MAST30034 2021 Semester 2

Assignment 1

Wai Lam Wong
Student ID: 1007788

April 27, 2022

1 Synthetic dataset generation, data preprocessing, data visualization

1.1 Question 1.1

A matrix **TC** of size 240 x 6 consisting of six temporal sources was created as shown in Figure 1. Following on from that, each time course (TC) will then be standardized by subtracting its mean and also by dividing it by its standard deviation. Standardization was used instead of normalisation by ridge regression because normalisation does not ensure the TCs are bias free (not centered around the origin or mean $\neq 0$) and equally important (no unit variance). Normalisation will only shift and rescale the TCs so that they are in the range of 0 and 1.

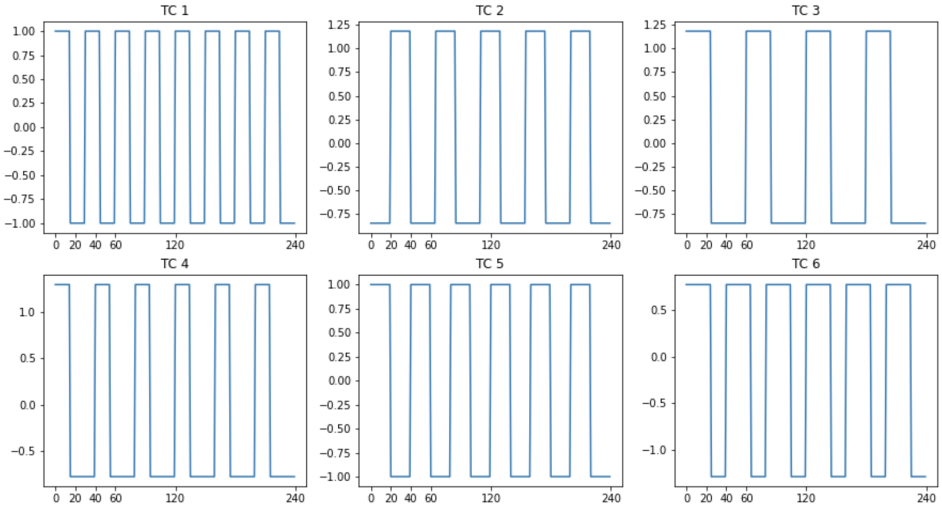

Figure 1: The 6 standardized TCs

1.2 Question 1.2

Figure 2 displays a correlation matrix (CM) that was generated using the 6 variables from **TC**. We can observe from Figure 1 that the 4th TC and 5th TC appear to be highly correlated with each other. However, after observing the CM from Figure 2, the 4th and 6th TC have equally high correlation values as compared to the 5th and 6th TC, with a correlation value in between 0.6 and 0.8.

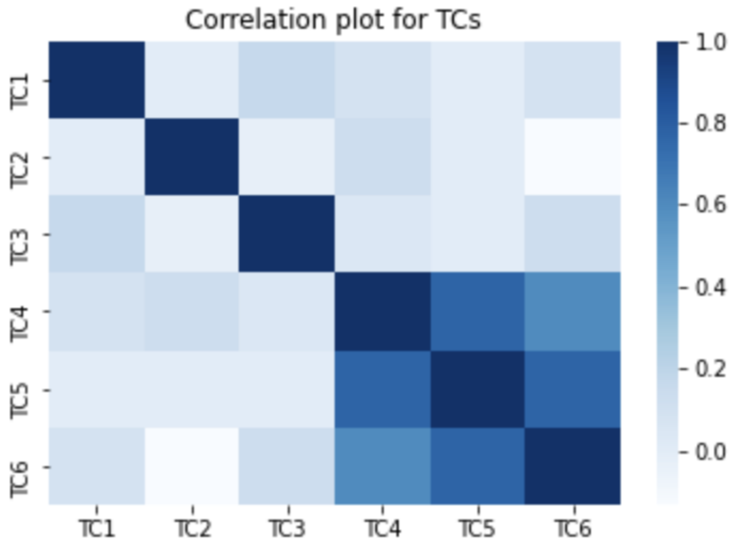

Figure 2: Correlation plot for TCs

1.3 Question 1.3

Figure 3 illustrates the 6 spatial maps (SM) that have been plotted in order to construct the **tmpSM** array. The CM shown in Figure 4 shows that the 6 SMs are independent from one another. Due to the independence of the SMs from each other, they do not need to be standardized like TCs as each SM are computed independently. Hence, rescaling the SMs to be on identical scales through standardization would not prove any importance.

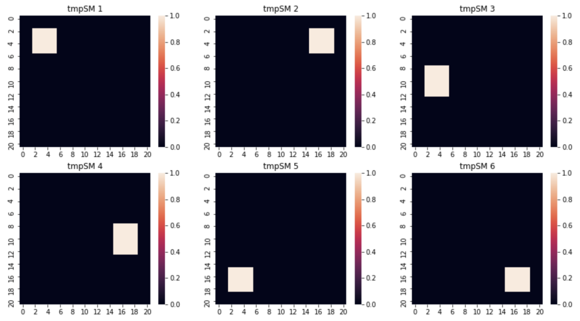

Figure 3: 6 SMs for tmpSM

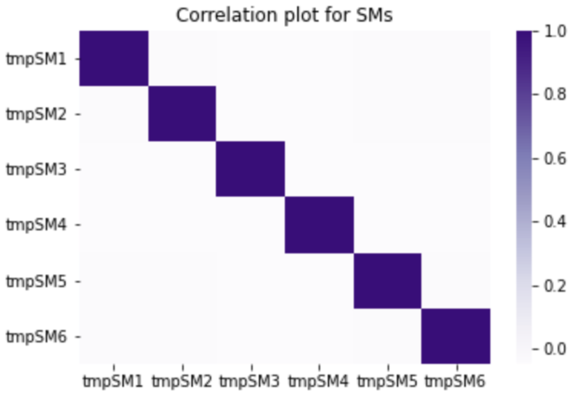

Figure 4: Correlation plot for SMs

1.4 Question 1.4

From the CMs in Figure 5 and Figure 6, we can observe that the temporal and spatial sources have either low correlation across the sources or have no correlation at all. Other than that, histograms for both temporal noise sources and spatial noise sources have been plotted, as shown in Figure 7 and Figure 8 respectively. From these graphs, we can see that both noise sources appear to be normally distributed. However, both the temporal and spatial noise sources do not fulfill the criteria of zero mean and variance $= 1.96\sigma$. The actual computed mean and variance for the temporal noise distribution is -0.00162 and 0.2578 respectively, whereas the value of 1.96σ is 0.9951. Furthermore, the actual computed mean and variance for the spatial noise distribution is -0.000771 and 0.0143 respectively, whereas the value of 1.96σ is 0.2343. Thus, we can conclude that variance $\neq$ 1.96σ for both sources. For Figure 9, the first 8 sources from the temporal noise and spatial noise sources were used as samples in order to plot the CM. We can see that some of the sources are correlated with each other whereas others are less correlated.

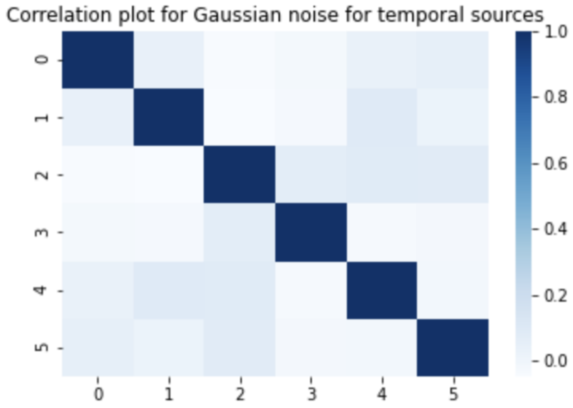

Figure 5: Correlation plot for Gaussian noise for temporal sources

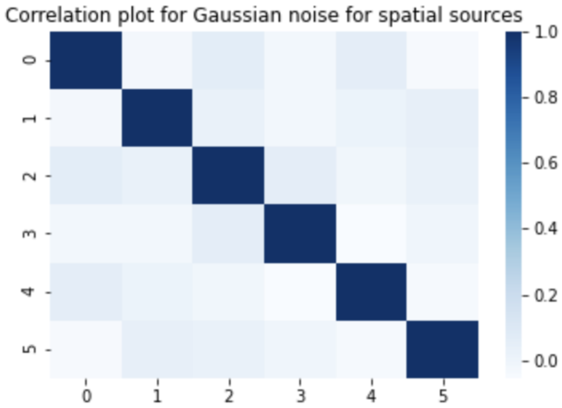

Figure 6: Correlation plot for Gaussian noise for spatial sources

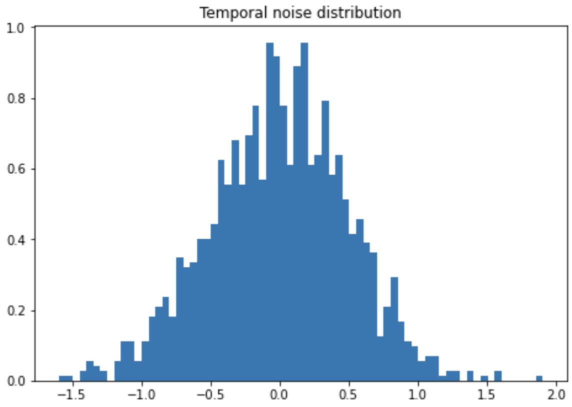

Figure 7: Temporal noise distribution

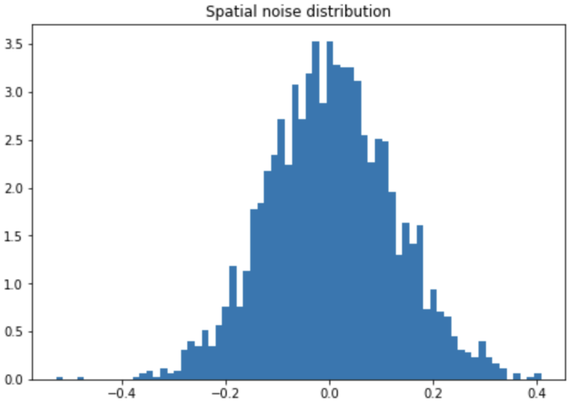

Figure 8: Spatial noise distribution

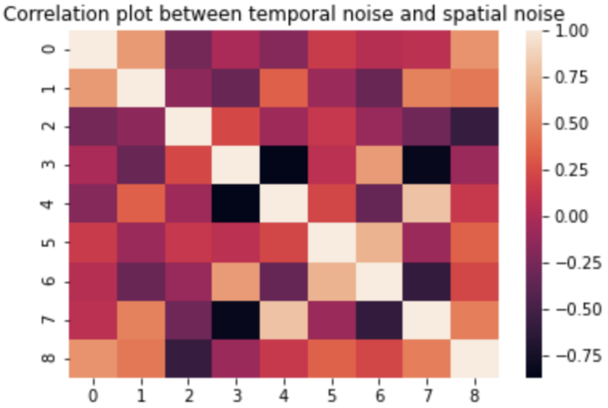

Figure 9: Correlation plot between temporal noise and spatial noise

1.5 Question 1.5

A synthetic dataset $\mathbf{X} = (\mathbf{TC} + \mathbf{\Gamma}_t) \times (\mathbf{SM} + \mathbf{\Gamma}_s)$ of size 240 x 441 was generated. Figure 10 shows a plot of 100 randomly selected time-series of X. The variances of all 441 variables was also plotted as shown in Figure 11. From the plot in Figure 11, we can see that most of the variables have low variances that are in the range from 0 to 0.5.

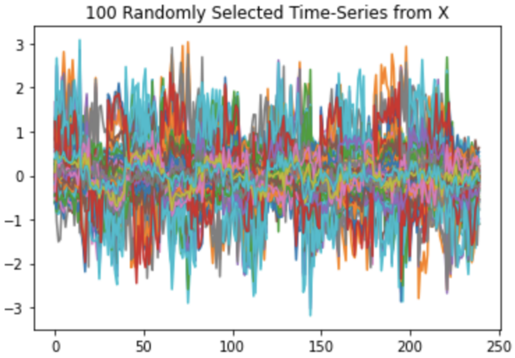

Figure 10: 100 randomly selected time-series from X

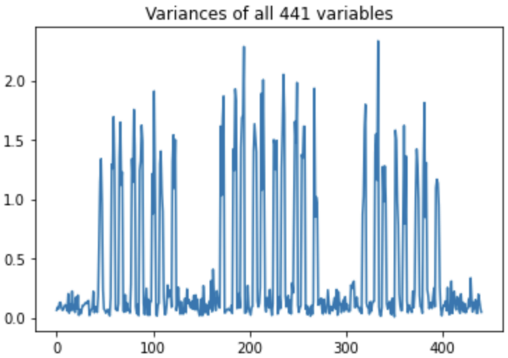

Figure 11: variances of all 441 variables

2 Data analysis, results visualization, performance metrics

2.1 Question 2.1

Least Square Regression (LSR) was used to estimate $\mathbf{A}$ using the least square solutions $\mathbf{A}_{LSR} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{X}$ and $\mathbf{D}_{LSR} = \mathbf{X}\mathbf{A}_{LSR}^T$. Figure 12 illustrates the retrieved SMs and retrieved TCs. A scatterplot between the 3rd column of $\mathbf{D}_{LSR}$ and the 30th column of standardised $\mathbf{X}$ as shown in Figure 13 shows a linear relationship between the two. However, Figure 14 shows that a linear relationship does not exist between the 4th column of $\mathbf{D}_{LSR}$ and the 30th column of standardised $\mathbf{X}$ due to the presence of noise in the data that may have an influence on the graph. Additionally, the 30th pixel position in the standardized X dataset is not filled up by the 4th SM, thus the 4th TC is unable to form a linear relationship with the 30th column from X.

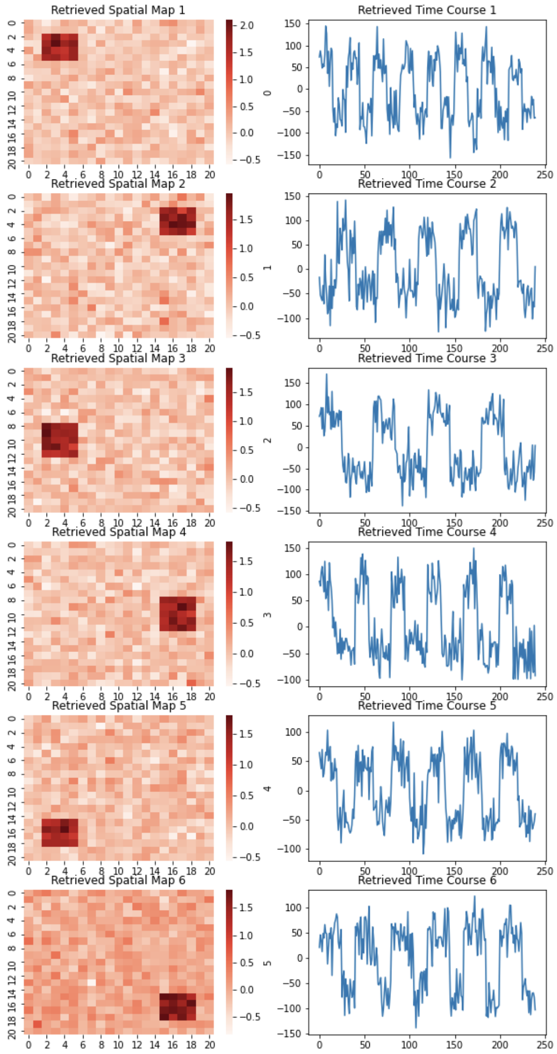

Figure 12: Retrieved Spatial Maps and Retrieved Time Courses

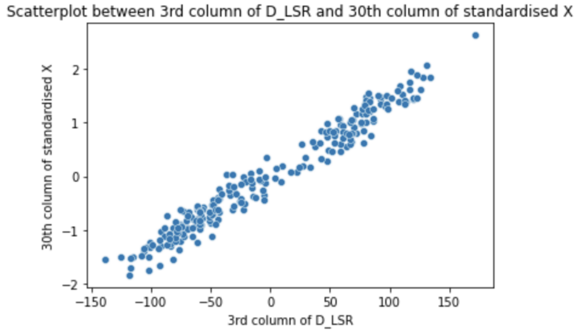

Figure 13: Scatterplot between 3rd column of Dlsr and 30th column of X

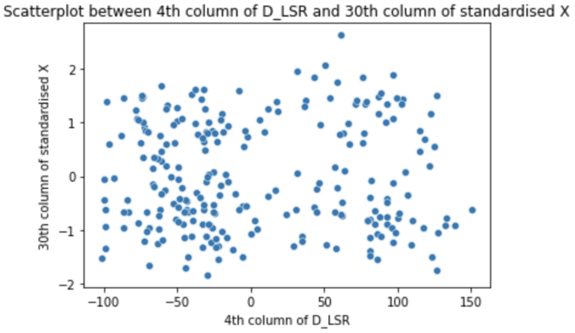

Figure 14: Scatterplot between 4th column of Dlsr and 30th column of X

2.2 Question 2.2

Ridge Regression (RR) parameters $\mathbf{A}_{RR} = (\mathbf{D}^T\mathbf{D}+\widetilde{\lambda}\mathbf{I})^{-1}\mathbf{D}^T\mathbf{X}$ and $\mathbf{D}_{RR} = \mathbf{X}\mathbf{A}_{RR}^T$ were estimated and subsequently compared to the corresponding LSR parameters. Two correlation vectors are estimated and the maximum absolute correlations between each $\mathbf{TC}$ and $\mathbf{D}_{LSR}$ as well as between each $\mathbf{TC}$ and $\mathbf{D}_{RR}$ are retained for the comparison. The maximum absolute correlations between each $\mathbf{TC}$ and $\mathbf{D}_{LSR}$ will be stored in $\mathbf{c}_{TRR}$ whereas the maximum absolute correlations between each $\mathbf{TC}$ and $\mathbf{D}_{RR}$ will be stored in $\mathbf{c}_{TLSR}$. In order to achieve $\sum \mathbf{c}_{TRR}$ is greater than $\sum \mathbf{c}_{TLSR}$, the value of λ chosen is 0.6. In Figure 15, a plot of the first vector from $\mathbf{A}_{RR}$ and the first vector from $\mathbf{A}_{LSR}$ is generated. We can clearly see that all the values in $\mathbf{a}_{rr}^1$ shrink towards zero.

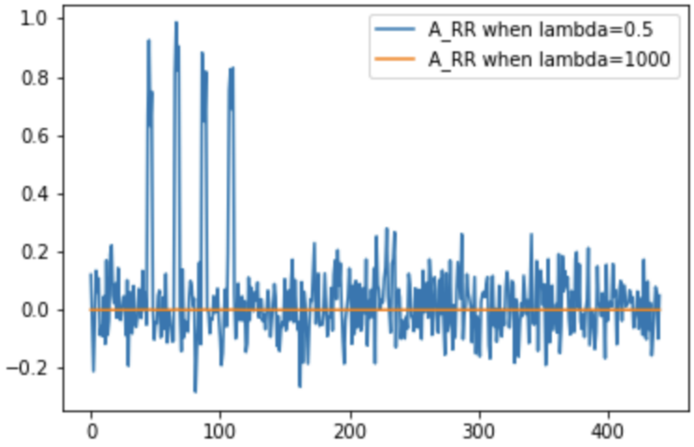

Figure 15: ARR vs ALSR

2.3 Question 2.3

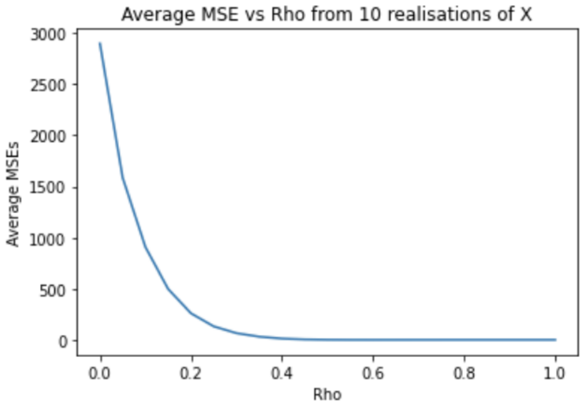

Figure 16: Average MSE vs. rho from 10 realisations of X

A list of 21 ρ values between 0 and 1 with 0.05 increments was created. Using Lasso Regression (LR) parameters as $\sum_{v=1}^{V}||\mathbf{X}-\mathbf{D}_{LR}\mathbf{A}_{LR}^2||/NV$, the $\mathbf{A}_{LR}$, $\mathbf{D}_{LR}$ and the sum of mean square error (MSE) values are estimated. This process is repeated 10 times which indicate 10 realisations, each time with a new standardized $\mathbf{X}$ that is in terms of $\mathbf{\Gamma}_t$ and $\mathbf{\Gamma}_s$. A graph of the average MSEs over the 10 realisations against each ρ value is then plotted, as illustrated in Figure 16. The minimum MSE was found when the rho value is 0.6. This value is appropriate as it is an unbiased estimator. The MSE value started to increase again when the rho value is approximately 0.3.

2.4 Question 2.4

Using $\rho = 0.6$ (selected value from Question 2.3), four correlation vectors are estimated and the maximum absolute correlations retained are as follows:

- between each **TC** and $\mathbf{D}_{RR}$ which will be stored in $\mathbf{c}_{TRR}$
- between each **SM** and $\mathbf{A}_{RR}$ which will be stored in $\mathbf{c}_{SRR}$
- between each **TC** and $\mathbf{D}_{LR}$ which will be stored in $\mathbf{c}_{TLR}$
- between each **SM** and $\mathbf{A}_{LR}$ which will be stored in $\mathbf{c}_{SLR}$

The sum of these four correlation vectors are then computed. $\sum \mathbf{c}_{TLR} > \sum \mathbf{c}_{TRR}$ and $\sum \mathbf{c}_{SLR} > \sum \mathbf{c}_{SRR}$ are achieved with $\rho = 0.6$. These four correlation vectors are plotted side by side in order to visualise the differences between each of them, as shown in Figure 17. It is important to note the large difference between the estimates $\mathbf{A}_{RR}$ and $\mathbf{A}_{LR}$ where false positives can clearly be seen. The false positives between the estimates of A can be attributed to the presence of noise in the data.

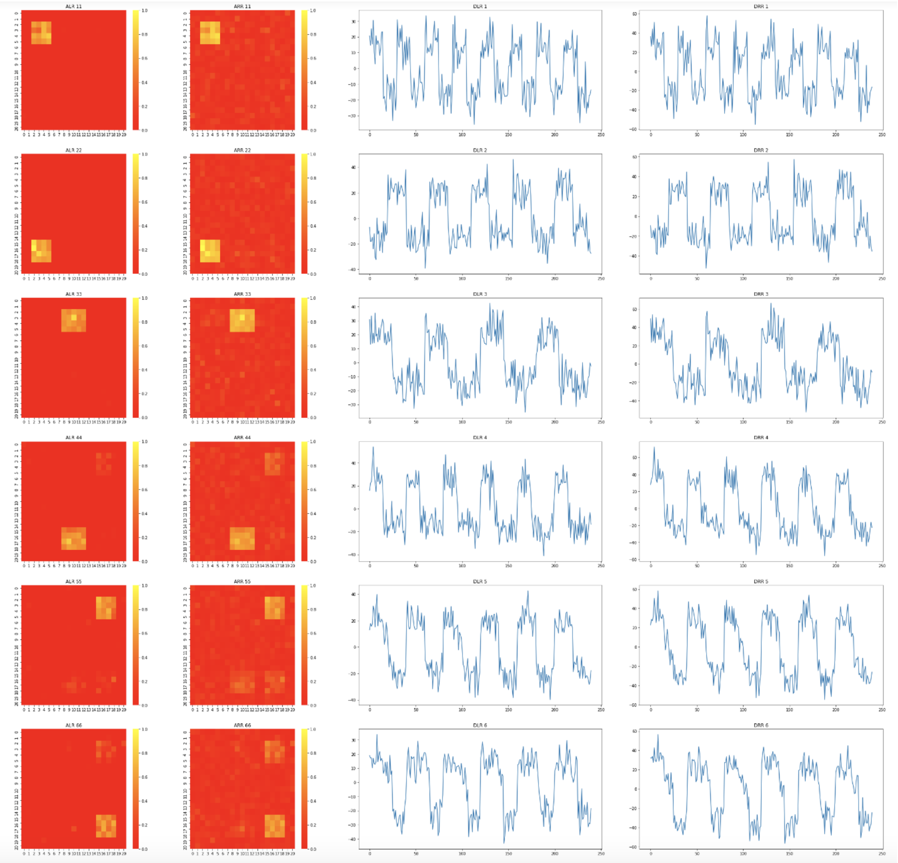

Figure 17: Estimates for ALR, ARR, DLR and DRR

2.5 Question 2.5

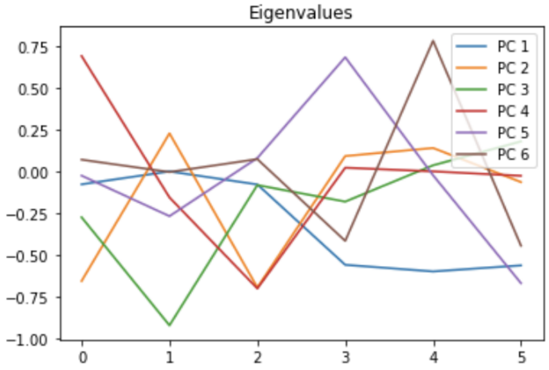

Figure 18: Eigenvalues for the PCs

Figure 18 shows a plot of the eigenvalues of principal components (PC). The 3rd PC was found to have the smallest eigenvalue with a value of -0.923. Subsequently, **Z** is computed by multiplying the **TC**s and the eigenvectors (u) obtained from Singular Value Decomposition (SVD). The regressors in **Z** and source TCs are then plotted side by side in order to visualise the differences, as shown in Figure 19. From the plots, we can see that the shape of the PCs have deteriorated and this can could be due to dimensionality reduction as the usage of PCR decreases the number of effective parameters in the model. Using $\rho = 0.001$, Figure 20 displays the results of the vectors $\mathbf{D}_{PCR}$ and $\mathbf{A}_{PCR}$. From the plots in Figure 20, it can be seen that Principal Component Regression (PCR) has a considerably low performance rate as compared to the other three regression models. This could be due to the simplicity of PCR. This method is able to generate results regardless of the nature of the input data or its underlying statistics which is the PCR method's main limitation.

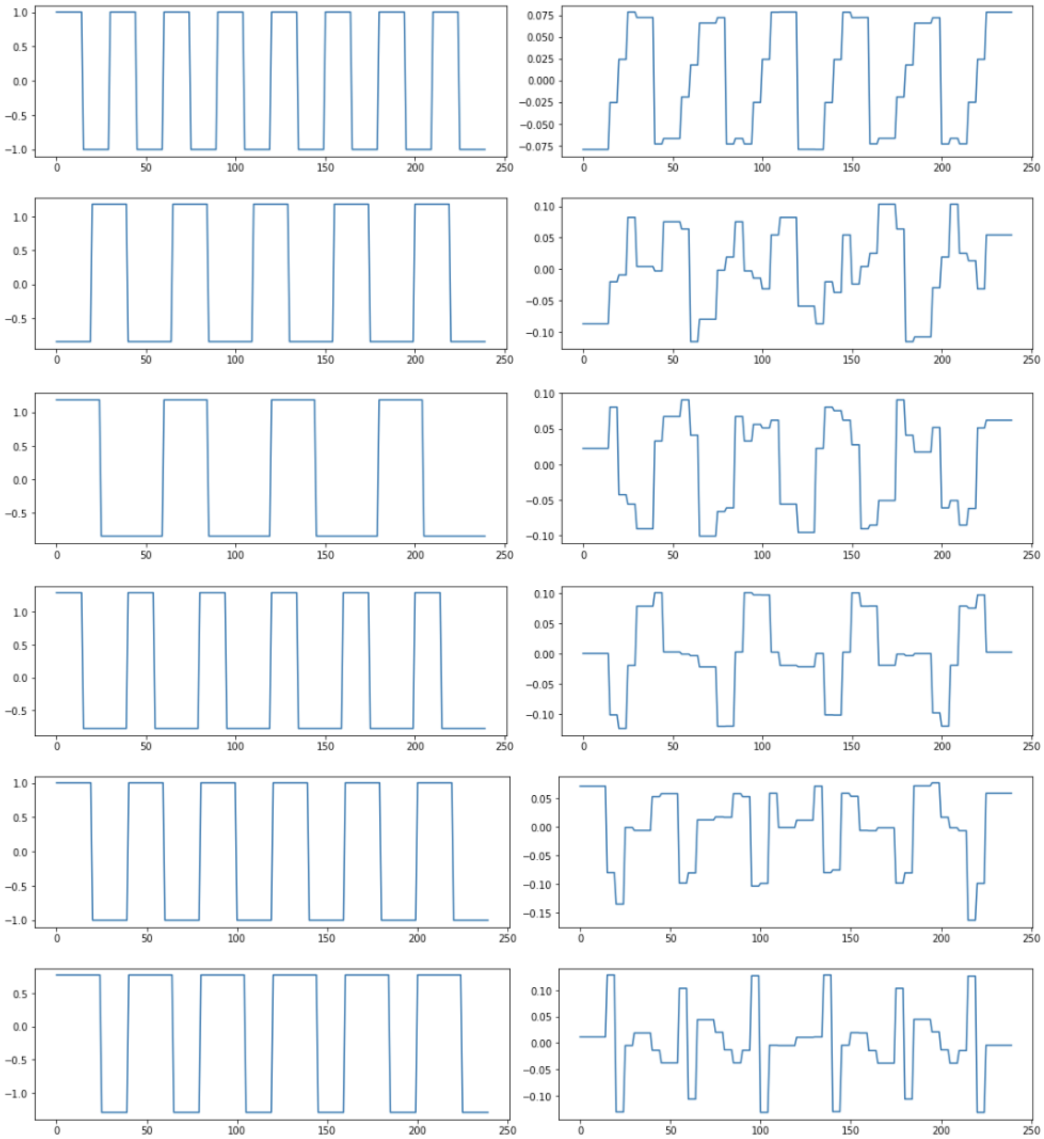

Figure 19: Regressors of Z and source TCs

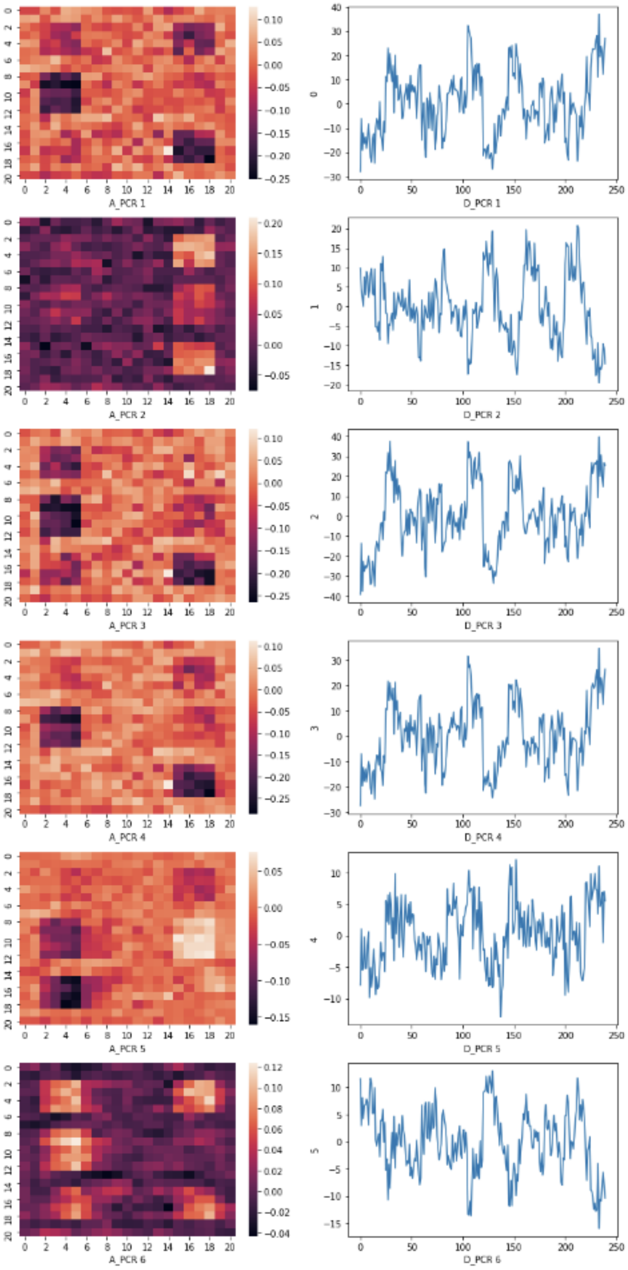

Figure 20: APCR vs. DPCR