Assignment 1 - MAST30034

Binh (Brian) Duc Vu
Student ID:1053531

September 9, 2021

1 Synthetic dataset generation, data preprocessing and data visualization.

1.1

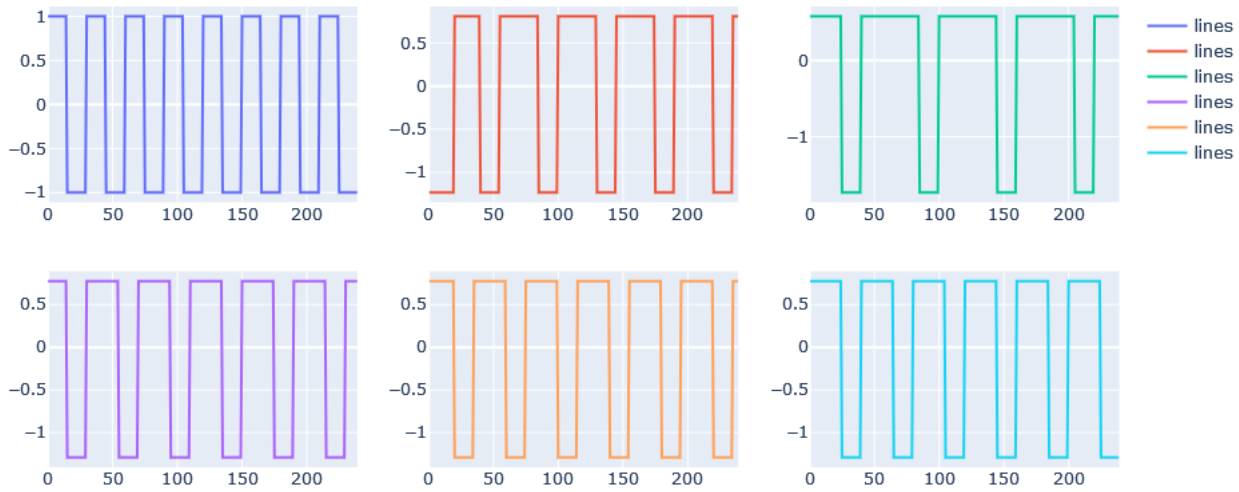


Figure 1: Constructed TC

The reason why the TC is standardized instead of normalized is because we do not want to change the distribution of the TC, which the normalization will do.

1.2

From the generated correlation matrix, we can see that the vector pairs 4-5 and 5-6 are relatively correlated, whereas in other pairs there is minimal correlation.

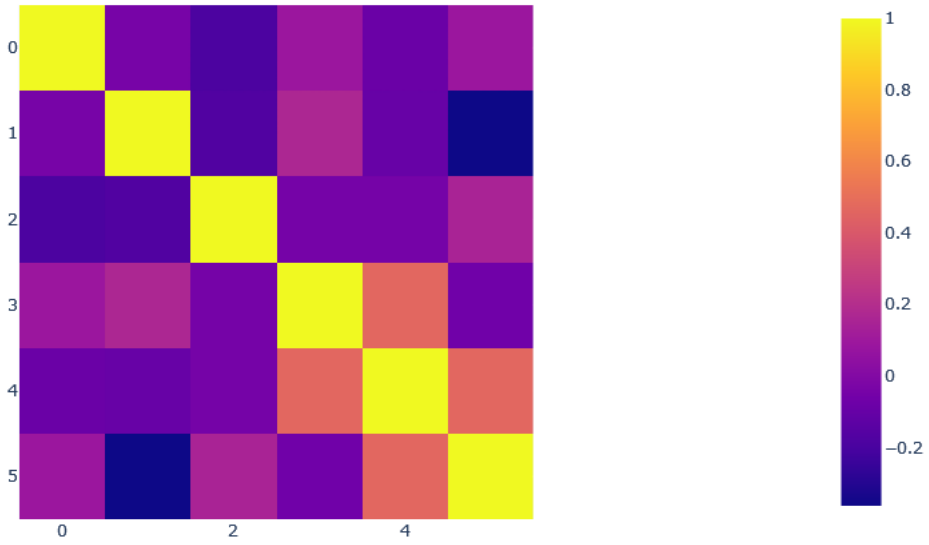
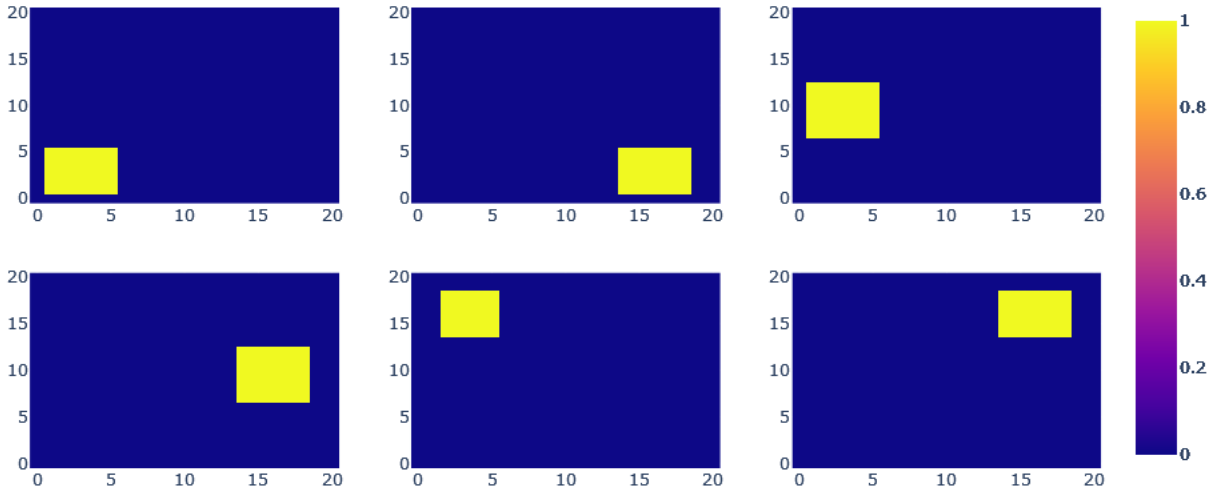
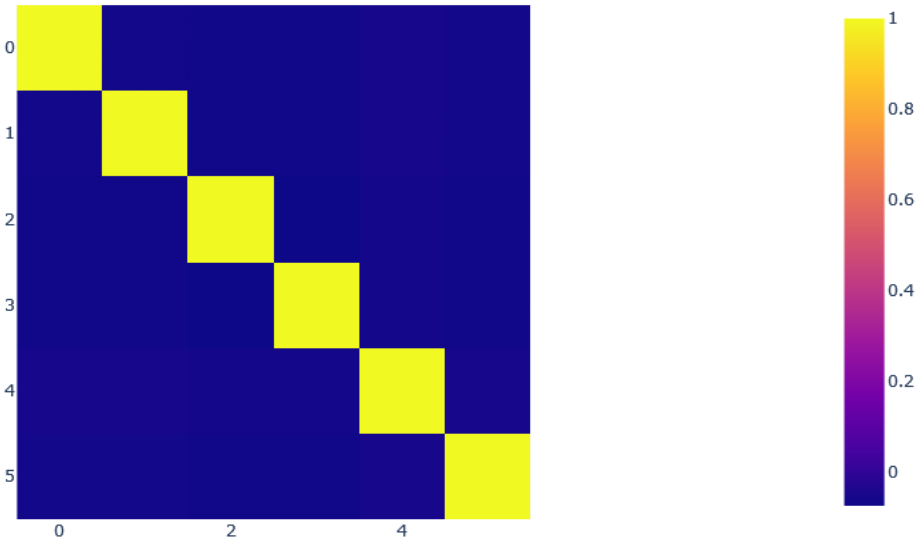


Figure 2: TC correlation heatmap

1.3





From the figure above, 6 vectored SMs are independent. The standardization of SMs is not important because the values are fixed at 0-1, and the number of 1 pixels are the same in each SM vector - so if we perform standardization, the vectors will have different values but the information it gives is unchanged.

1.4

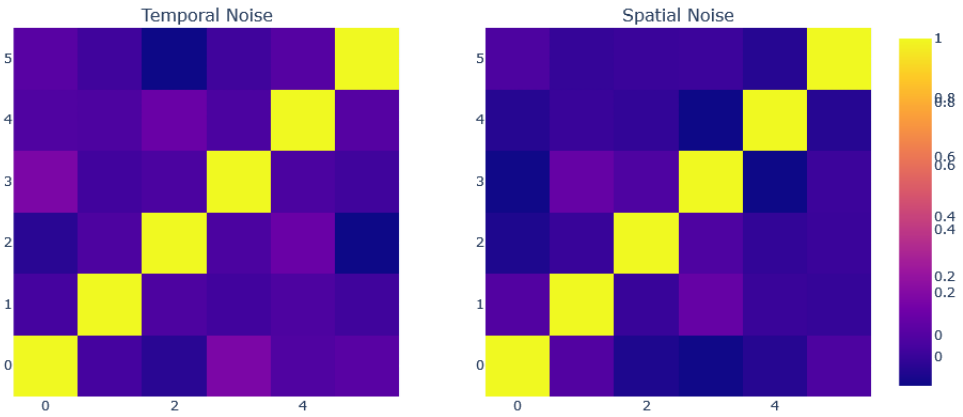


Figure 3: 6x6 CM for each noise source

There doesn't seem to be any correlation across sources for the generated spatial and temporal noise.

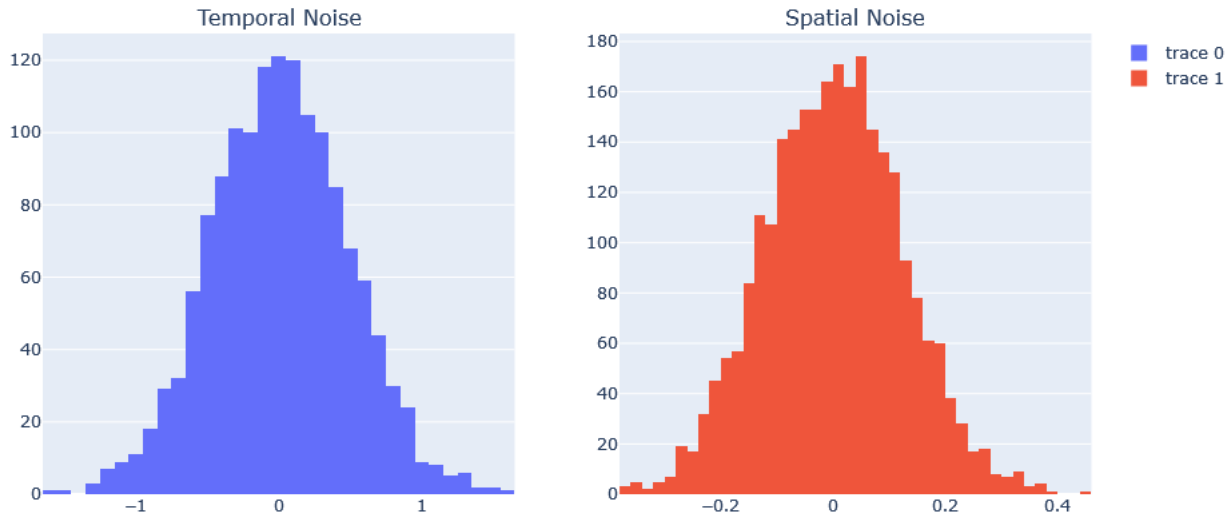


Figure 4: Histogram for noise sources

From the histogram of temporal and spatial noise, it is evident that the data follows the normal distribution - which is expected, as we generated it from the normal distribution.

```
The mean of the Temporal Noise matrix is:  0.010619502329591095
The mean of the Spatial Noise matrix is:  0.0010058573933117334
The variance of the Temporal Noise matrix is:  0.23031643820594225
The variance of the Spatial Noise matrix is:  0.014999804646216138
The 1.96sigma is for Temporal Noise is 0.98
The 1.96sigma is for Spatial Noise is 0.24004999479275144
```

Figure 5: Mean and Variance of noise

From the values calculated, the mean and variance are close to the desired distribution, and the variance of the matrices falls within the 1.96σ range, so all our requirements are satisfied.

1.5

The two products aforementioned can exist, and in fact will be kept in the matrix X. These two additional products just serve as additional variance to our X matrix. Furthermore, since they are still correlated to the variables SM and TC, they will be reduced as part of any Regression algorithm that is used.

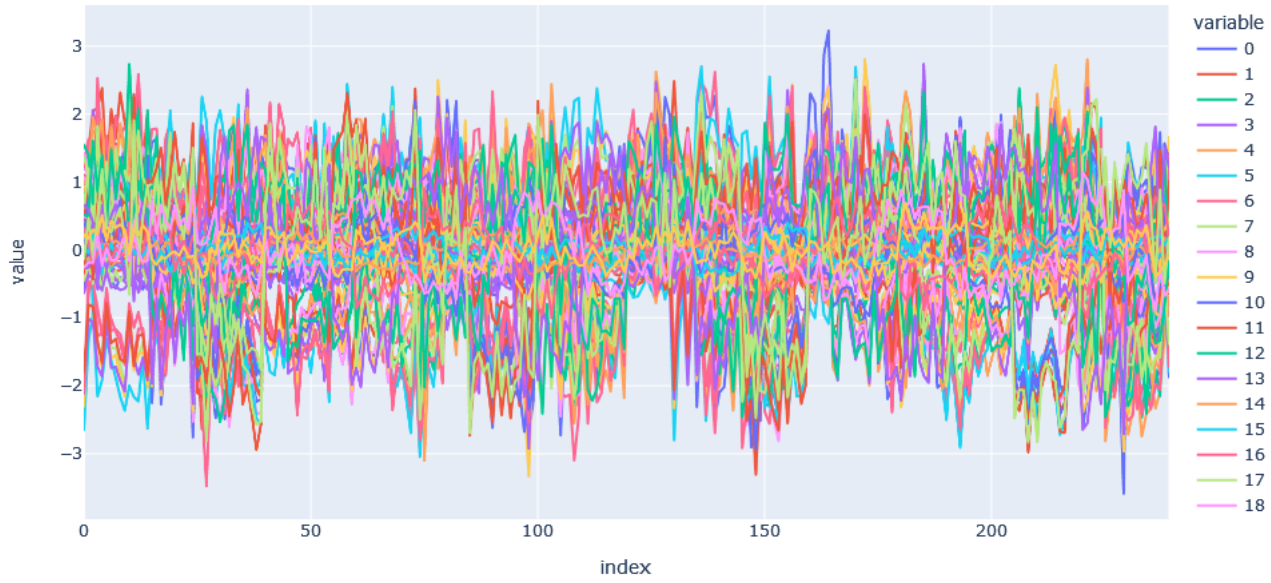
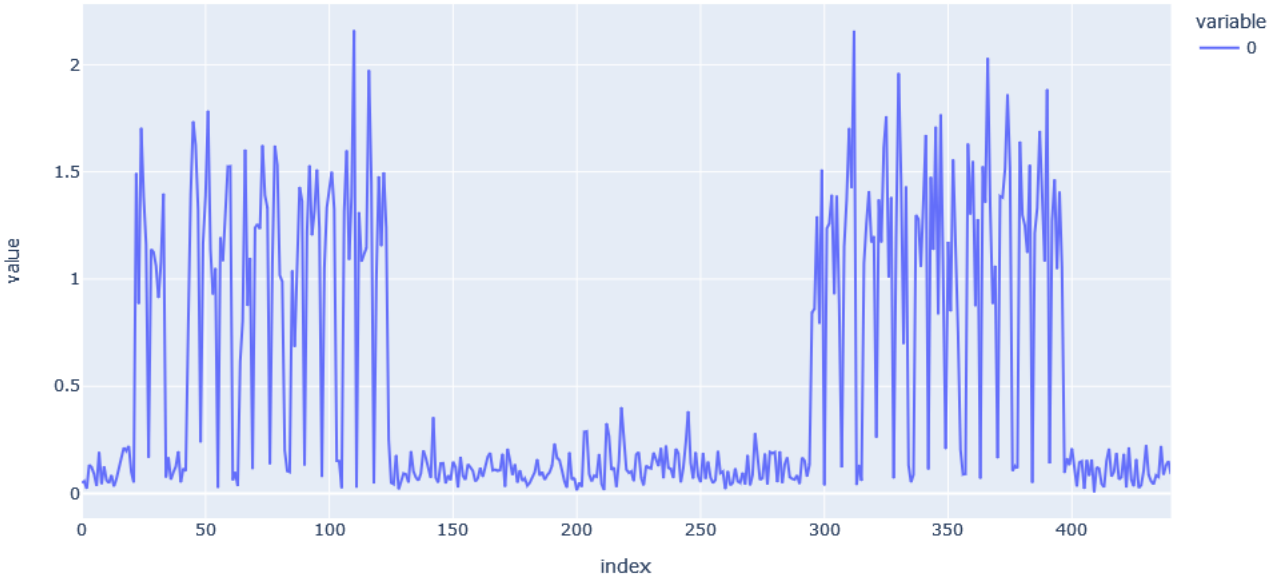


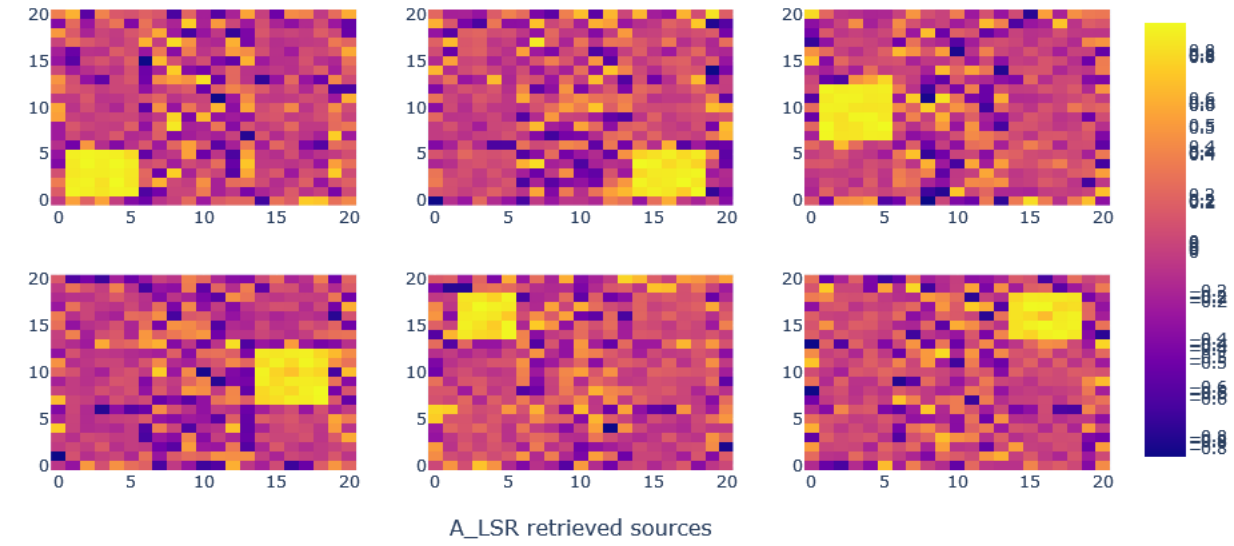
Figure 6: 100 randomly generated time series

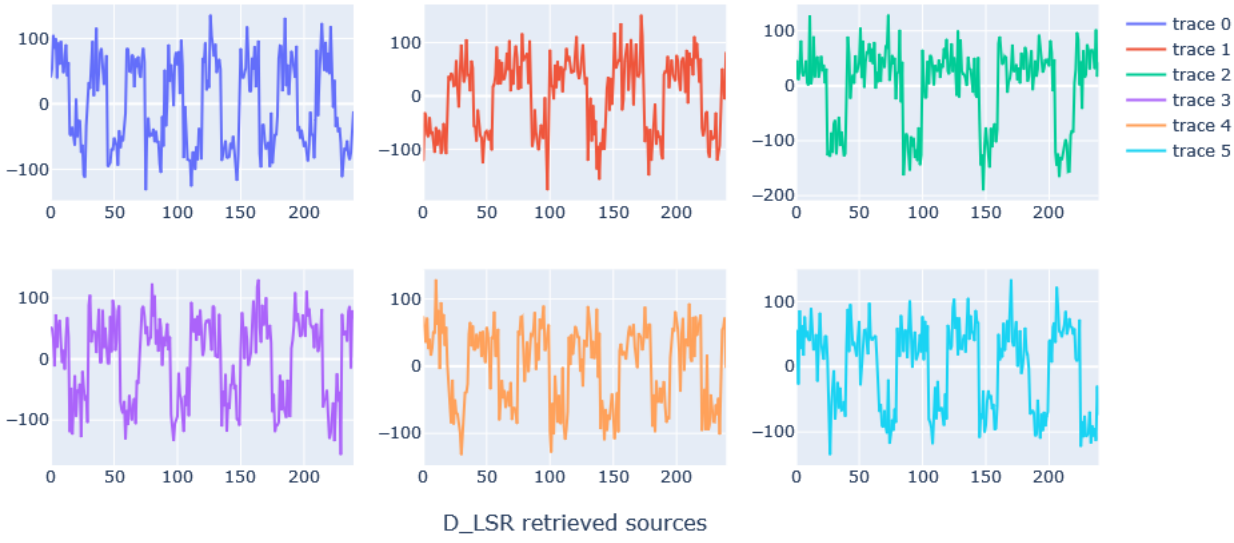


From the variance plot above, we can observe that most of the variance of the columns of X is relatively low. However, in the first and final thirds of the matrix, there are multiple jumps in variance in variables - and these jumps are relatively consistent in height.

2 Data analysis, results visualization, performance metrics

2.1





The reason why the linear relationship does not exist for column 4 when it does for column 3 is because these two columns were sliced differently - with different increment vectors and different duration of ones. With little relationship between the two columns themselves it is unlikely that column 4 will also be correlated with X's column 30.

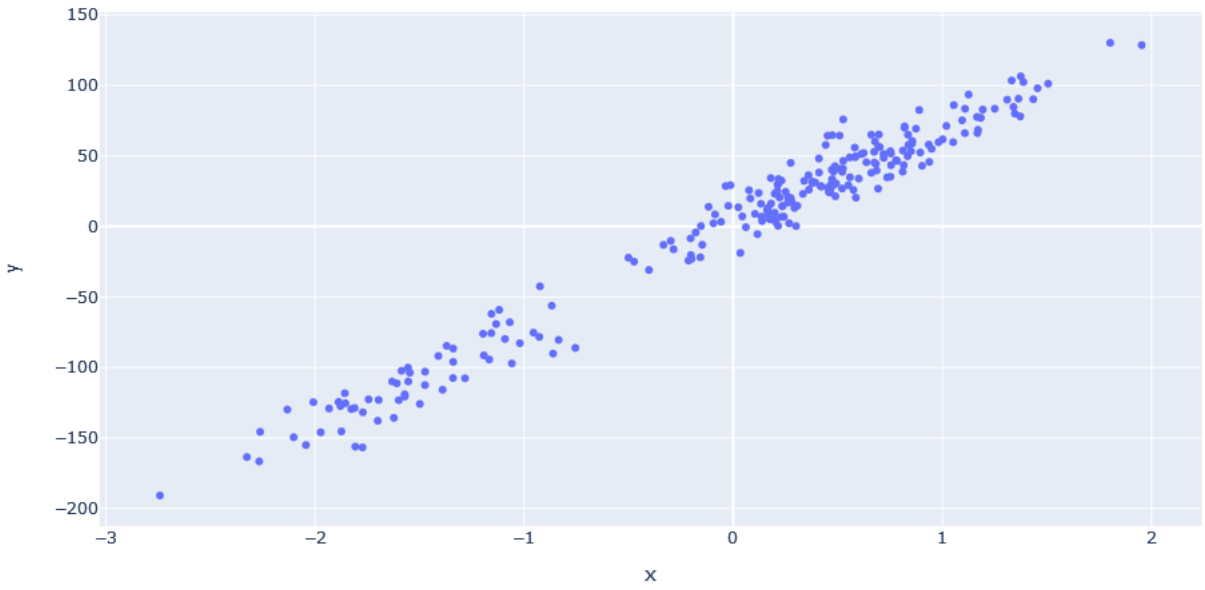


Figure 7: Column 3 of D and column 30 of standardized X

2.2

Performing Ridge Regression and selecting the penalty term of 116 gave the sum of correlation vectors for Ridge Regression to be 5.42, where as LSR gave 5.32 - Ridge Regression ¿ LSR, as expected.

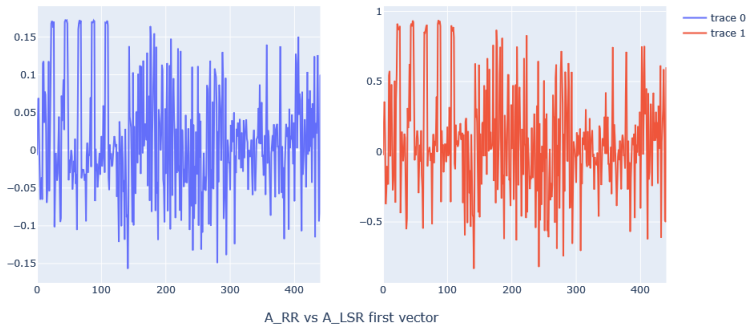


Figure 8: First vectors of A

From the above figure, the values of A in Ridge Regression is shrinking towards zero.

2.3

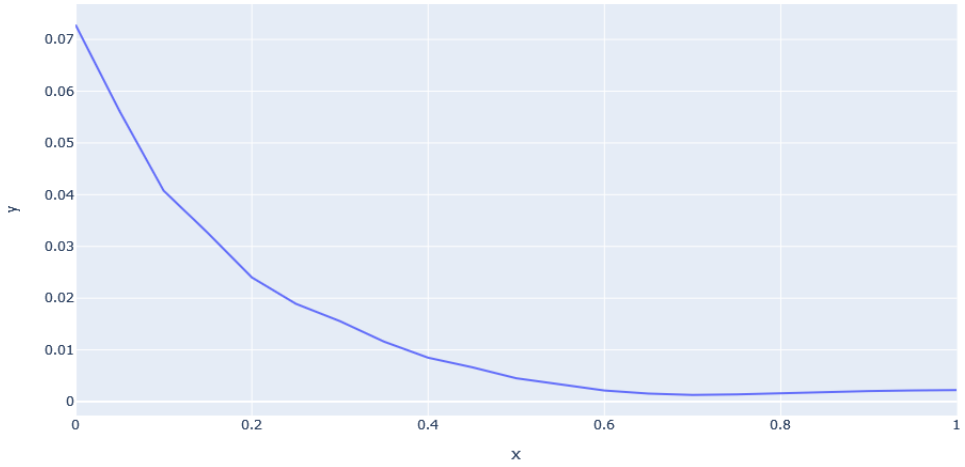


Figure 9: MSE against rho graph

From the plot above, the value at which minimum MSE was found is 0.7 - which is okay to select. After this point, the MSE began to increase again.

2.4

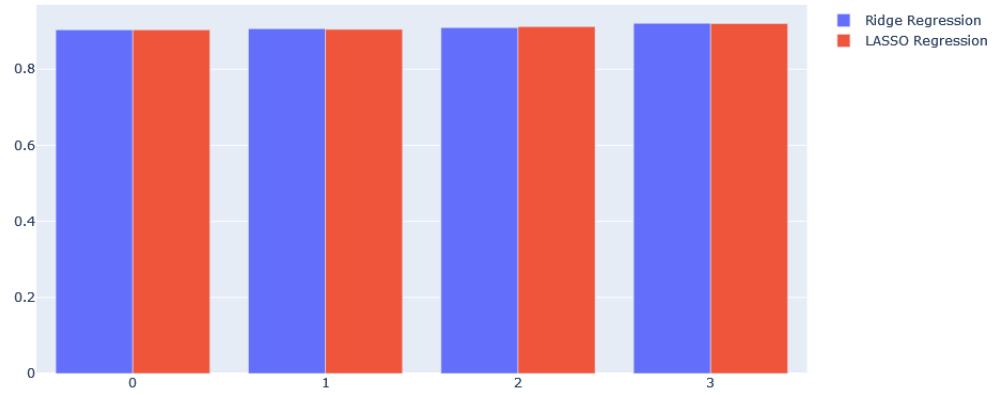


Figure 10: Correlation bar chart for TC

The correlation sum for TRR was 3.6406 against TLR's 3.6408 - a marginal improvement. However, the correlation sum for SRR was 2.24, where as SLR was 3.74 - a substantial increase. This is indicative of there being a lot of 'noisy' parameters in Ridge Regression A that were not important - and that most of these parameters were reduced to zero in the LASSO Regression A.

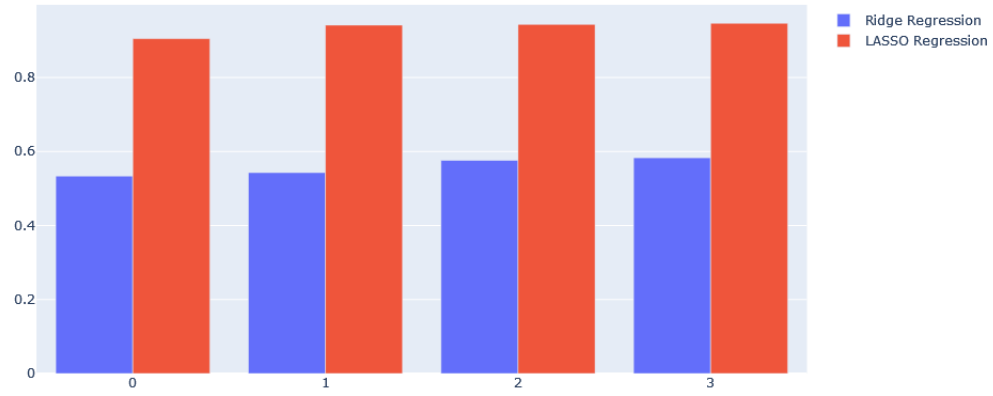


Figure 11: Correlation bar chart for SM

2.5

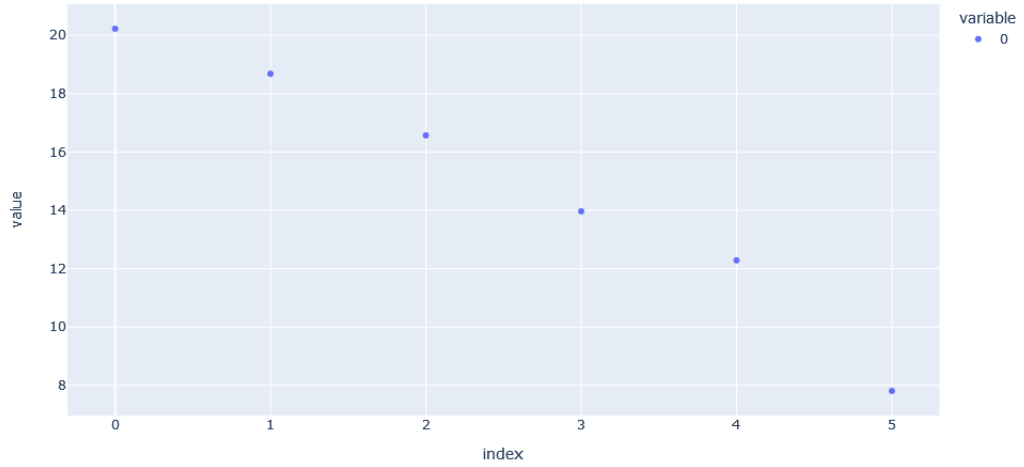


Figure 12: Eigenvalues for principal components

The 6th principal component has the smallest eigenvalue, as indicated by the plot.

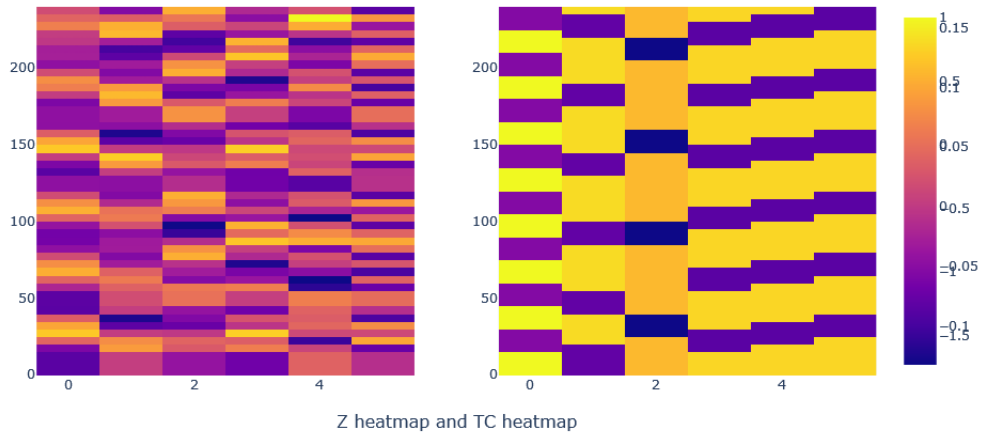
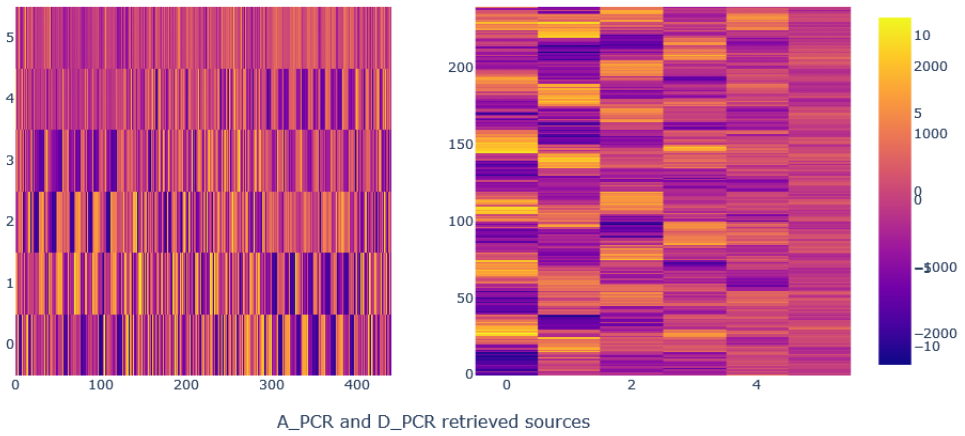


Figure 13: Plot of Z and TC

From the heatmap of Z and of TC, we see that the PCs of Z do not really closely follow the shape of TC. This is likely because, in construction, the PC method imposed restrictions on TC without knowing how TC affects X. In this case, the TC constructed has a lot of variables with similar values, and therefore some of these will not seem valuable. However, its nature as a time series data means that the sequence the variables appear is very important - but PC do not know this, and may not fully capture it.



Evident in the plot of PCR retrieved sources, the realisations of A and D from PCR is quite poor - and much poorer than the retrieved sources for other regressions investigated. The underlying cause for this is likely due to the problems with PC explained above. The rho value also did not undergo careful selection, and thus may be an imperfect choice.

- End of Document -