MAST30034 Assignment 1

Zhi Hern Tom
School of Mathematics and Statistics
The University of Melbourne
Student ID: 1068268

September 10, 2021

Question 1: Synthetic dataset generation, data preprocessing, & data visualization

Q1.1)

A matrix TC of size 240×6 consisting of six temporal sources using onsets arrival vector, increment vector, and duration of ones was constructed as shown in Figure 1. The data was assumed to be Gaussian distributed as we will be implementing linear regression analysis. Also, if we normalized the data, the variance of the data will become much smaller. Thus, we standardized the TCs.

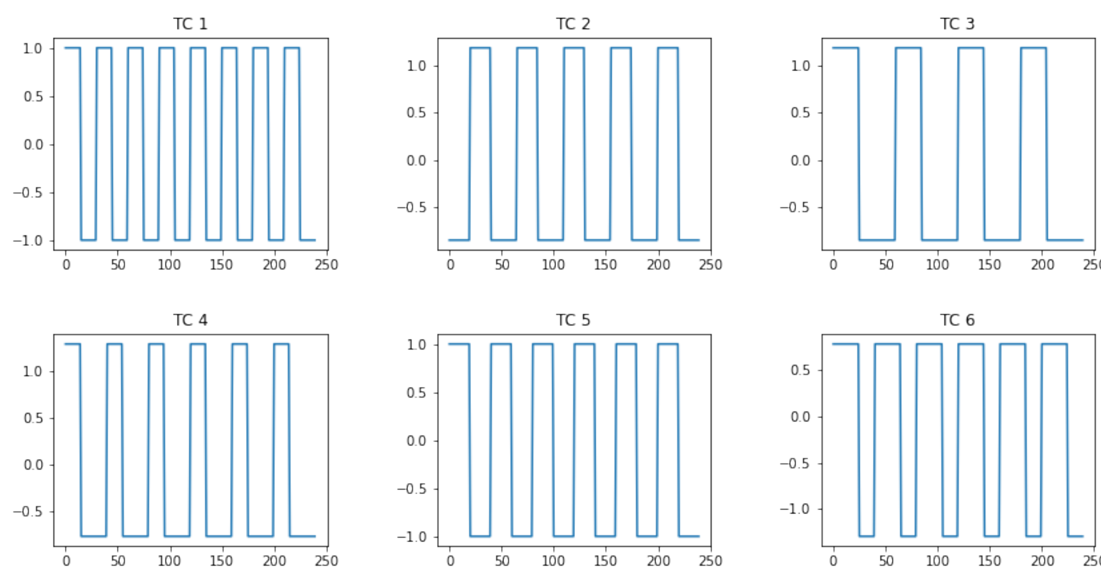


Figure 1: TCs

Q1.2)

A CM that represents correlation values between all 6 variables was constructed. In Figure 2, TC 4, TC 5 and TC 6 were highly positively correlated with each others. TC 4 has a positive correlation of 0.77 and 0.6 with TC 5 and TC 6 respectively while TC 5 has a correlation of 0.77 with TC 6.

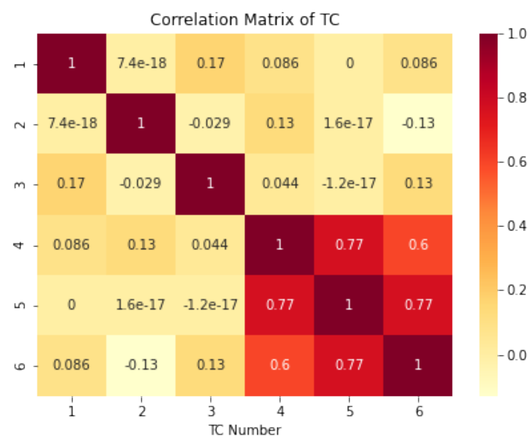


Figure 2: Correlation Matrix of TC

Q1.3)

In Figure 4, there is no correlation among all SMs. Standardization of SMs like TCs is not crucial as each vector has similar mean and standard deviation.

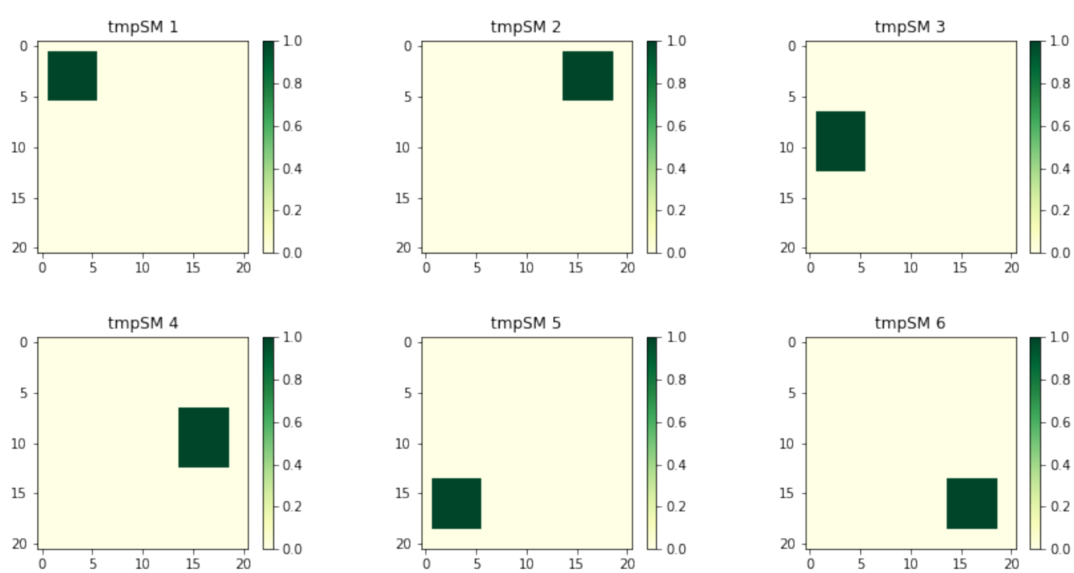


Figure 3: Image Plot of tmpSMs

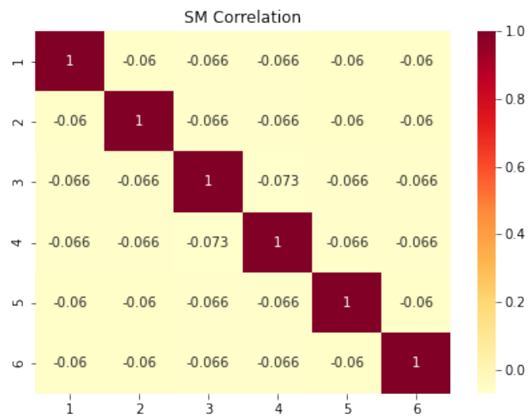


Figure 4: Correlation Matrix of SM

Q1.4)

In Figure 5, spatial and temporal noises are not correlated across sources. In Figure 6, both noise have a nice normal distribution curve as they are both sampled from normal distribution. The red line is the boundary of $\mu \pm 1.96\sigma$. Both normal distribution fulfil the mean and variance$= 1.96\sigma$ criteria relating to 0.25, 0.015, and zero mean. Also, according to Figure 7, it does not show that $\Gamma_t\Gamma_s$ has significant correlation across 441 variables as most of them have a relatively small correlation.

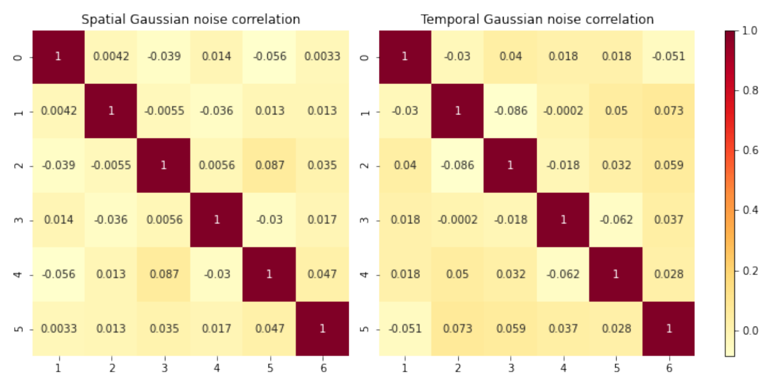


Figure 5: Spatial and Temporal Gaussian Noise Correlation Matrix

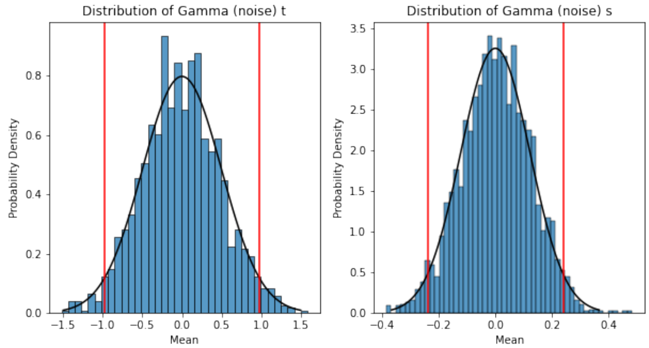


Figure 6: Distribution of Spatial and Temporal Noises

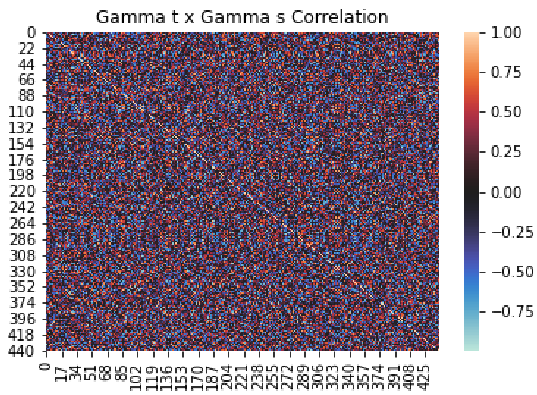


Figure 7: Correlation Matrix of Noises

Q1.5)

As we can see **(TC $\times$ SM)** is a linear combination of sources, **(Γt $\times$ Γs)** produces a structured noise. Second and third term will either produce structured noise or straight zeros on pixels with no values. Hence we can incorporate it into last term $\mathbf{E} = (\mathbf{TC} \times \Gamma\mathbf{s}) + (\Gamma\mathbf{t} \times \mathbf{SM}) + (\Gamma\mathbf{t} \times \Gamma\mathbf{s})$ to simplify the model. Apart from that, Figure 9 shows 2 clusters of variance, which are around 0 and 1.5. The reason behind this is that the noise sources were generated from 2 Gaussian distribution with both $\mu = 0$ and $\sigma^2 = 0.25$ and 0.015 respectively.

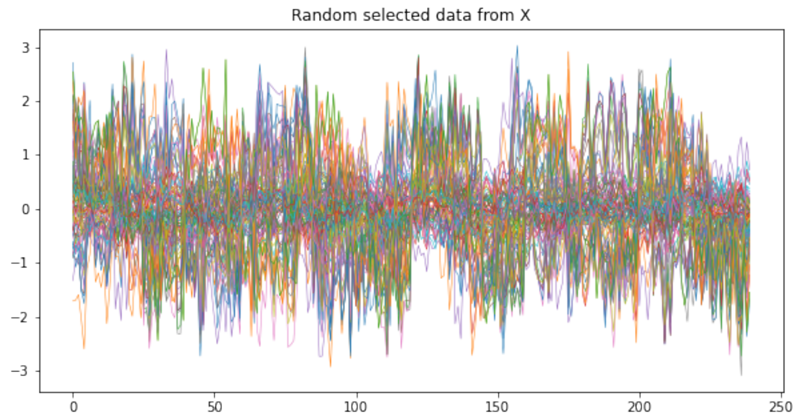


Figure 8: 100 Randomly Selected Time-series from X

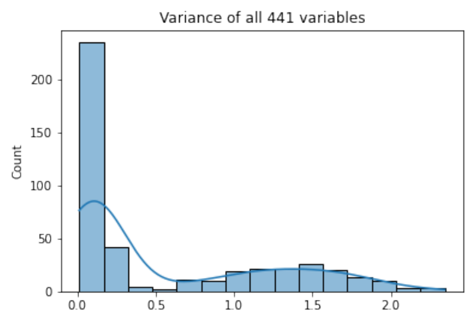


Figure 9: Variance of all 441 Variables

Question 2: Data analysis, results visualization, & performance metrics

Q2.1)

A scatter plot between 3^{rd} column of D_{LSR} and 30^{th} column of standardized X was constructed as shown in Figure 11. The 30^{th} pixel position is filled by the 3rd SM, thus the third TC is the only time course that constructs 30^{th} column of X. Also, it shows a distinct linear relationship between them. However, according to Figure 12, there is no significant linear relationship between 4^{th} column of D_{LSR} and 30^{th} column of X.

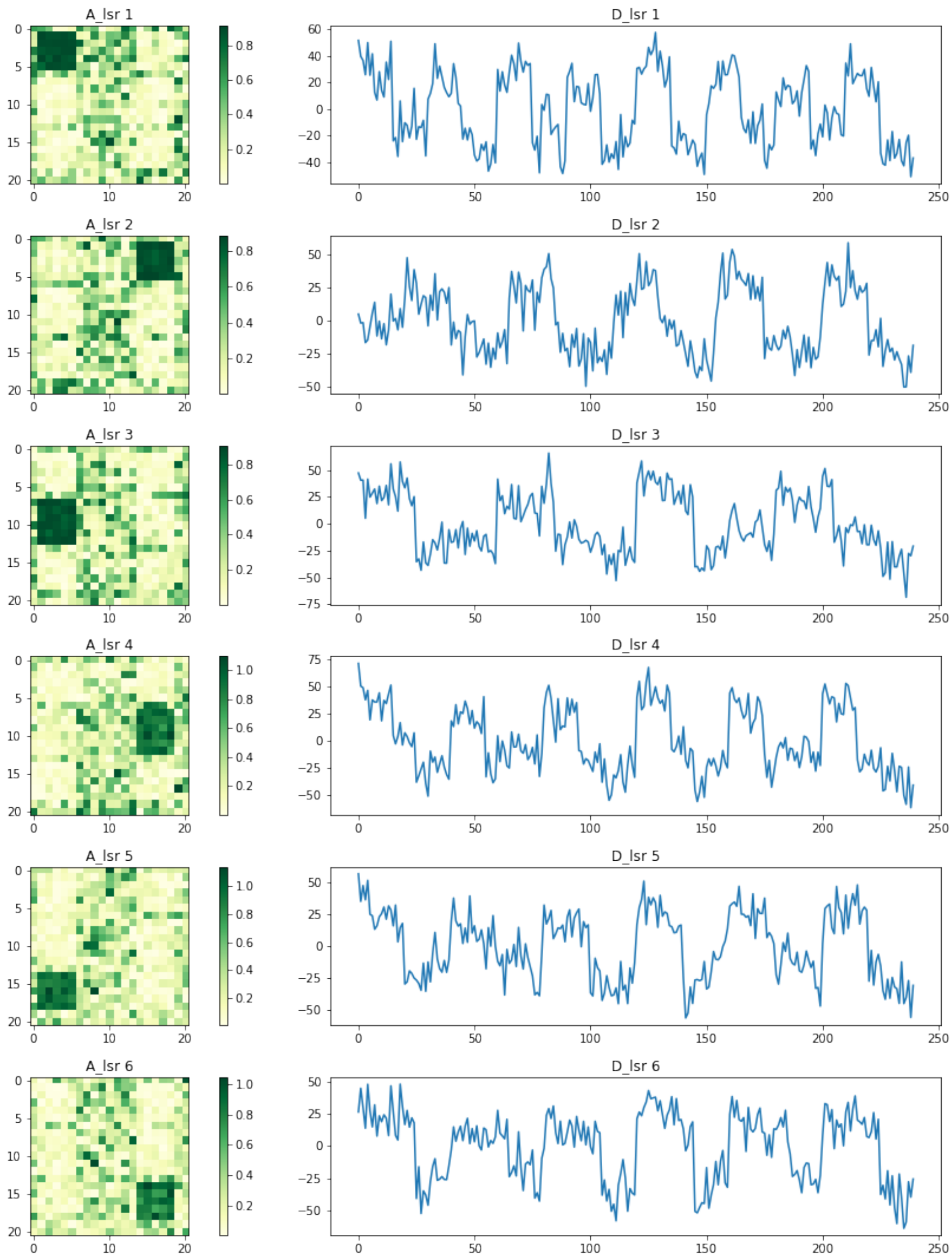


Figure 10: A_{LSR} (left) and D_{LSR} (right)

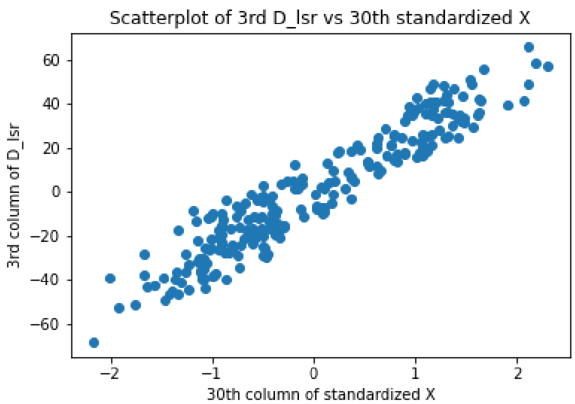


Figure 11: Scatterplot of 3^{rd} D_{LSR} vs 30^{th} standardised X

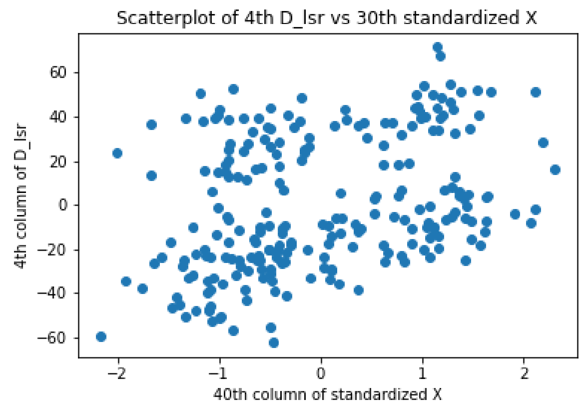


Figure 12: Scatterplot of 4^{th} D_{LSR} vs 30^{th} standardised X

Q2.2)

D_{RR} was estimated using $\lambda = 0.5$. The sum of correlation of c_{TLSR} and c_{TRR} have been calculated and plotted in Figure 13. In Figure 13, we can see that $\sum c_{TRR} = 5.42$ is slightly higher than $\sum c_{TLSR} = 5.11$. Futhermore, a^1_{RR} and a^1_{LSR} have been plotted as shown in Figure 14. According to Figure 14, we can see clearly that all values in a^1_{RR} shrink towards zero while a^1_{LSR} does not.

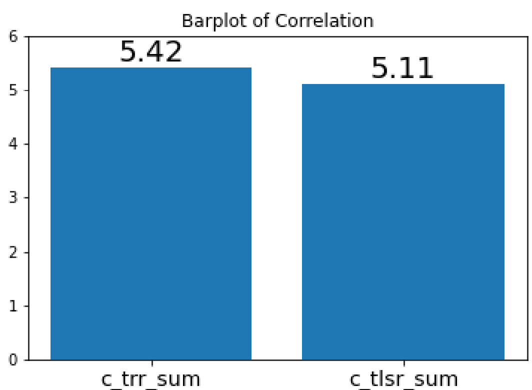


Figure 13: Barplot of c_{TLSR}_sum vs c_{TRR}_sum

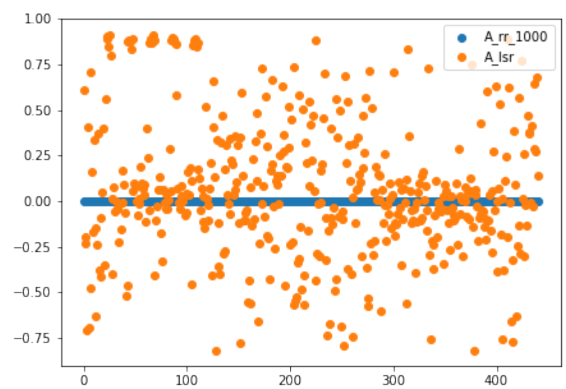


Figure 14: Scatterplot of $A1_{RR}$ (left) and $A1_{LSR}$ (right)

Q2.3)

LR parameters A_{LR}, D_{LR} were estimated using 21 values of ρ selected between 0 and 1 with an interval of 0.05. Then the average of MSE over 10 realizations against each value of ρ was plotted. In Figure 15, we can see that it reaches a minimum point at $\rho = 0.6$ and after that the MSE started to increase again and consequently converge at around $\rho = 0.9$. However, the interval of ρ is quite large, it would be better if we decrease the interval of ρ to get a better result.

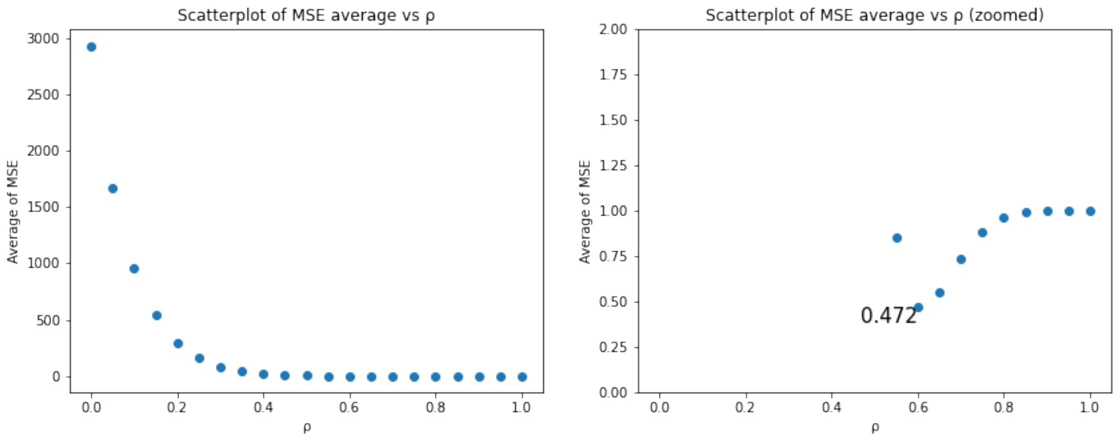


Figure 15: Scatterplot of MSE average vs rho

Q2.4)

LR parameters was estimated using $\rho = 0.6$ (as selected in Question 2.3). The sum of correlation vectors between

1. TC and D_{LR} (stored in c_{TRR})
2. SM and A_{RR} (stored in c_{SRR}), and
3. SM and A_{LR} (stored in c_{SLR})

were calculated and plotted using barplot. In Figure 16, it can be shown that $\sum c_{TLR}$ is higher than $\sum c_{TRR}$ and $\sum c_{SLR}$ is higher than $\sum c_{SRR}$. In Figure 17, 4 columns estimates of D and A for both RR and LR have been plotted. It shows that A_{LR} has much less false positives than A_{LSS}. It might be the reason that the Lasso Regression encourages coefficients shrink to zero, therefore pixels with zero value which contribute nothing to the model will remain zero. On the other hand, in Ridge Regression, constraint was put on the sum of squares of coefficients, which brings the value of coefficients close to zero. This is why A_{RR} is not sparsity and has more false positive term.

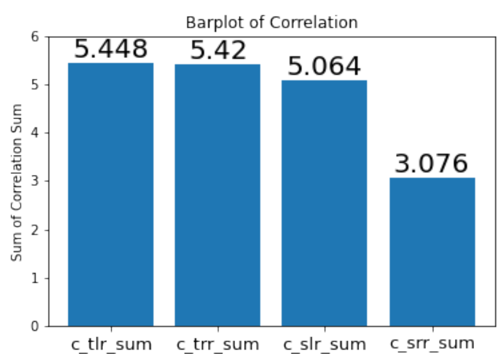


Figure 16: Barplot of Correlation Sum

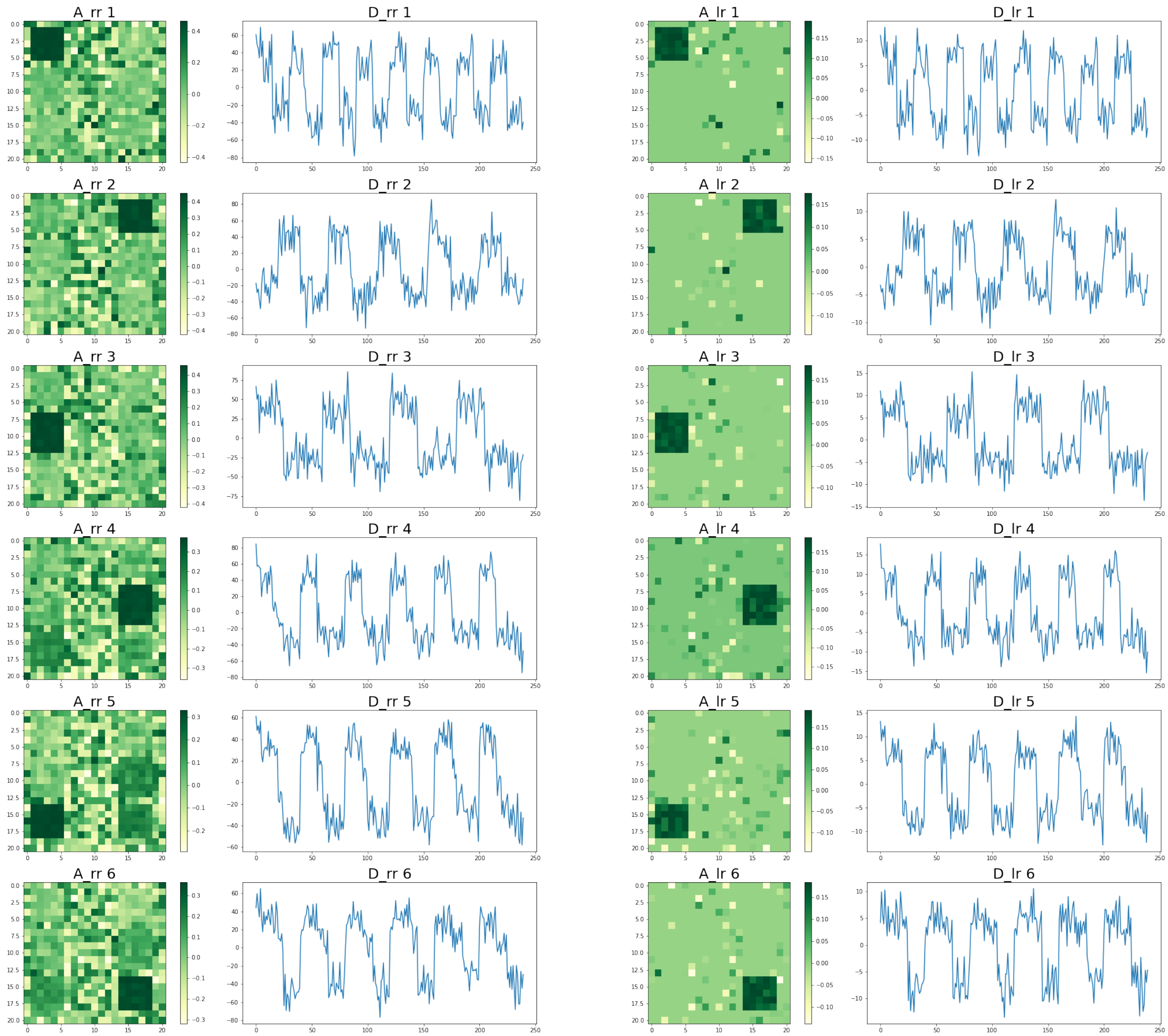
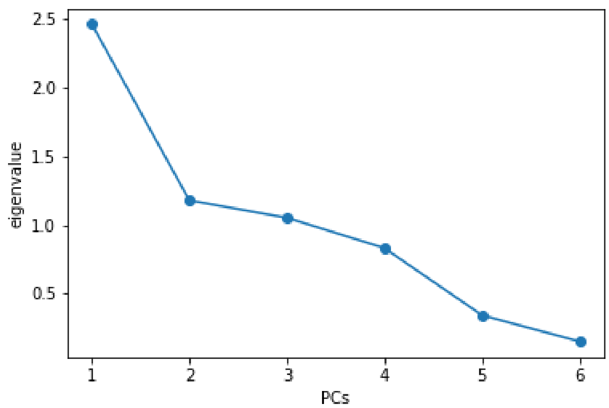


Figure 17: Plot of A_{RR} (1st), D_{RR} (2nd), A_{LR} (3rd) and D_{LR} (4th)

Q2.5)

Estimate PCs of the TCs were estimated and their eigenvalues were plotted as shown in Figure 18. In Figure 18, we can find that PC_{6th} has the smallest eigenvalue. Also, according to Figure 19, each of the Z (PC) was out of shape comparing to TCs. It was because the principal components of Z could not explained as much information as the original TC did since the dimension has been reduced. In Figure 20, we can see that the performance of PCR is worse than the other three regression models. The reason behind this is that both A_{PCR} and D_{PCR} were estimated using Z_{PC} which part of the information was deleted, hence resulting in inferior performance or PCR.



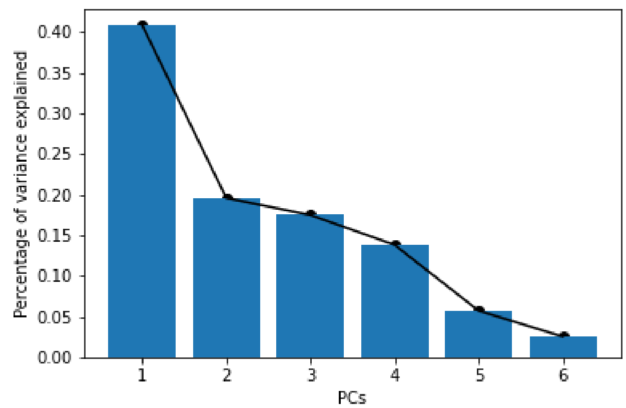


Figure 18: Eigenvalues of each PCs

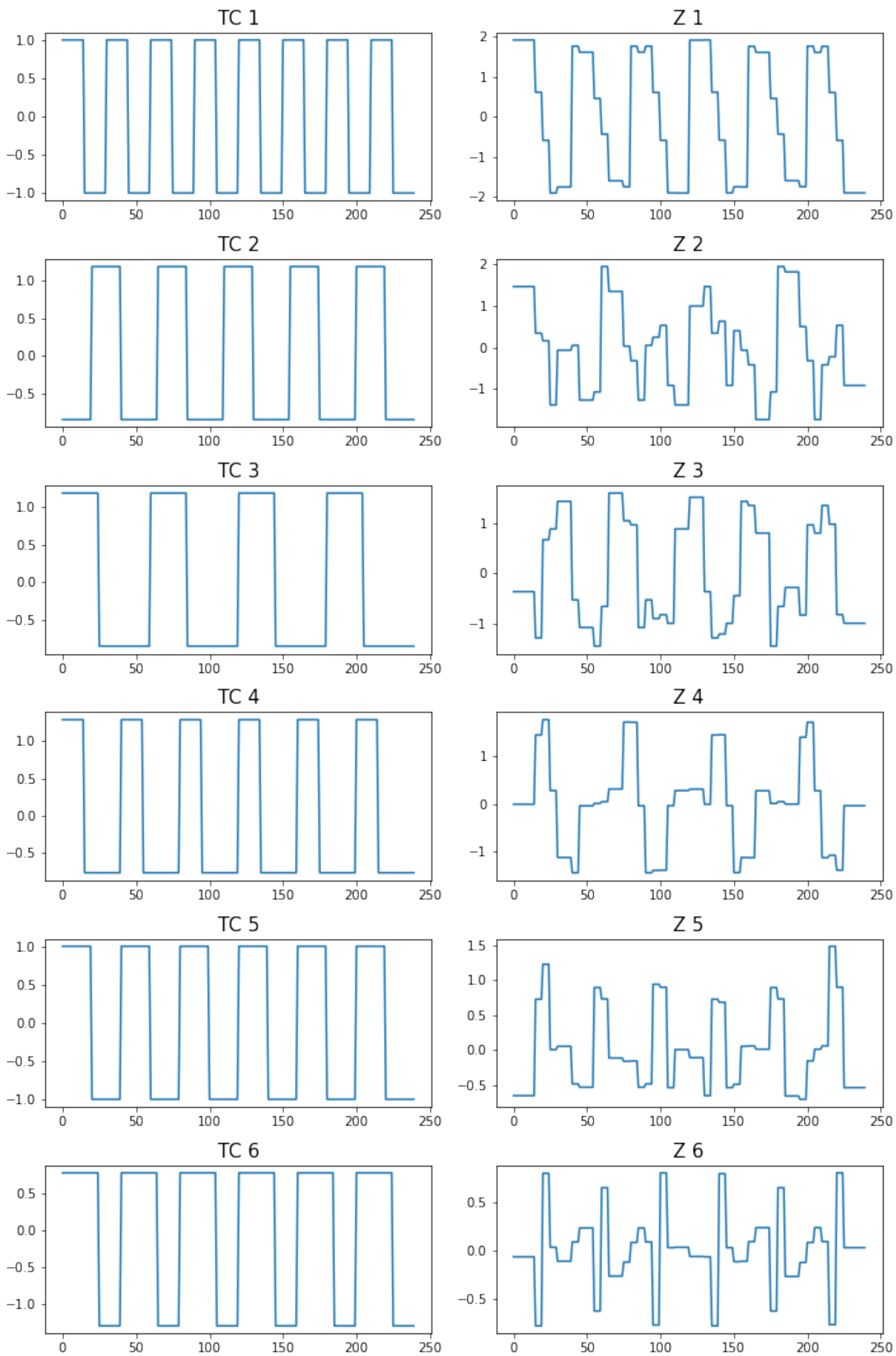


Figure 19: TCs and PCc

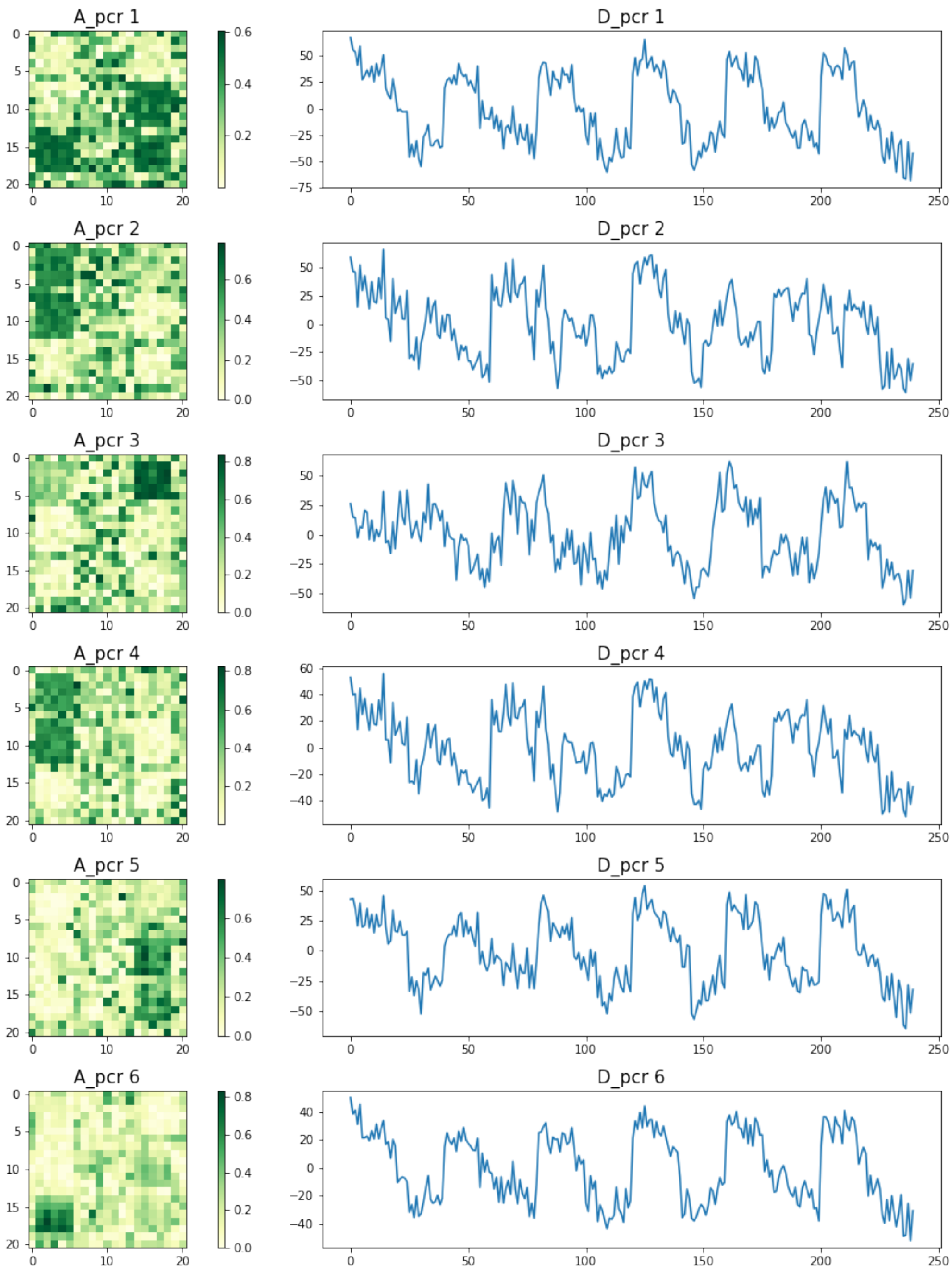


Figure 20: A_{PCR} (left) and D_{PCR} (right)