MAST30034

Applied Data Science Assignment 1

Juan Jesse Holiyanto
Student ID: 1001932

September 9, 2021

1 Question 1

1.1 Question 1.1

i) Plot all TCs as six subplots

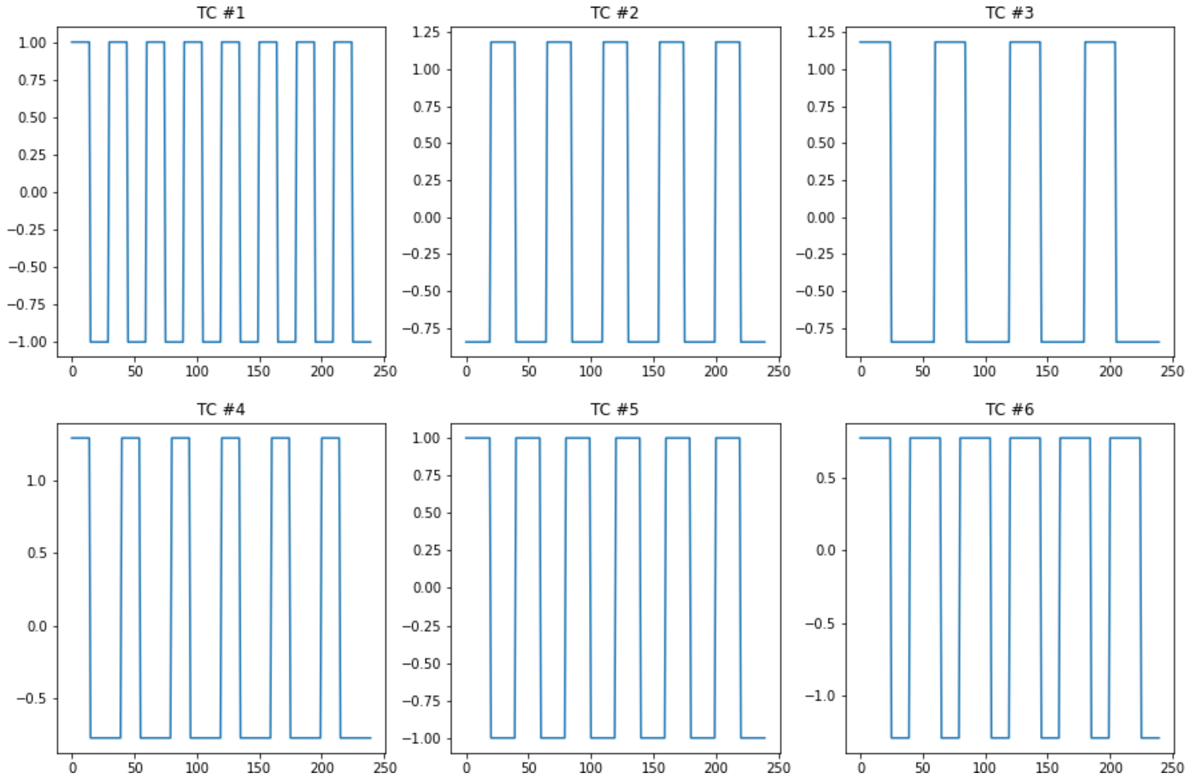


Figure 1: 6 TCs

ii) Why not normalize (divide by l-2 norm) the TCs instead of standardizing it?
If l2-norm normalization is used, the mean will not be centered around 0 and the solution will not be sparse. This is because the TC originally already has the range (0, 1). Thus, the resulting graph for the TCs will be the same. As such, normalization is not used

1.2 Question 1.2

i) Show its plot

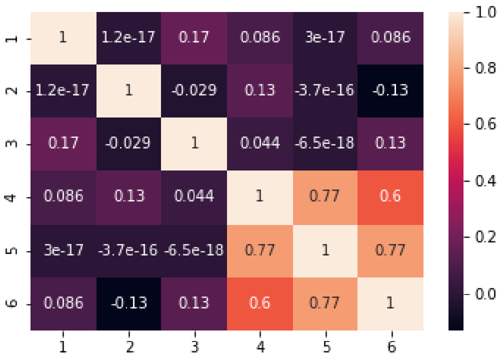


Figure 2: CM between 6 variables

ii) Can you tell visually which two TCs are highly correlated? If not, can you tell this from CM? We can see that TC 4, TC 5, and TC 6 have similar shape in 1.1. using the correlation matrix, we are able to see that TC 4 is correlated with TC 5 and 6. TC 5 IS correlated with TC 4 and TC 6. TC 6 is correlated with TC 4 and TC 5.

1.3 Question 1.3

i) Plot these SMs in six subplots

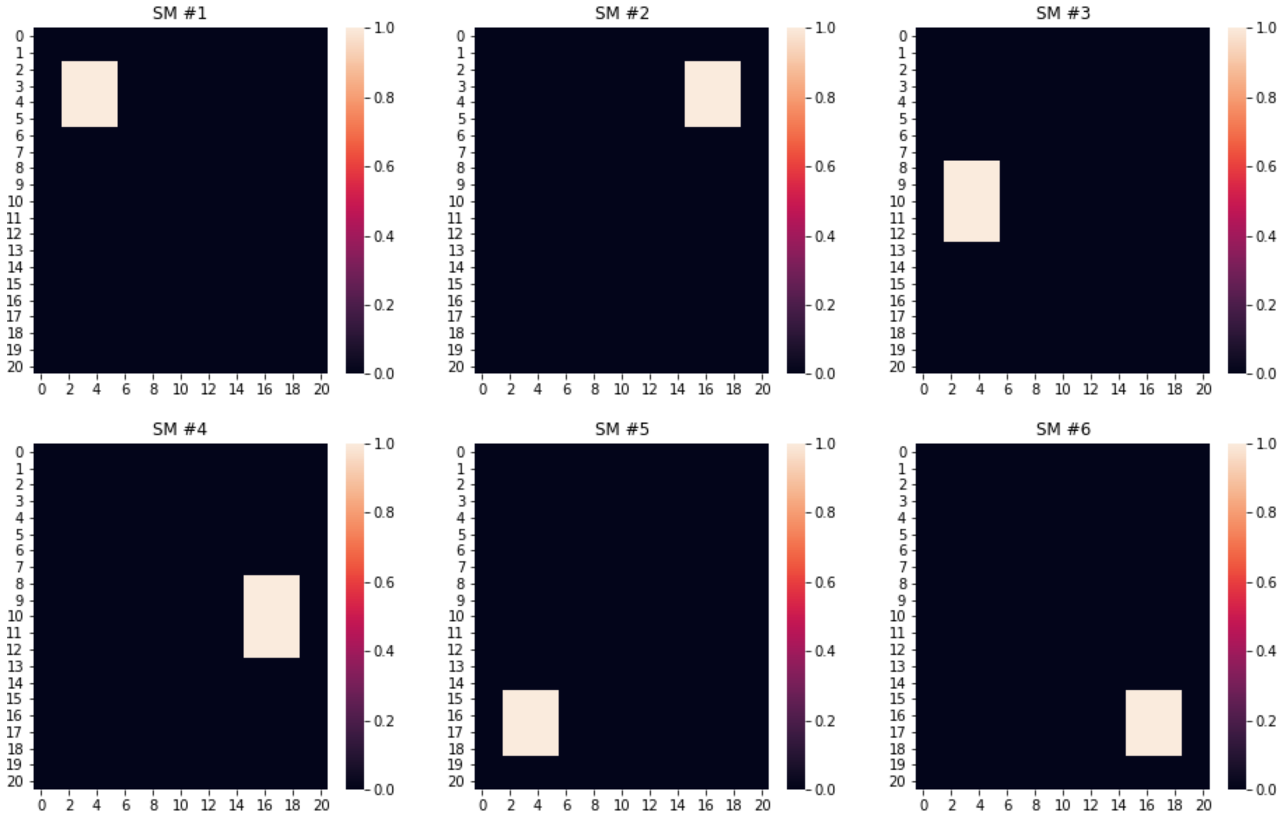


Figure 3: SMs in six subplots

ii) Using CM show if these 6 vectored SMs are independent

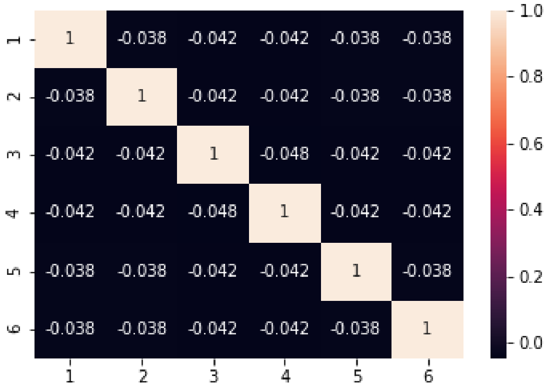


Figure 4: CM between SMs

The 6 vectors are not correlated, but we cannot fully say that they are independent as no correlation does not imply independence.

iii) For our particular case, why is standardization of SMs like TCs is not important?
Standardization is not needed as the value 1 represents the location of the SM. Furthermore, the mean and standard deviation of the SMs are similar, thus standardization is not needed.

1.4 Question 1.4

i) Using a 6 × 6 CM for each noise type (spatial and temporal) can you show if they are correlated across sources?

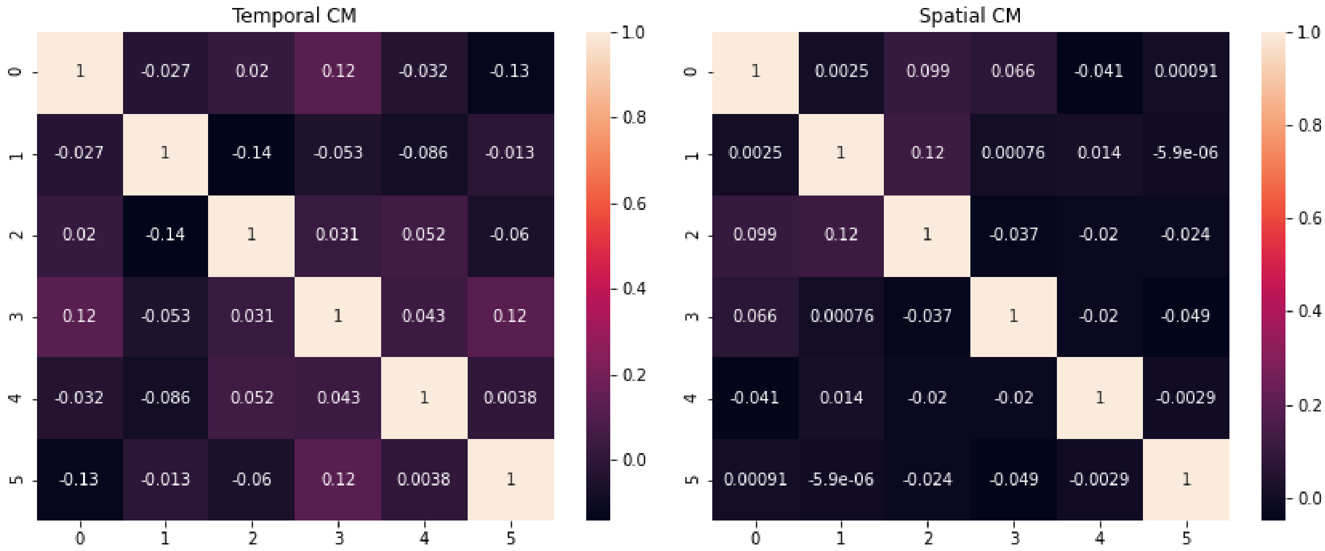


Figure 5: CM for each noise type

The sources within the noise are either not correlated or weakly correlated. However, we can't know if the 2 noises, temporal and spatial, are correlated by looking at them separately.

ii) Also plot the histogram of both noise sources to see if they have a normal distribution.

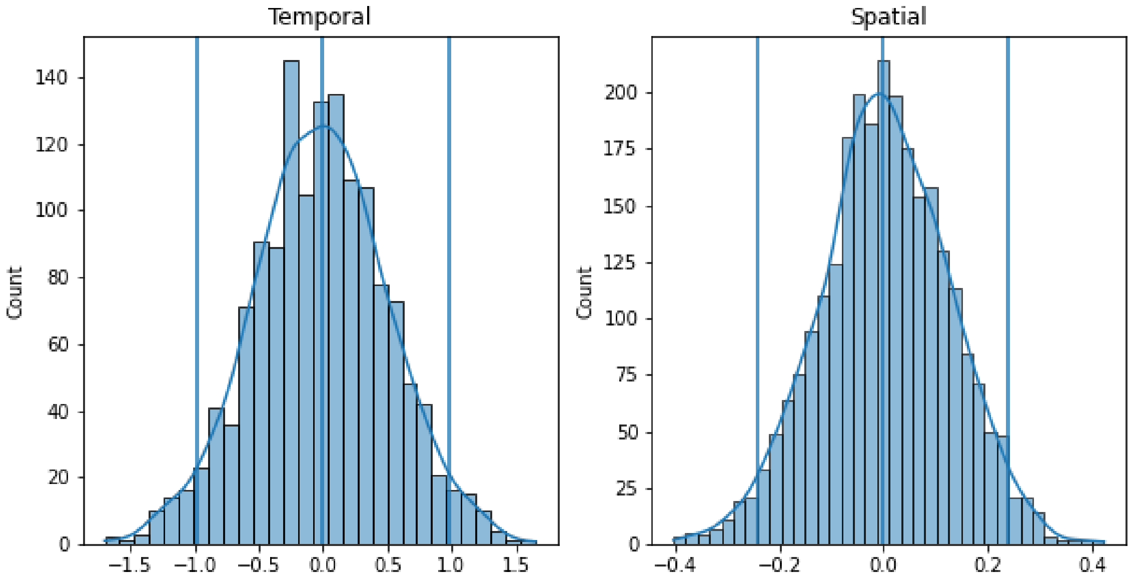
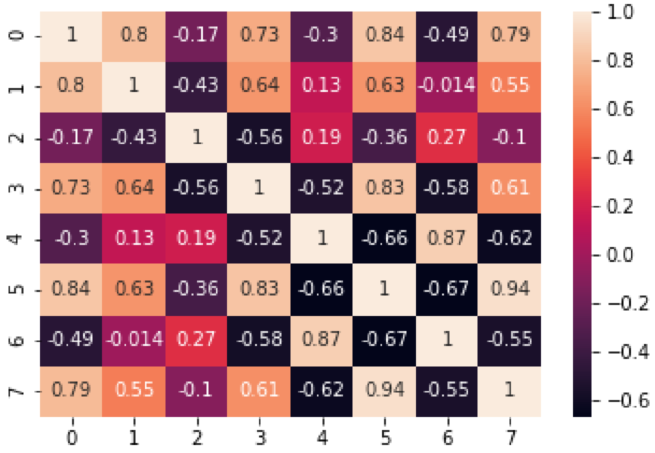


Figure 6: Histogram for each noise type

iii) Does this normal distribution fulfils the mean and variance=1.96σ criteria relating to 0.25, 0.015, and zero mean?
Yes, as the histograms follows the bell shaped curve, we can assume that both noise sources have a normal distribution. The normal distribution fulfills the mean and variance = 1.96σ criteria relating to 0.25, 0.015, and zero mean with a greater density in the middle.

iv) Is there product $\Gamma_t\Gamma_s$ correlated across V number of variables? As 441 variables is too much to map unto a CM, a small sample of 8 variable was taken instead. From the small sample, we can see that some variables from the product $\Gamma_t\Gamma_s$ are correlated across V number of variables.



	0	1	2	3	4	5	6	7
0	1	0.8	-0.17	0.73	-0.3	0.84	-0.49	0.79
1	0.8	1	-0.43	0.64	0.13	0.63	-0.014	0.55
2	-0.17	-0.43	1	-0.56	0.19	-0.36	0.27	-0.1
3	0.73	0.64	-0.56	1	-0.52	0.83	-0.58	0.61
4	-0.3	0.13	0.19	-0.52	1	-0.66	0.87	-0.62
5	0.84	0.63	-0.36	0.83	-0.66	1	-0.67	0.94
6	-0.49	-0.014	0.27	-0.58	0.87	-0.67	1	-0.55
7	0.79	0.55	-0.1	0.61	-0.62	0.94	-0.55	1

Figure 7: CM across V number of variables

1.5 Question 1.5

i) Can these products (TC × Γ_t) and (SM x Γ_s) exist, If yes what happened to them because if we keep them then we cannot fit our model onto (1)?

Yes the products of (TC $\times$ Γ_t) and (SM x Γ_s) exist as the shapes of the matrix match. They produce noise or straight zeros. As such, we can include them in the error term E in X = DA + E.
ii) Plot atleast 100 randomly selected time-series from X.

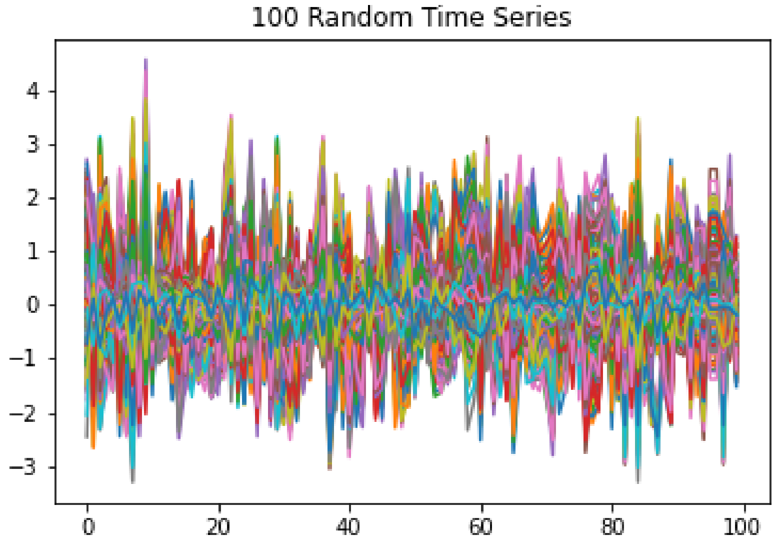


Figure 8: 100 Random Time Series

iii) Plot variance of all 441 variables on a separate plot. What information does this plot give you?

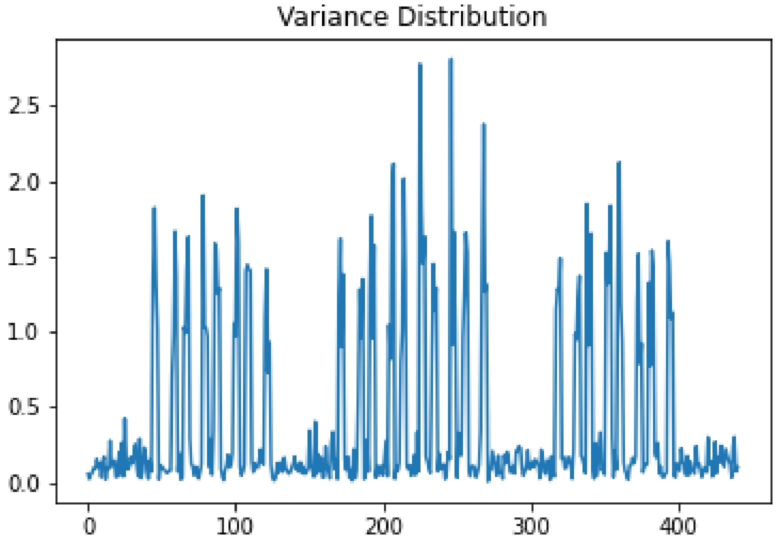


Figure 9: Variance distribution across 441 variables

The plot shows the variance distribution throughout the variables. There are multiple peaks, this denotes that there are multiple distributions used to generate X. We can also see that variance in X is inconsistent.

2 Question 2

2.1 Question 2.1

i) Plot six retrieved sources using A_{LSR} and D_{LSR} side by side

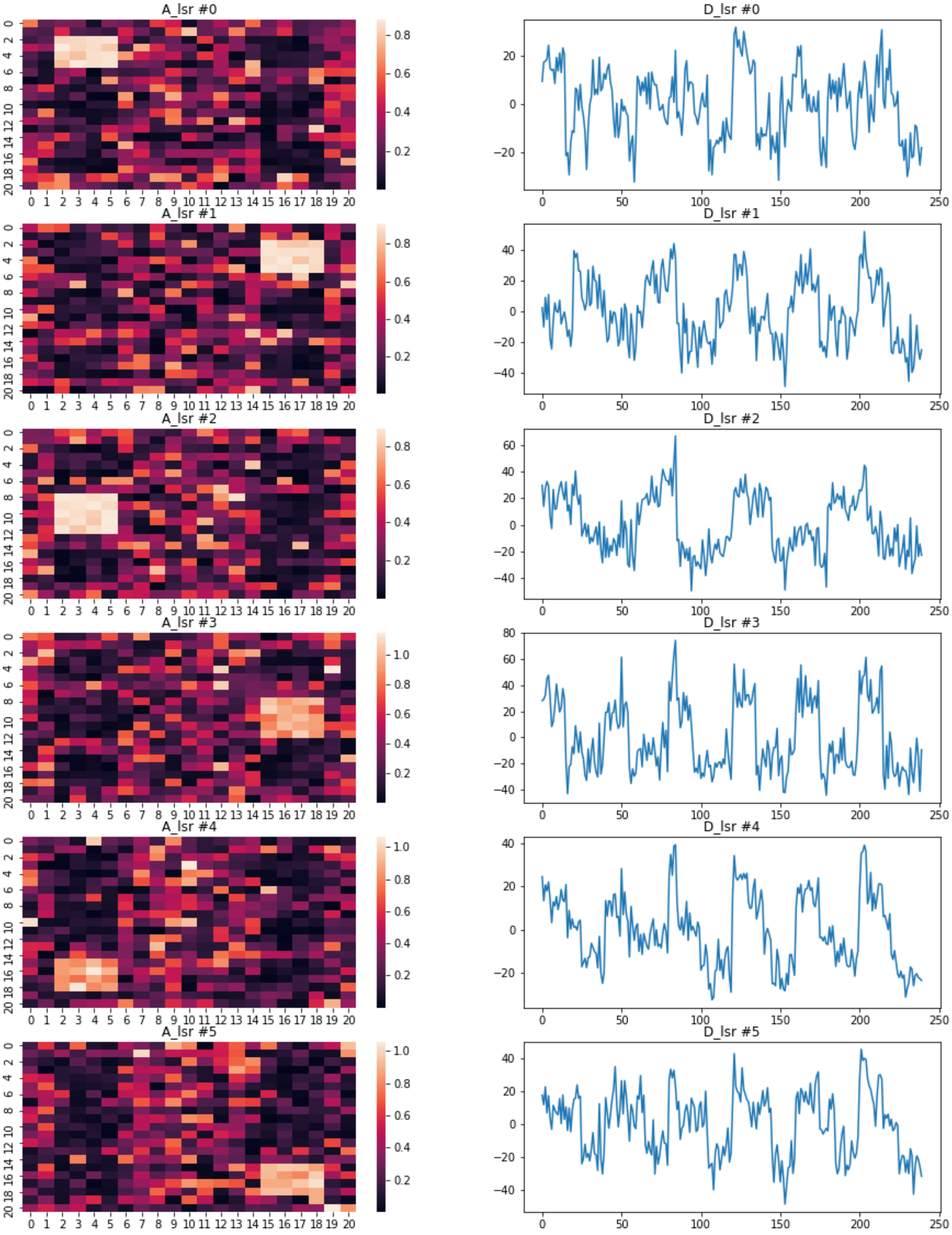


Figure 10: Six Sources of A_{LSR} and D_{LSR}

ii) 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?

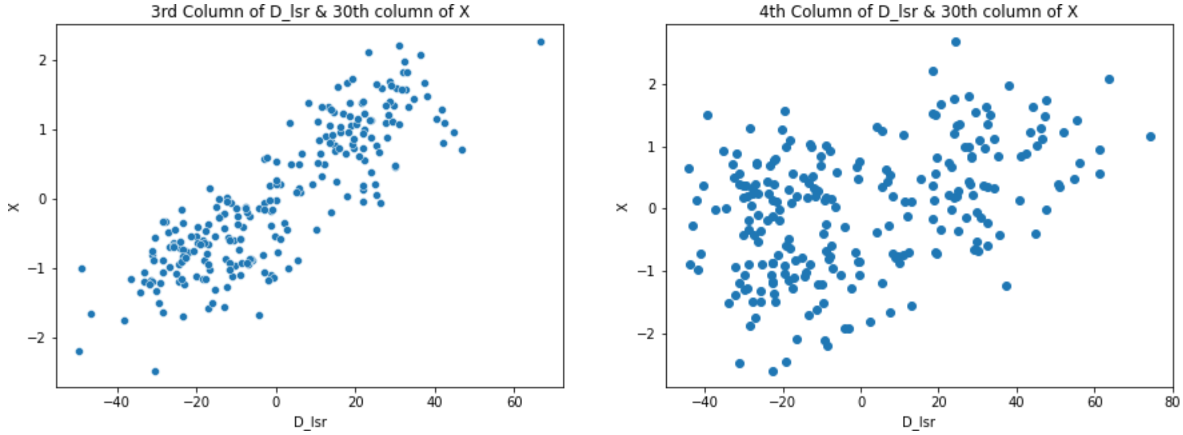
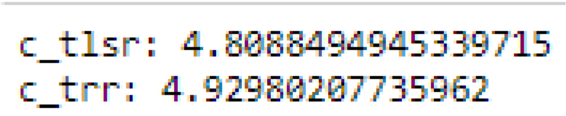


Figure 11: Scatter Plot

The 30th pixel position is filled by the 3rd SM. While the 30th pixel is not filled by the 4th SM. Thus, the third TC is the only time course that constructs 30th column of X. This 3rd TC is obtained from 3rd D_{LSR}. As such, there is no linear relationship from 4th column of D_{LSR} and the 30th pixel.

2.2 Question 2.2

i) Calculate the sum of these two correlation vectors (C_{TLSR} and C_{TRR}).



```
c_tlsr: 4.8088494945339715
c_trr: 4.92980207735962
```

Figure 12: Sum of two correlation vectors

ii) Also, for $\lambda = 1000$, 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?

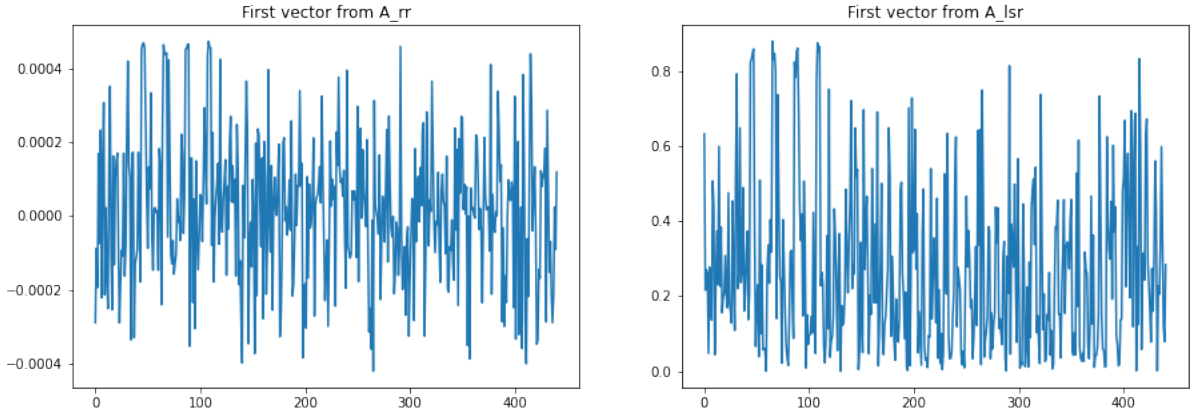


Figure 13: First Vector from A_{rr} and A_{lsr}

From the graph, we can see that the values in a^1_{rr} are indeed shrinking towards zero. This is because Ridge Regression shrink the coefficients of A towards zero when λ is large enough.

2.3 Question 2.3

i) 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)?

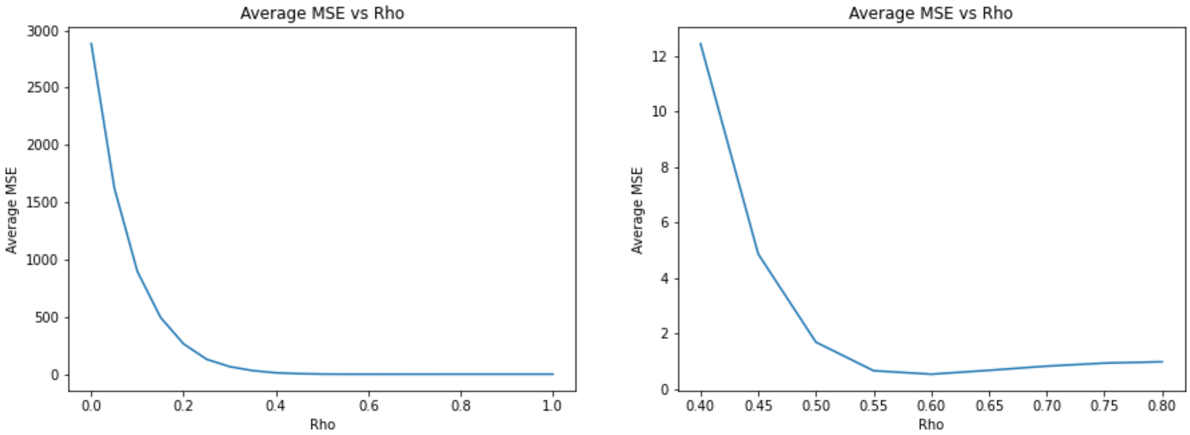
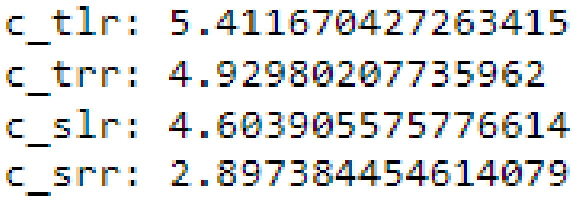


Figure 14: Average MSE over 10 realization

The minimum MSE is found when $\rho = 0.6$. In addition, ρ is obtained by averaging MSE over 10 different realisations with different sets of randomly generated noise. Thus, it is okay to select this value as we are aiming for the smallest MSE to yield the optimal result. MSE started to increase again when ρ is greater than 0.6

2.4 Question 2.4

i) Calculate the sum of these four correlation vectors. If you have carefully selected the value of ρ you must end up with $\sum C_{TLR} > \sum C_{TRR}$ and $\sum C_{SLR} > \sum C_{SRR}$



```
c_tlr: 5.411670427263415
c_trr: 4.92980207735962
c_slr: 4.603905575776614
c_srr: 2.897384454614079
```

Figure 15: Sum of four correlation vectors

ii) 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?

We can clearly see that there is significantly more false positives in A_{rr} in comparison to A_{lr}. Other than the inability of LSR to handle MC, the main reason behind the bad performance of LSR and RR (both of them producing many false positives while recovering coefficients) is that they incorporate the undesired (noise carrying) pixels into the estimate of A and this also effects the estimate of D in terms of overfitting. (Plot in next page)

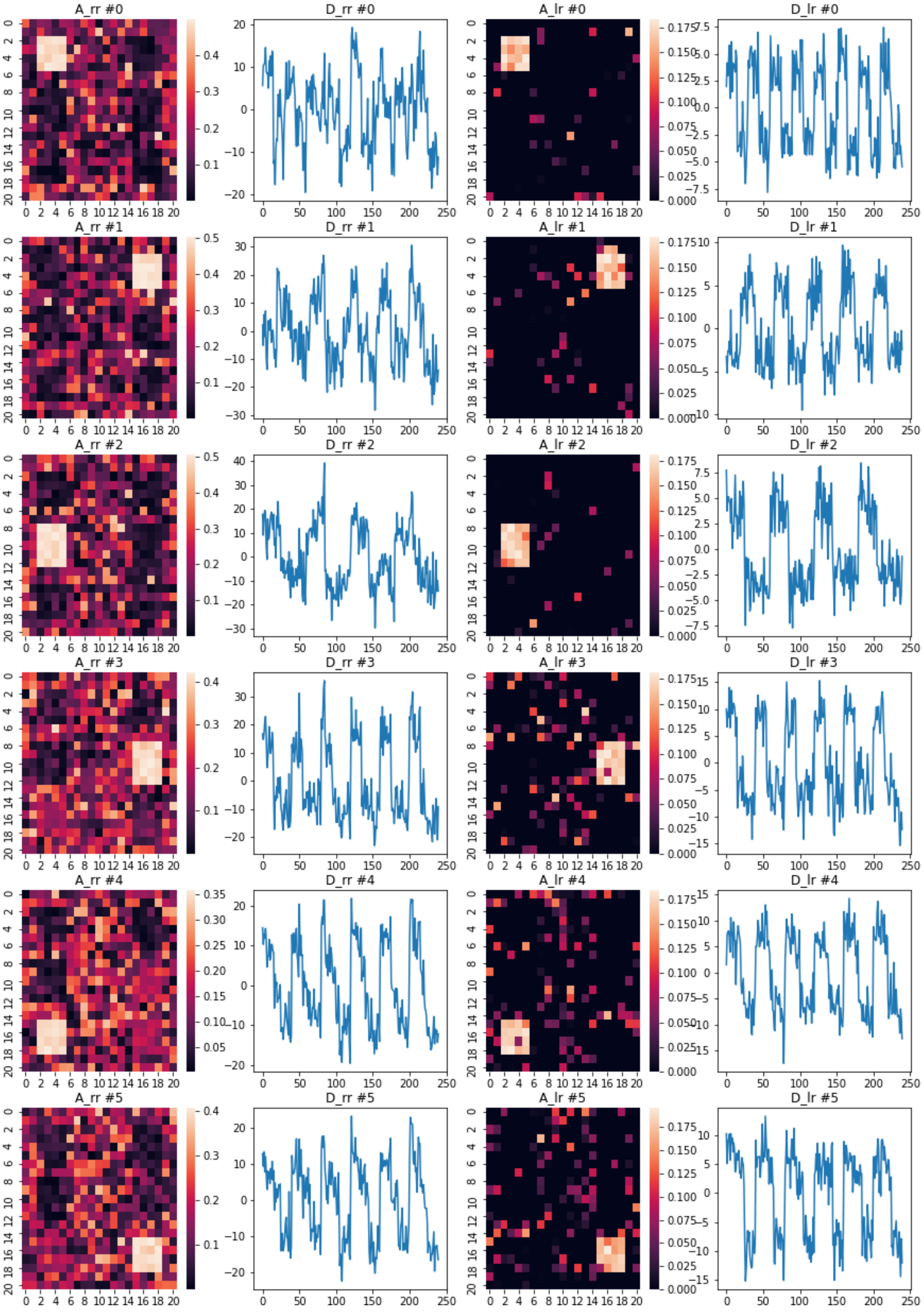


Figure 16: Estimates of D and A for RR and LR

2.5 Question 2.5

i) Estimate PCs of the TCs and plot their eigen values. For which PC the eigen value is the smallest?

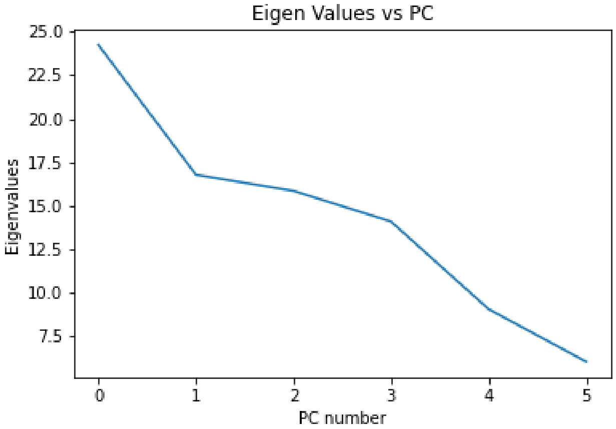


Figure 17: PC Estimates

Eigenvalue is least for the 6th PC.

ii) 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?

From the plot, we can see that the PCs are deteriorated. Yes, the PCs are deteriorated. The shape of TC has been lost because the PCs are actually a linear combination of the TCs. Furthermore the shapes of the TCs are lost as it has been projected to the direction of the loading vectors. Thus, not all variances of the ground truth are kept in the PC features. (Plot in next page)

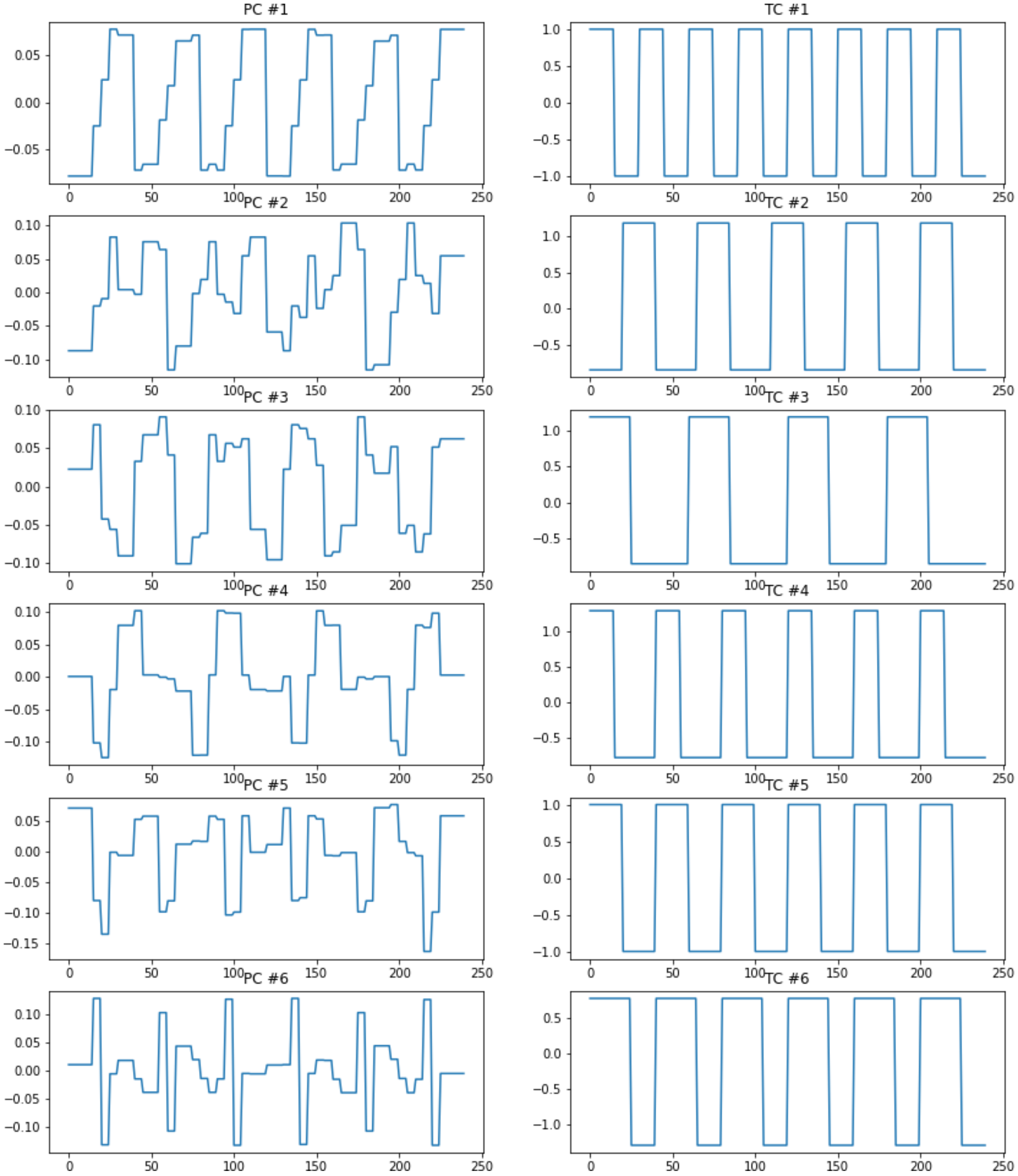
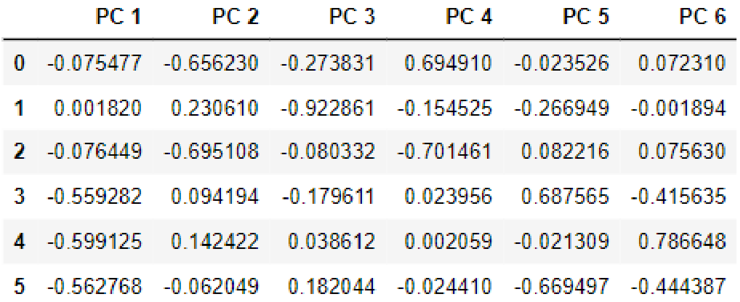


Figure 18: PC and TC side by side

iii) 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?

It can be seen that the PCR is inferior compared to the other three regression model. This is due to the fact that A is generated from Z (PCs from TC), a linear combination of the TCs.



	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
0	-0.075477	-0.656230	-0.273831	0.694910	-0.023526	0.072310
1	0.001820	0.230610	-0.922861	-0.154525	-0.266949	-0.001894
2	-0.076449	-0.695108	-0.080332	-0.701461	0.082216	0.075630
3	-0.559282	0.094194	-0.179611	0.023956	0.687565	-0.415635
4	-0.599125	0.142422	0.038612	0.002059	-0.021309	0.786648
5	-0.562768	-0.062049	0.182044	-0.024410	-0.669497	-0.444387

Figure 19: PC Results

The filled part (yellow part) of the first Retrieved Spatial Maps can be seen to come from SM 4, SM 5, and SM 6 (As seen on figure 3). This is because the first PC was utilised as the regressor, and the fourth, fifth, and sixth values in PC 1 are higher when absolute and compared to the other three values. In addition, the first A_{PCR} shows red blocks, corresponding to SM 1, SM 2, and SM 3. We can infer this as the first, second, and third coefficients of PC 1 as it is close to zero. (D_{PCR} and A_{PCR} plot in next page)

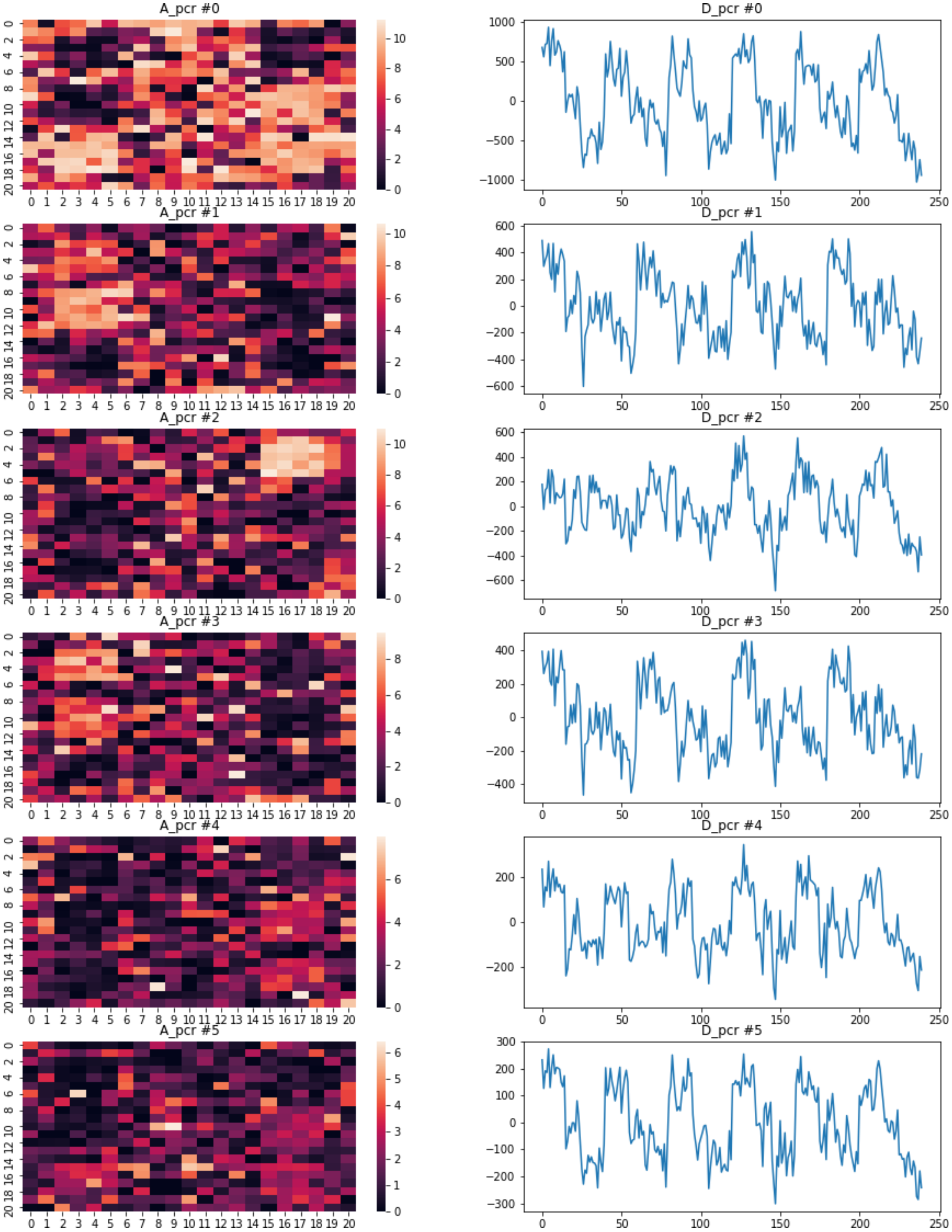


Figure 20: D_{PCR} and A_{PCR} side by side