Yuxuan Yang
815565
yuy15@student.unimelb.edu.au

Assignment 1
MAST30034: Applied Data Science

2021-09-09

Question 1

1.1 Plot all TCs as six subplots.Why not normalize (divide by *l-2 norm*) the TCs instead of standardizing it?

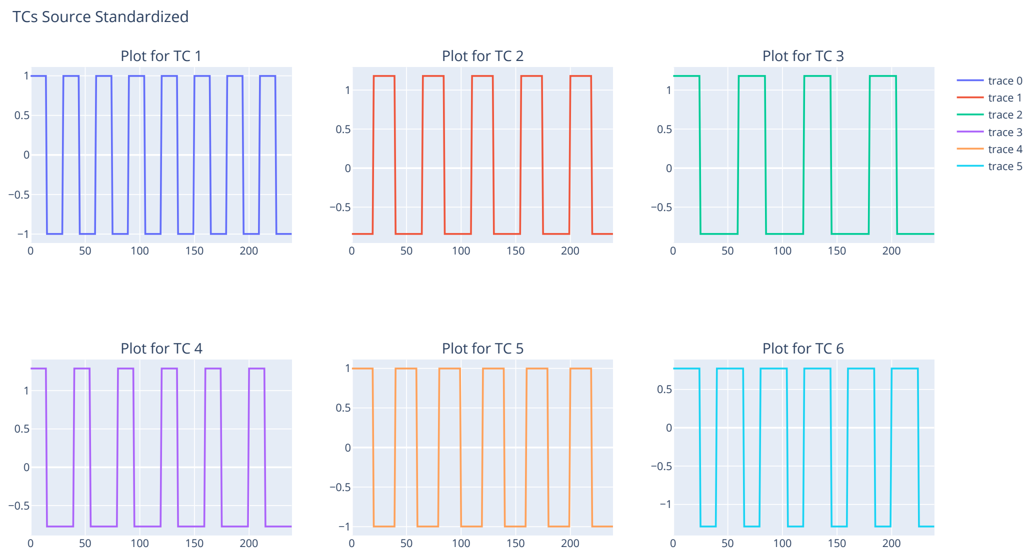

Figure 1: TC Matrix Sources

The l2-norm $\frac{X - X_{min(i)}}{X_{max(i)} - X_{min(i)}}$ take account of maximum and minimum value within i-th column of X and it is prone to outliers, therefore Standardization $\frac{X_i - \mu_{X_i}}{\sigma_{X_i}}$ is preferable options in this case as it utilizes standard deviation as its fractional denominator and therefore is less likely effected by range of extreme local minimum or local maximum.

1.2 Construct a CM that represents correlation values between 6 variables. Show its plot, and can you tell visually which two TCs are highly correlated? If not, can you tell this from CM?

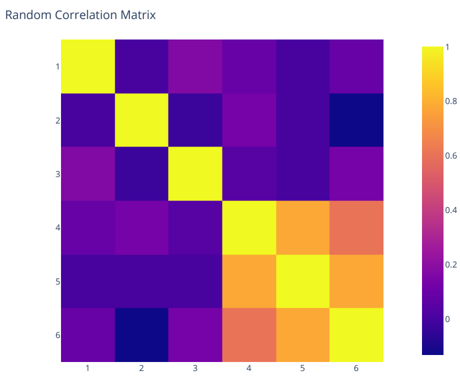

Figure 2: Random Correlation Matrix

The 4th, 5th, 6-th components of TCs are highly correlated as shown through their brighter color on the color scale. This is likely due to similar cycles contributed by identical increment values.

1.3 Plot these SMs in six subplots. Using CM show if these 6 vectored SMs are independent? For our particular case, why standardization of SMs like TCs is not important?

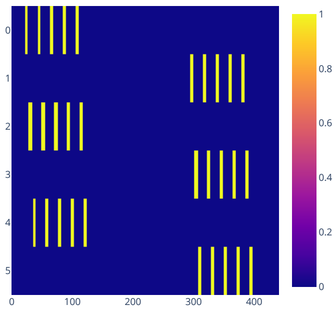

Figure 3: Correlation Matrix for 6-vector SMs

According to SMs' correlation matrix (**Figure 3**) above, the 1s from the associated 6 vectors are non-overlapping from other respective vectors. Given that SM inputs are associated with binary data, the phenomenon implies that the occurrence of 1s are mutual exclusive in terms of associated vectors. By definition of independence where overlapping occurrence between the vectors are the product of probability of a single vector. Since the overlapping of occurrence is zero, is non-equivalent to that of the respective product of vector-wise probability.

Regarding standardization of SMs, all 6 vectors are in the same dimension with binary inputs implies that the values regardless of frequency or spatially are on the same scale. Hence standardization is not required compare to TCs.

1.4 Using a 6 $\times$ 6 CM for each noise type (spatial and temporal) can you show if they are correlated across sources? Also plot the histogram of both noise sources to see if they have a normal distribution? Does this normal distribution fulfils the mean and variance = 1.96 σ criteria relating to 0.25,0.015, and zero mean? Is there produce $\Gamma_t \Gamma_s$ correlated across V number of variables?

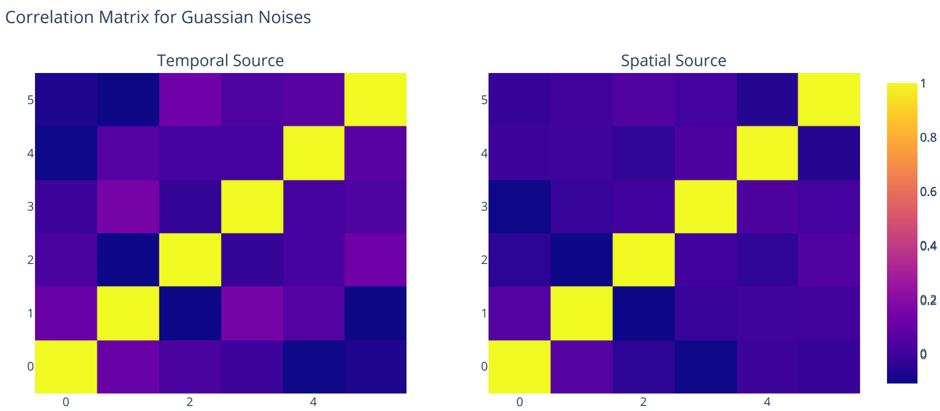

Figure 4: Correlation Matrix for Gaussian Noises

The respective CMs for spatial and temporal in **Figure 4** exhibit moderate correlation across the sources. This is shown in the light purples prevalent across the matrix blocks, indicating presence of correlation greater than 0.

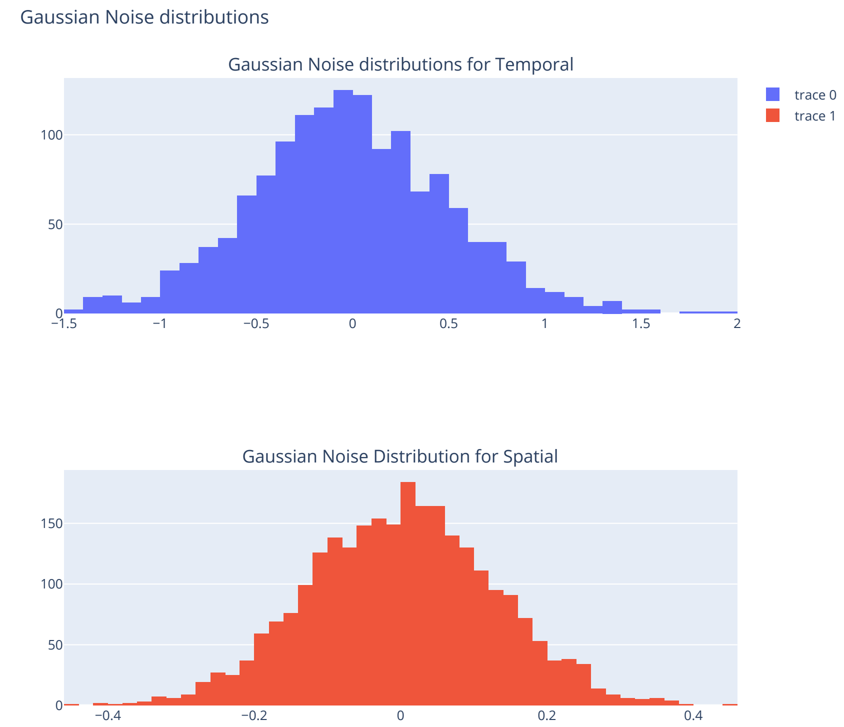

Figure 5: Histogram for $\mathbf{\Gamma_t}$ and $\mathbf{\Gamma_s}$

Through the histograms (**Figure 5**) which consist of Gaussian noises for both spatial and temporal data, both models visually illustrates normal distribution through their approximately equal tails and resemblance of a bell curve.

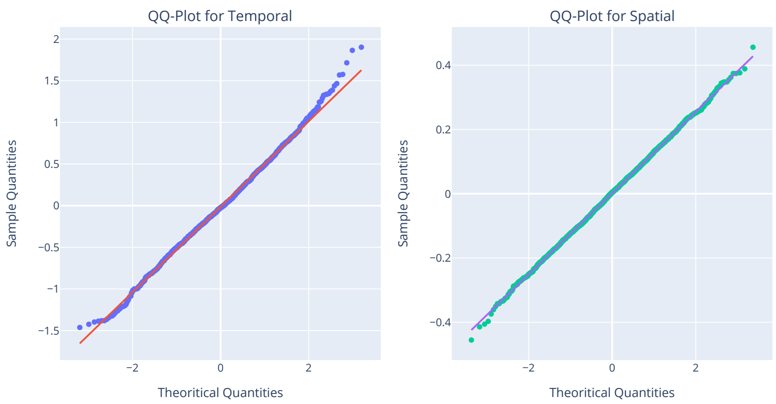

Figure 6: QQ-plot for $\mathbf{\Gamma_t}$ and $\mathbf{\Gamma_s}$

The QQ-plot of the respective model further demonstrated the models' normality as the model's quantile-wise value closely align to the theoretical value of a normal distribution with identical mean and standard deviation. The models' distributions thereby fulfill the mean and variance resembling 1.96σ criteria relating to 0.25, 0.015, and zero mean.

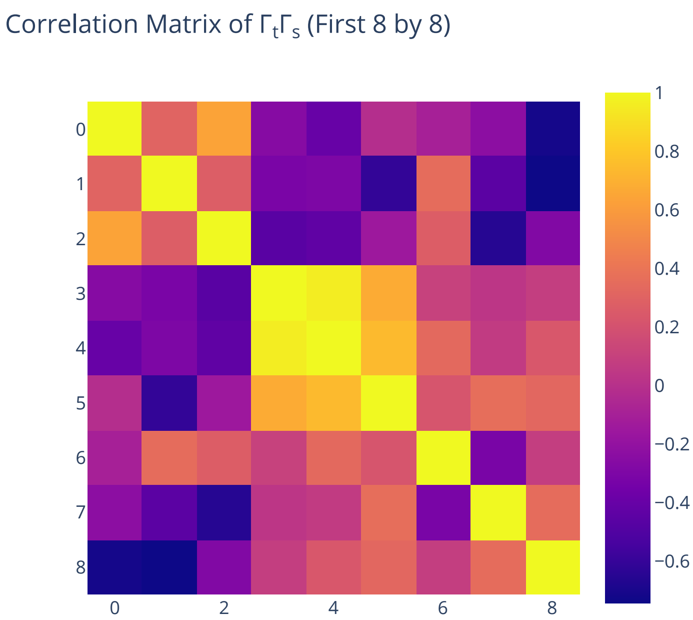

Figure 7: First 8×8 Correlation Matrix for $\mathbf{\Gamma_t \Gamma_s}$

According to the Correlation matrix of product $\mathbf{\Gamma_t \Gamma_s}$ (**Figure 7**), some vectors demonstrated moderate correlations. Examples are between 1st and 8th column, and 3rd and 4th column where the correlation is greater than 0.75

1.5 Can these products $\mathrm{TC} \times \mathbf{\Gamma_s}$ and $\mathbf{\Gamma_t} \times \mathrm{SM}$ exist, If yes what happened to them because if we keep them then we cannot fit our model onto LSR? Plot at least 100 randomly selected time-series from X as few of them are shown in Figure 2 (in the specification sheet). Also plot variance of all 441 variables on a separate plot. What information does this plot give you?

```
     TC + Gamma_t Shape:  (240, 6)
     TC + Gamma_s Shape:  (6, 441)
Gamma_s Shape:  (6, 441) TC shape:  (240, 6)
Gamma_t Shape:  (240, 6) SM shape:  (6, 441)
```

Figure 8: Jupyter Notebook Outputs

Both $\mathbf{TC} \times \mathbf{\Gamma_s}$ and $\mathbf{\Gamma_t} \times \mathbf{SM}$ exist due to the distributive law of matrices and that the number of column for the first product term is identical to the number of row in the second term (**Figure 8**). Therefore, dot product is satisfiable.

In regards with these two components, it would be reduced as part of Ordinary Least Square (OLS) process, as it minimizes the difference between labelled value and predicted value. In another word, the cost function will be optimized during regression fitting process where $TC \times \Gamma_s$ and $\Gamma_t \times SM$ will be reduced to 0.

Figure 9: 101 randomly selected time-series from **X**

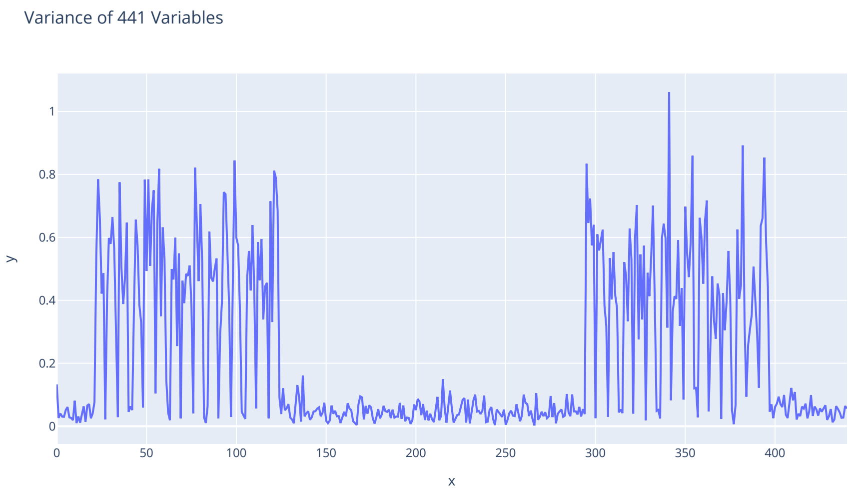

Figure 10: Variance of 441 variables

The variance graph shown in **Figure 10** demonstrated two clusters of variance for SMs and TC vectors where positional value equals to 1.

Question 2

2.1 Plot six retrieved sources using A_{LSR} and D_{LSR} side by side. Do a scatter-plot between 3rd column of D_{LSR} and 30th column of standardized X, you will find a linear relationship between them, why this does not exist between 4th column of D_{LSR} and same column of X.

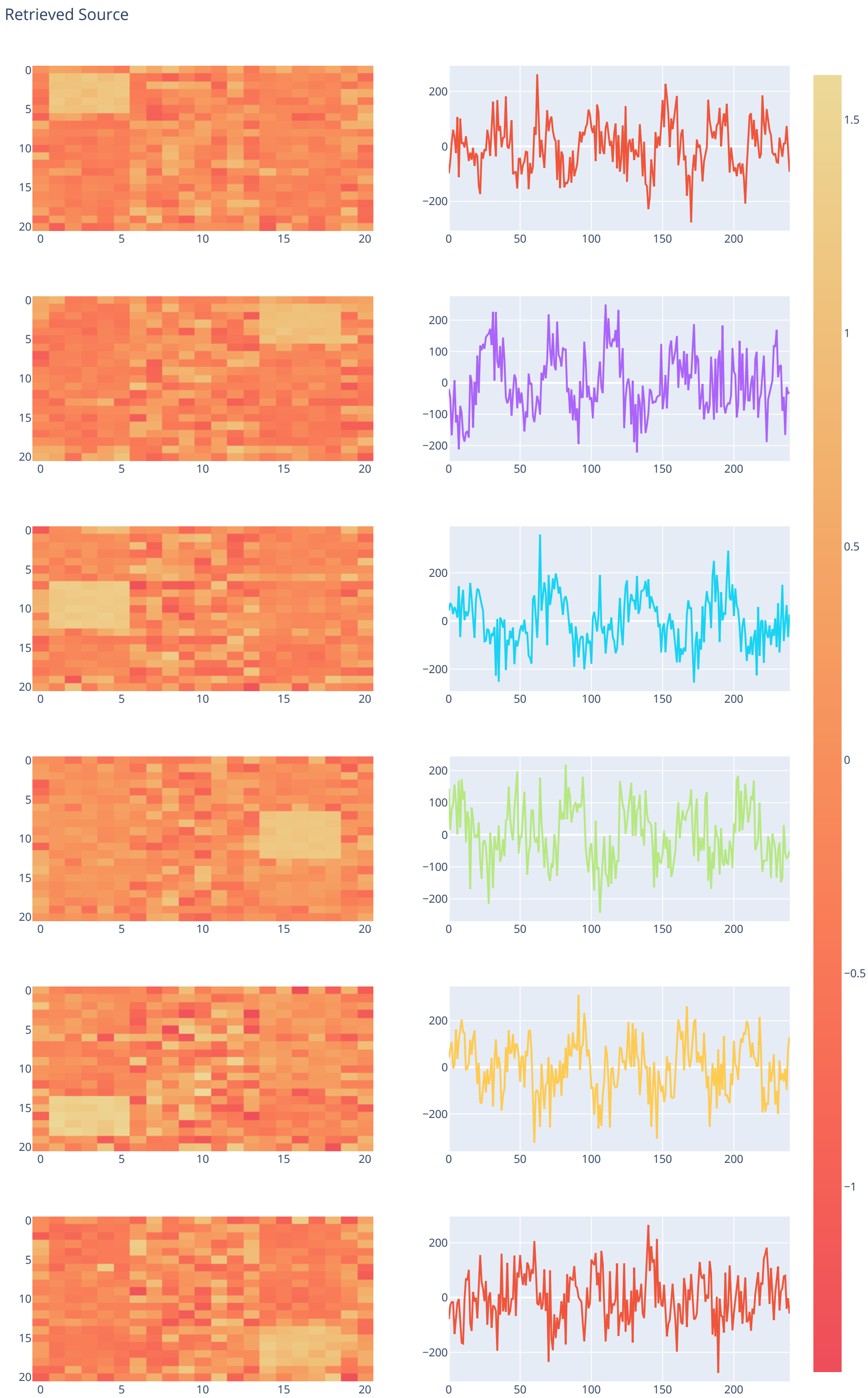

Figure 11: Six Retrieved Sources

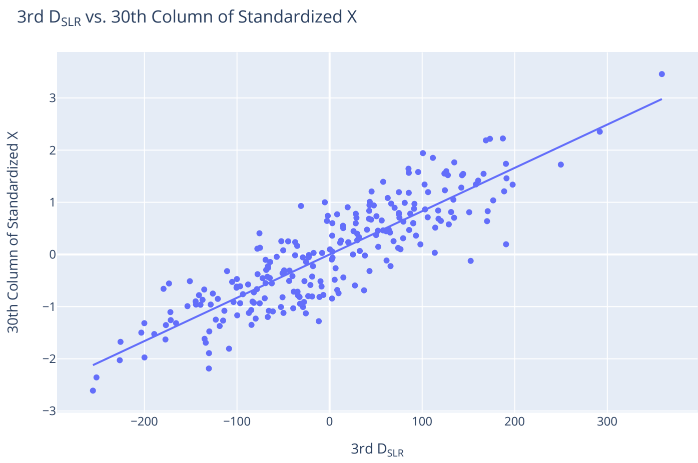

Figure 12: 3rd $\mathbf{D_{SLR}}$ vs. 30th Column of Standardized $\mathbf{X}$

A relationship exists between 3rd column of D_{LSR} and 30th column of standardized X but not the 4th column of D_{LSR} because at the 30th column of **SM** there is 1 at the 3rd column but 0 at the 4th column (slice output of the 30th column were `[0, 0, 1, 0, 0, 0]`), therefore D at the 4th column would not be exhibit linear relationship due to null space.

2.2 Calculate the sum of these correlation vectors c_{TLSR} and c_{TRR}. Also, for λ= 1000, plot first vector from A_{RR} and the corresponding vector from A_{LSR}, Do you find all values in a^1_{RR} shrinking towards zero?

The values for $\sum c_{TLSR}$ and $\sum c_{TRR}$ respectively are 3.5697 and 3.5297 and it satisfies the condition that $\sum c_{TLSR} > \sum c_{TRR}$

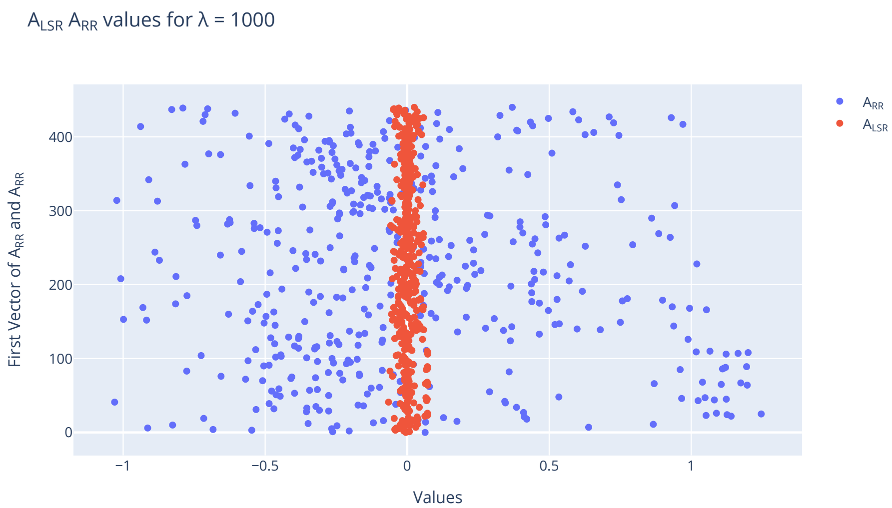

Figure 13: 1st vector of $\mathbf{A_{RR}}$ vs . 1st vector of $\mathbf{A_{LSR}}$

Most values of $\mathbf{a^1_{RR}}$ are shrinking towards to zero as λ coefficients applies a penalty rate on the co-efficients to prevent overfitting.

2.3 Plot average of MSE over these 10 realizations against each value of ρ. At what value of ρ do you find the minimum MSE? Is it okay to select this value? At what value of ρ did MSE started to increase again (LR diverged).

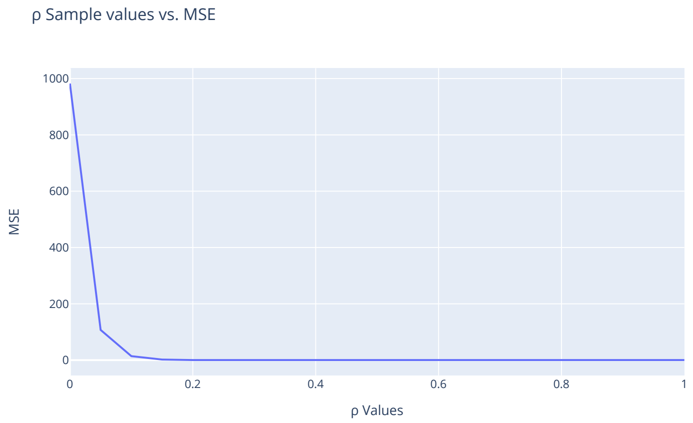

Figure 14: Average MSE over 10 realizations against ρ

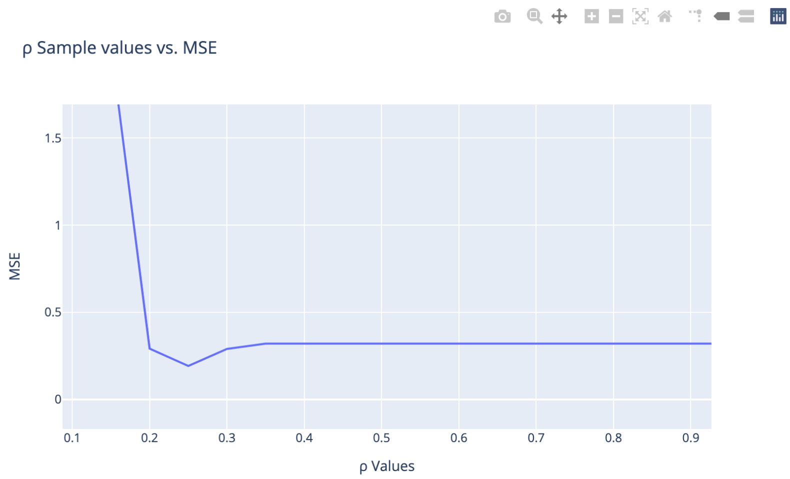

Figure 15: Average MSE over 10 realizations against ρ (zoomed-in)

Based on the MSE computed via generated standardized X in **Figure 14-15**, minimal MSE occurs at the approximate $\rho = 0.25$.

The values of ρ although can be selected as regularization coefficient given that it provides LASSO regression with the approximate minimal MSE, it would be optimal to run more iterations to precisely find out the rho beyond in more digits. According to the graph above, MSE started increased again and saturated after $\rho = 0.25$

2.4 Calculate the sum of these four correlation vectors c_{TLR}, c_{TRR}, c_{SLR} ,and c_{SRR} respectively. Plot side by side in form of 4 columns estimates of D and A for both RR and LR to know the difference visually. You will see a major difference in estimates of A in terms of false positives. Can you mention the reason behind this difference?

The sums of correlation vector $\sum c_{TLR}$, $\sum c_{TRR}$, $\sum c_{SLR}$, and $\sum c_{SRR}$ are 4.4238, 3.5698, 4.9781, and 3.2297 respectively and these values satisfies the constraint $\sum c_{TLR} > \sum c_{TRR}$ and $\sum c_{SLR} > \sum c_{SRR}$.

According to 4 randomly selected column estimates for both RR and LR from the graph above, A_{RR} and D_{RR} have noticeable larger value range compare to its LR counterpart. Specifically, the columns estimates $A_L R$

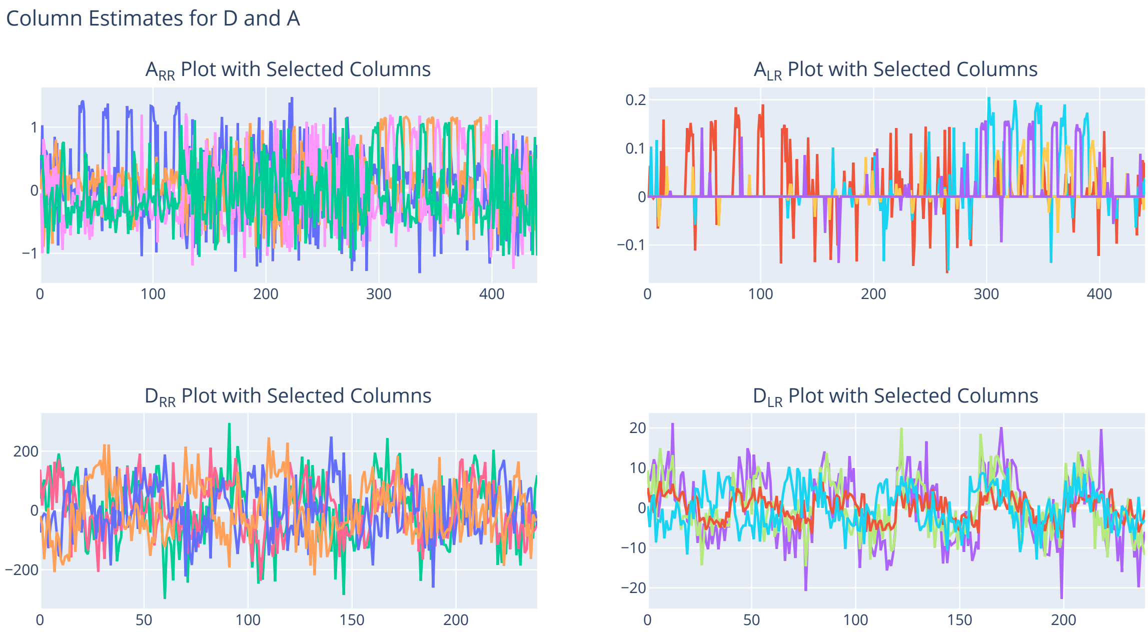

Figure 16: 4 Columns estimates of **D** and **A** for both RR and LR

demonstrated larger tendencies to produce false positives than A_RR as the graph in A_LR resembles closely that of the Least Square Regression. The reason this is because L1 shrinkage promote sparsity i.e. characteristic of storing highly significant coefficients that are either close to zero and far from zero, and L1 would shrunk those that are close to zero to zero. Thereby, such as shrinkage will mostly likely contribute to prediction error and hence higher false positives.

2.5 Estimate PCs of the TCs and plot their eigen-values. For which PC the eigen value is the smallest? Plot the regressors in Z and source TCs side by side. Did you notice deteriorated shape of PCs? Why the shape of TCs has been lost? Now keeping all components in Z apply lasso regression on X using ρ = 0.001 and then Plot the results of D_{PCR} and A_{PCR} side by side (note that $A_{PCR} = B$ and your regressors are in Z (PCs of the TCs)). Did you notice the inferior performance of PCR compared to the other three regression models? Why is that so?

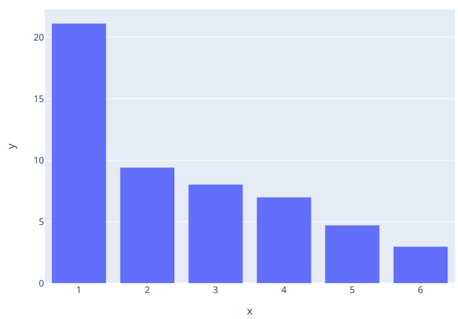

Figure 17: Eigenvalues of principal components of TCs

According to the **Figure 17**, the 6th PC has the smallest eigenvalue.

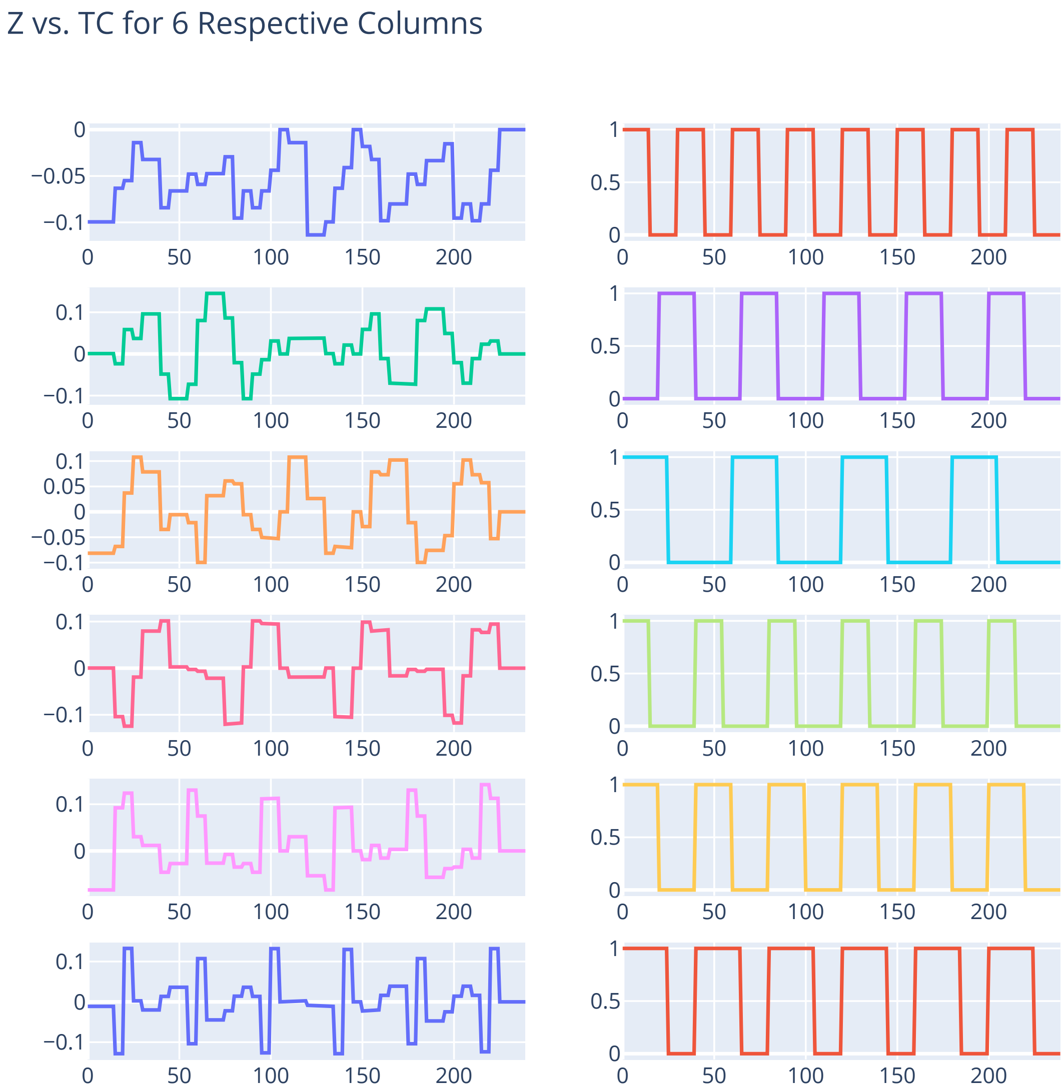

Figure 18: Regressors in **Z** and Source **TC**

By Looking at the graph above in comparison the regressor **Z** and source **TC**, there are noticeable but chiselness in Z compare to the rectangular graphs in TCs. The loss of TCs' shape are likely due to the information loss during the dimensional reduction caused by unexplained variance in the original dataset.

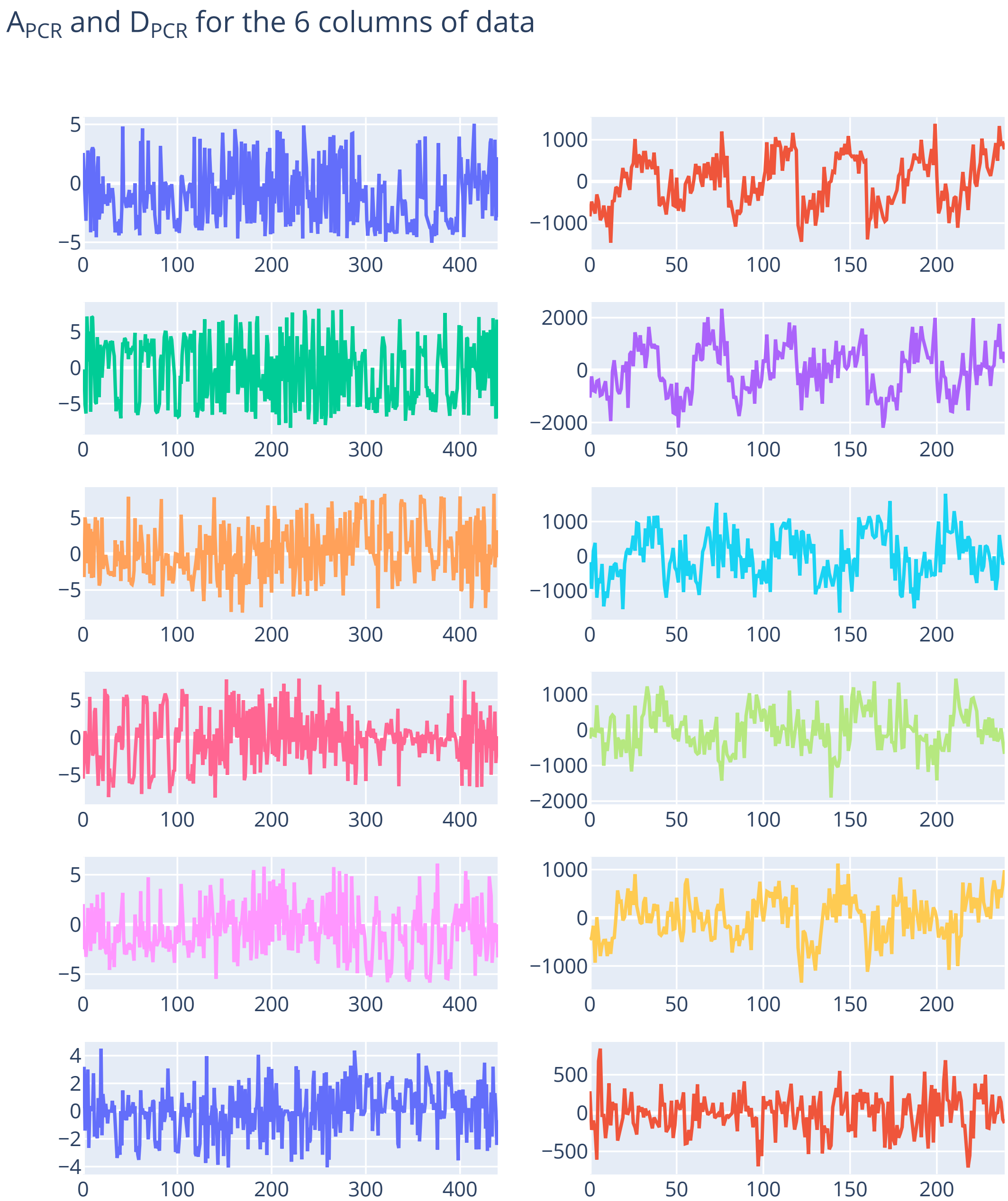

Figure 19: $\mathbf{D_{PCR}}$ and $\mathbf{A_{PCR}}$

Through visual examinations of $\mathbf{D_{PCR}}$ and $\mathbf{A_{PCR}}$ in **Figure 19**, both exhibit inferior performances compare to the previous regression models due to the regression only take a subset of all the principal components and that lasso shrinkage is dependent on the size of corresponding eigenvalues. Hence, the discard of smaller eigenvalue components in principal components regression might likely resulted its sub-par performances.