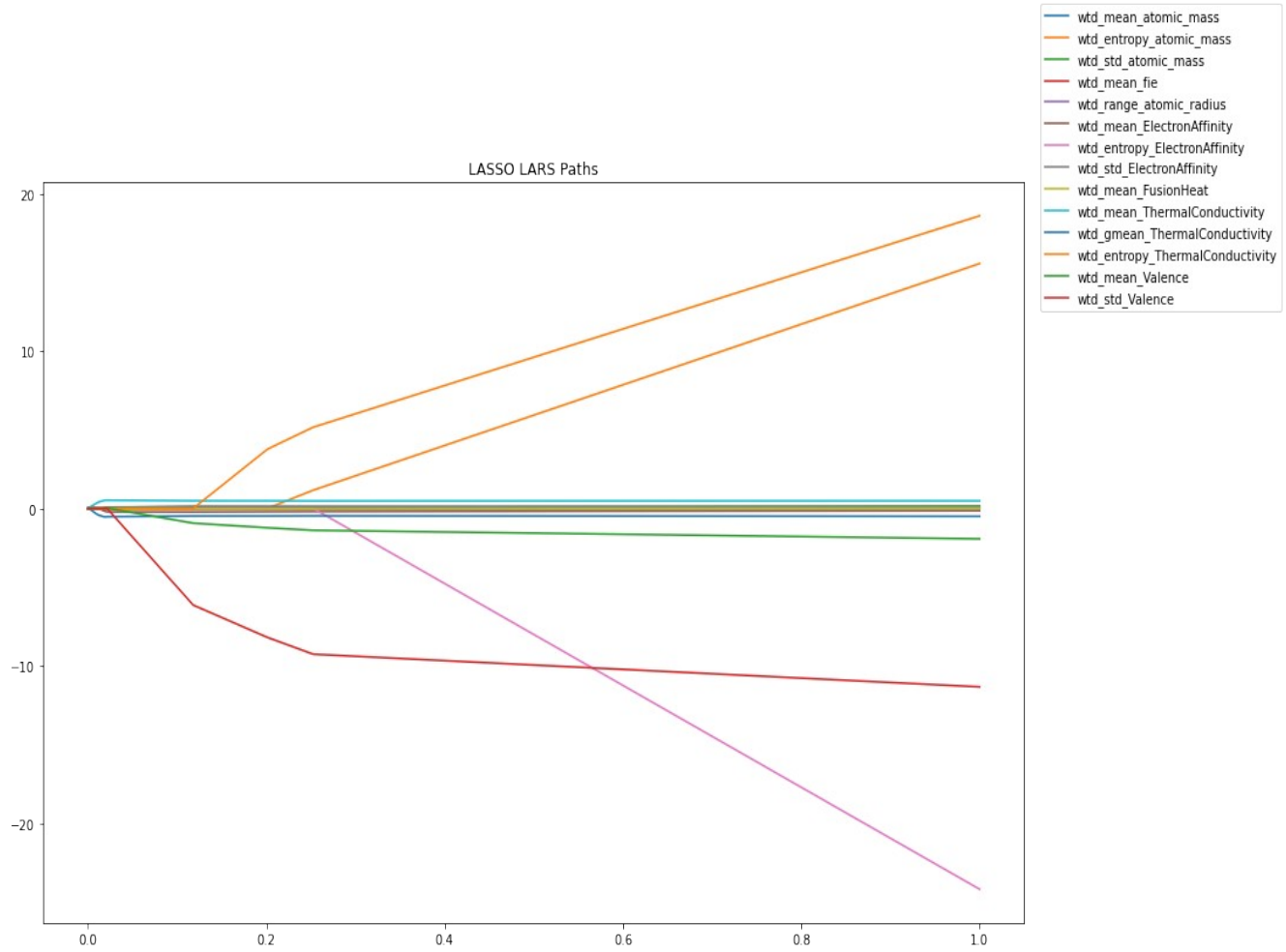
Comparing OLS , LASSO, Ridge,Elastic Net Coefficients:



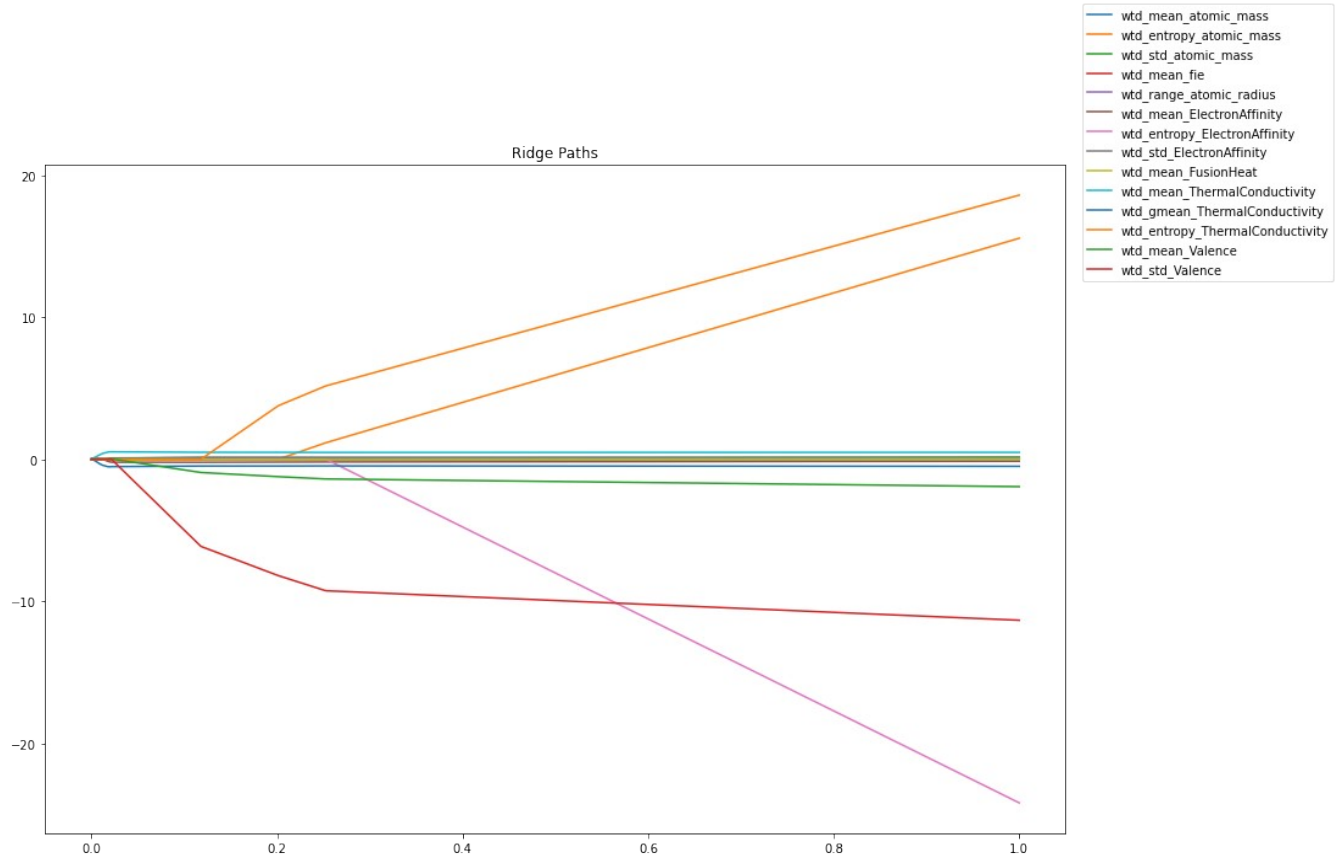
Comparing LASSO & Elastic Net Coefficients:



LASSO LRAS Paths:



Ridge Paths:-


Ridge Paths
wtd_mean_atomic_mass
wtd_entropy_atomic_mass
wtd_std_atomic_mass
wtd_mean_fie
wtd_range_atomic_radius
wtd_mean_ElectronAffinity
wtd_entropy_ElectronAffinity
wtd_std_ElectronAffinity
wtd_mean_FusionHeat
wtd_mean_ThermalConductivity
wtd_gmean_ThermalConductivity
wtd_entropy_ThermalConductivity
wtd_mean_Valence
wtd_std_Valence
20
10
0
−10
−20
0.0
0.2
0.4
0.6
0.8
1.0