MAST30034: Applied Data Science

Assignment 1

Haonan Zhong
Student ID: 867492

September 10, 2021

Question 1 Synthetic dataset generation, preprocessing & visualization

Question 1.1

Plot all TCs as six subplots

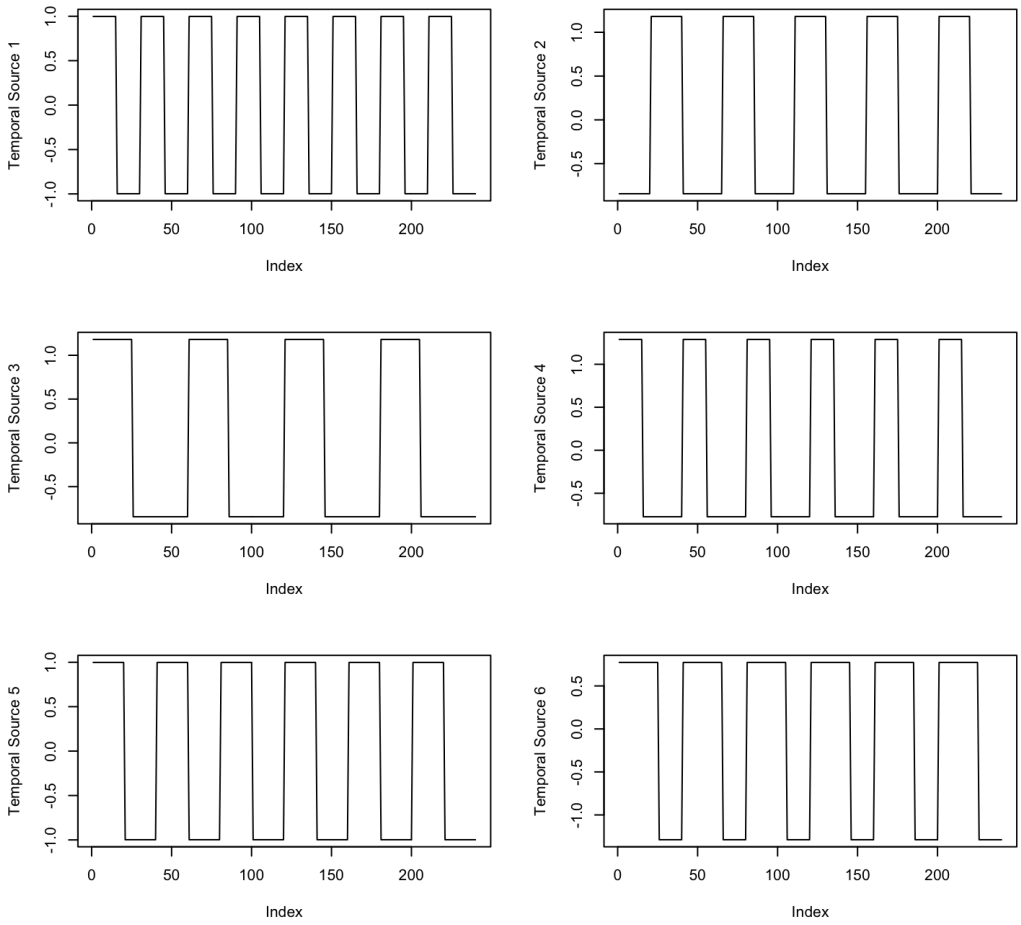

Figure 1: Standardised Temporal Sources

Why not normalize (divide by l-2 norm) the TCs instead of standardizing it?

If we normalize using L2-norm, the mean will not be centered around zero and the solution will not be sparse, given the TCs only contains 0 and 1. Hence, the resulting graph for the TCs will be the same.

Question 1.2

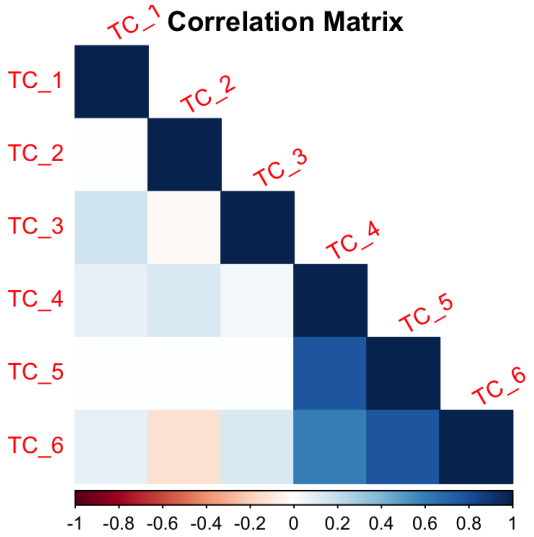

Figure 2: Correlation between 6 Temporal Sources

The correlation heatmap in Figure 2 shows that TC 4, TC 5, and TC 6 are highly correlated between each other, while other variables did not show significant correlation with each other.

Question 1.3

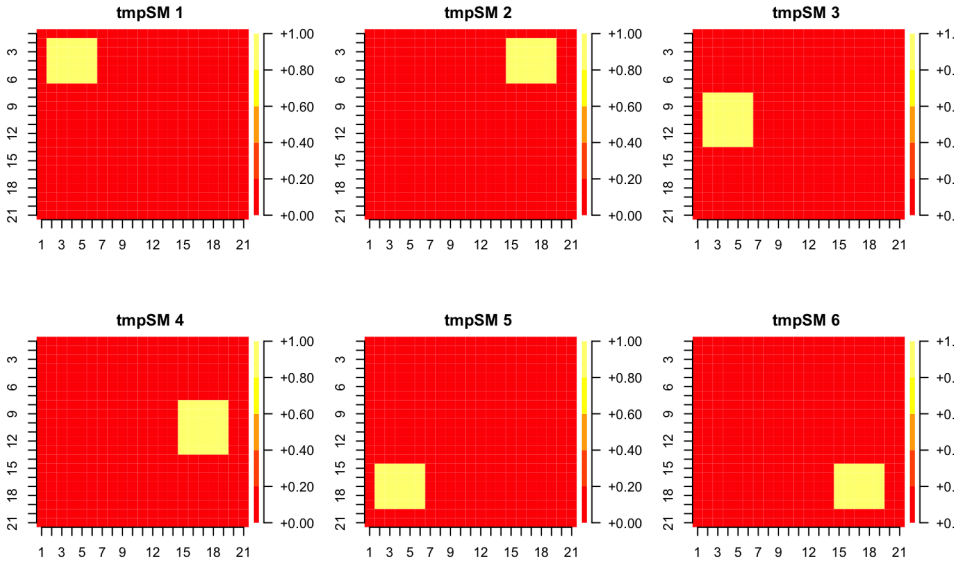

Figure 3: SM Sources

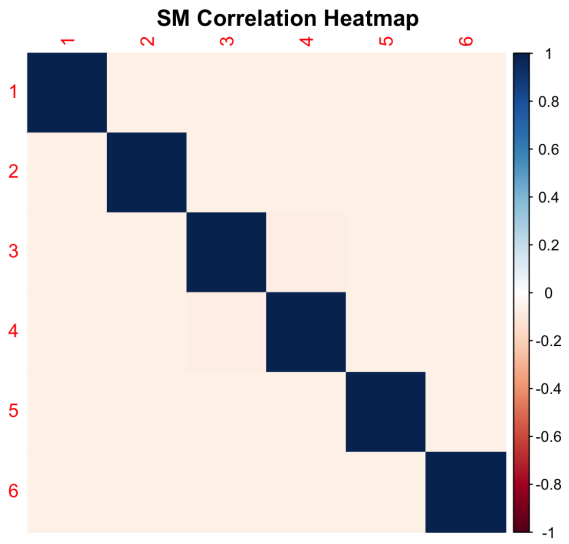

Figure 4: Correlation Between SMs

The correlation heatmap in Figure 4 suggests that all six vectored SMs are uncorrelated with each other. However, we cannot say that they are independent as no correlation does not imply independence. And standardization of SMs like TCs is not important because each vector has similar mean and standard deviation, thus standardization rendering is useless, we could directly compare them.

Question 1.4

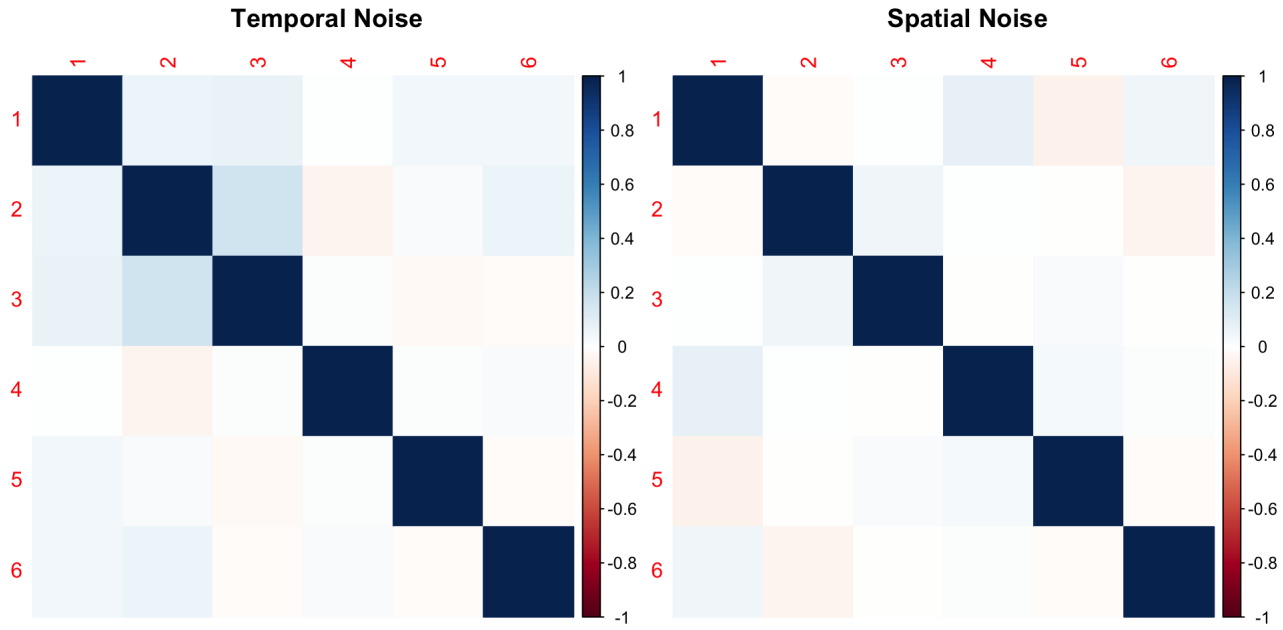

Figure 5: Correlation Matrix for each Noise Type

As we can see from Figure 5 above, no significant signs of correlation are shown across both sources.

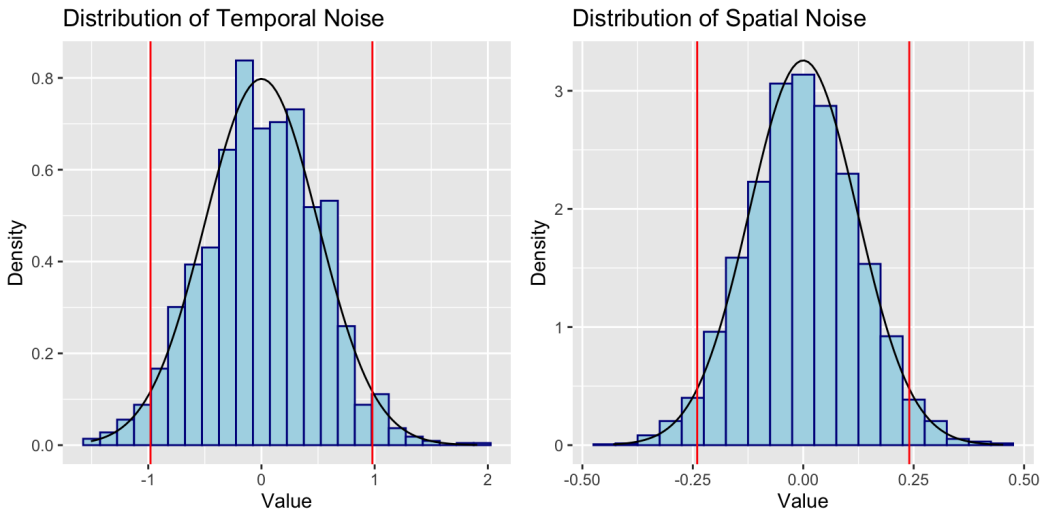

Figure 6: Distribution of Both Noise Sources

Both noise sources follows the normal distribution as the plots shown a symmetric bell shape curve. As we can see, distribution of $\mathcal{N}(0,\, 0.25)$ and $\mathcal{N}(0,\, 0.015)$ are also shown in the plots, respectively. These distribution of noises does fulfils the zero mean and variance $= 1.96\sigma$ criteria relating to 0.25, 0.015.

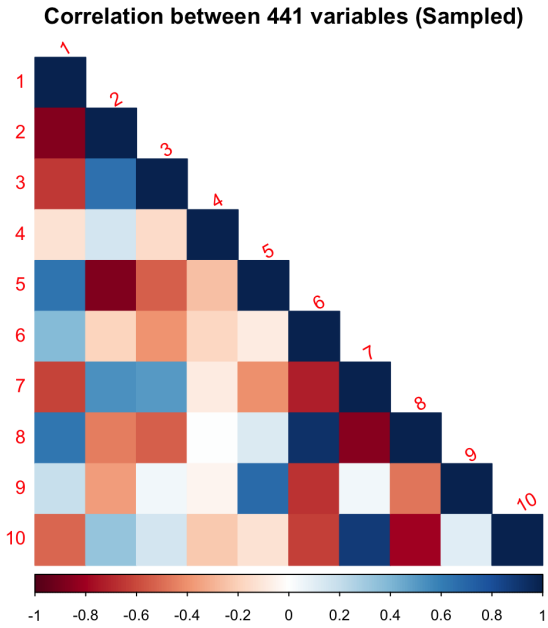

Figure 7: Correlation Between 441 Variables of $\Gamma_t\Gamma_s$ (Sampled)

Given the original heatmap is large, here we only sampled 10 variables for better visualisation. As Figure 7 shown, we can see that some variables are highly correlated in $\Gamma_t\Gamma_s$. However, we cannot visualise any obvious patterns, hence it will be unlikely that there are true correlation within $\Gamma_t\Gamma_s$.

Question 1.5

Can these products $\mathbf{TC} \times \Gamma_s$ and $\Gamma_t \times \mathbf{SM}$ exist?

Yes, both of these terms exist, as the dimension of the matrix matches.

$\mathbf{X} = (\mathbf{TC} + \Gamma_t) \times (\mathbf{SM} + \Gamma_s) = (\mathbf{TC} \times \mathbf{SM}) + (\mathbf{TC} \times \Gamma_s) + (\Gamma_t \times \mathbf{SM}) + (\Gamma_t \times \Gamma\mathbf{s})$
As we can see, $(\mathbf{TC} \times \mathbf{SM})$ is a linear combination of sources, $(\Gamma_t \times \Gamma_s)$ produces a structured noise. Second and third term of $\mathbf{X}$ will either produce structured noise or straight zeros on pixels with no values. Hence we can incorporate it into last term $\mathbf{E} = (\mathbf{TC} \times \Gamma_s) + (\Gamma_t \times \mathbf{SM}) + (\Gamma_t \times \Gamma_s)$ to simplify the model.

Plot 100 randomly selected time-series from X and variance of all 441 variables

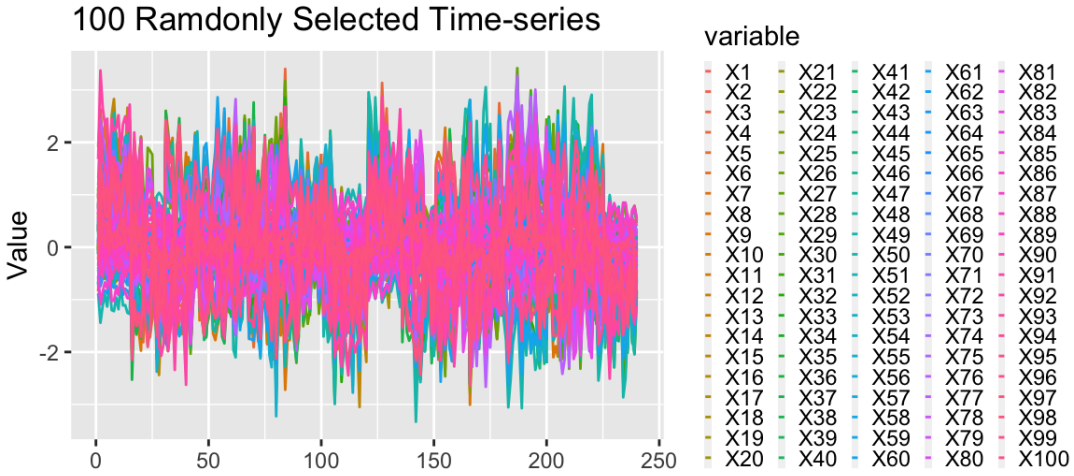

Figure 8: 100 Randomly Selected Time-series from X

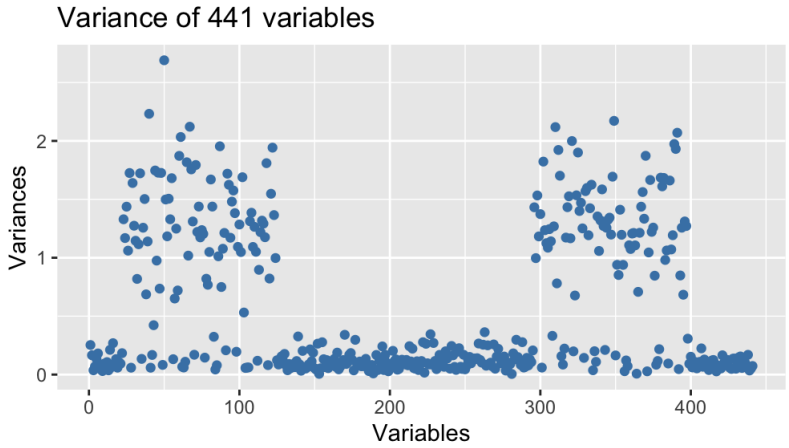

Figure 9: Variance of 441 Variables

Figure 8 shows 100 randomly selected time-series, Figure 9 shows variance of 441 variables, majority of the variables has zero variance or close to zero variance, and some variables has higher variance which forms two cluster, making it inconsistent.

Question 2 Data analysis, results visualization & performance metrics

Question 2.1

Plot six retrieved sources using $\mathbf{A}_{LSR}$ and $\mathbf{D}_{LSR}$ side by side

Retrieved Spatial Map and Time Courses are shown below

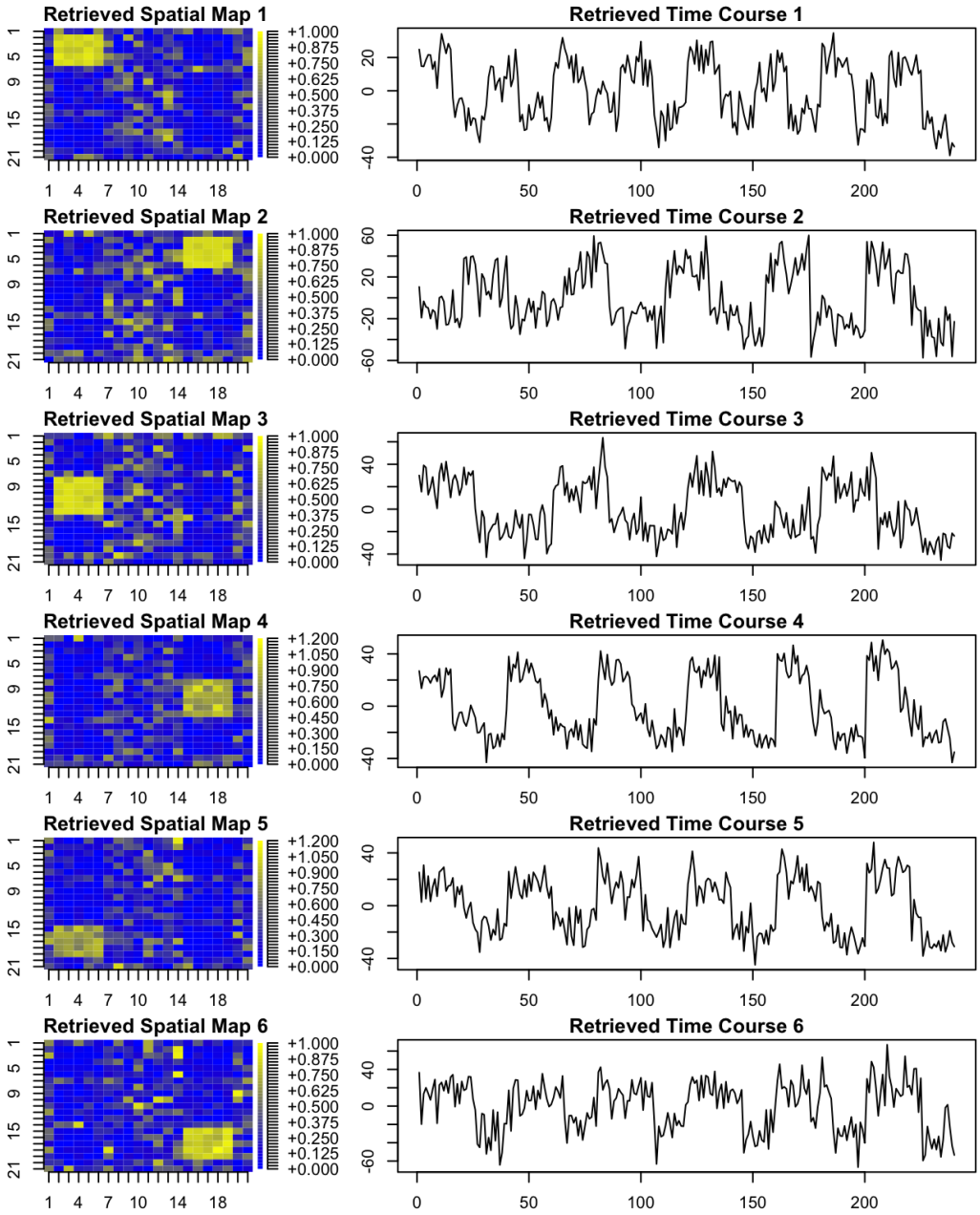

Figure 10: Retrieved Sources

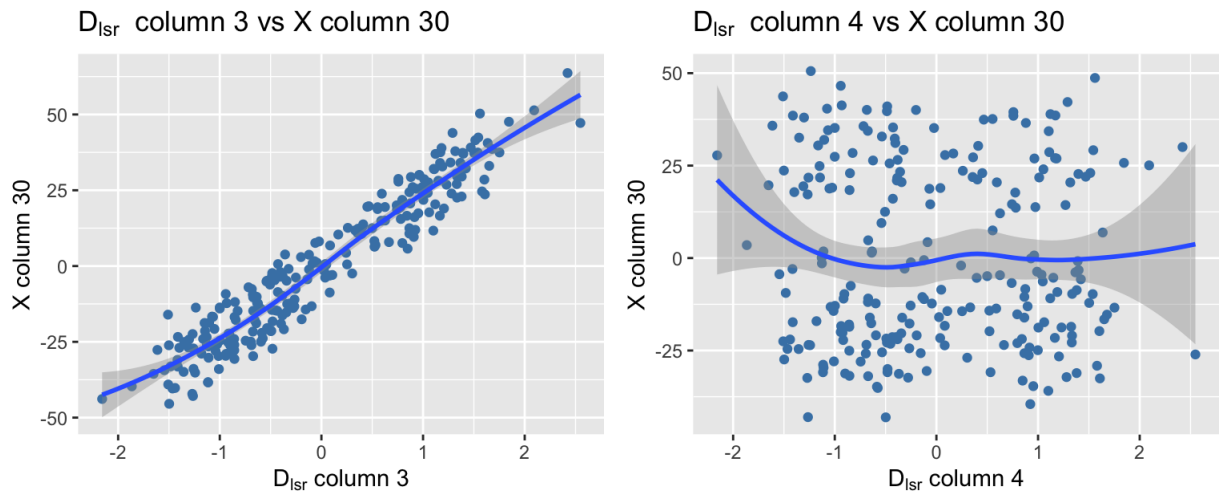

Figure 11: Scatter Plot betweem $\mathbf{D}_{LSR}$ column 3 and $\mathbf{X}$ column 30

The 30th pixel position is filled by the third SM, so the third TC is the only time course that constructs 30th column of X. As Figure 11 shown, we can observe a positive linear relationship between between 3rd column of $\mathbf{D}_{LSR}$ and 30th column of $\mathbf{X}$. As such, there's no relationship can be seen between 4th column of $\mathbf{D}_{LSR}$ and 30th column of $\mathbf{X}$.

Question 2.2

After estimating RR parameters $\mathbf{A}_{RR}$ and $\mathbf{D}_{RR}$ using $\lambda = 0.5$. Calculate the correlations between each **TC** and $\mathbf{D}_{LSR}$ and store it in $\mathbf{c}_{TLSR}$, the correlation between each **TC** and $\mathbf{D}_{RR}$ and store it in $\mathbf{c}_{TRR}$. We've obatined the sum of these two correlation vectors. $\Sigma\mathbf{c}_{TLSR} = 5.097481$ and $\Sigma\mathbf{c}_{TRR} = 5.162858$. We can see that $\Sigma\mathbf{c}_{TRR}$ are indeed greater than $\Sigma\mathbf{c}_{TLSR}$.

For $\lambda = 1000$, plot first vector from $\mathbf{A}_{RR}$ and the corresponding vector from $\mathbf{A}_{LSR}$

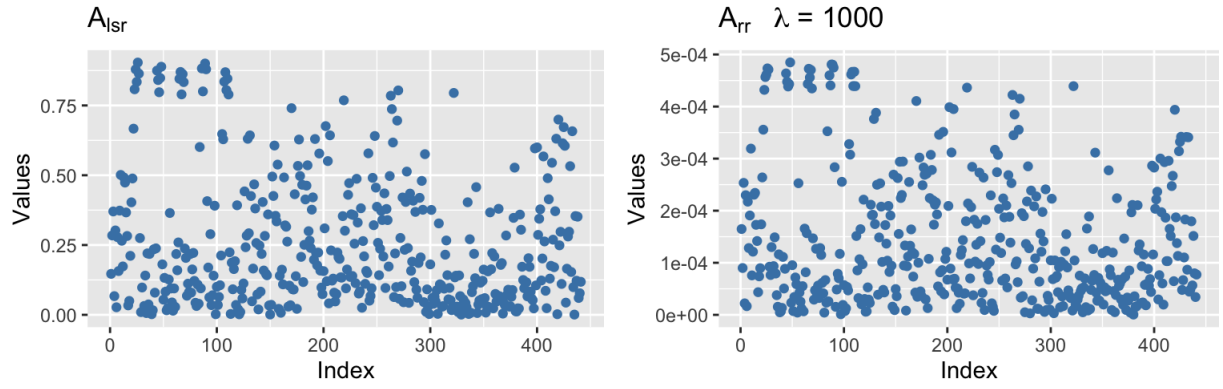

Figure 12: First vector of $\mathbf{A}_{LSR}$ and First vector of $\mathbf{A}_{RR}$

As we can see from Figure 12, all variables in the first vector of $\mathbf{A}_{RR}$ are significantly shrinked towards zero.

Question 2.3

Plot the average MSE over these 10 realizations against each value of ρ

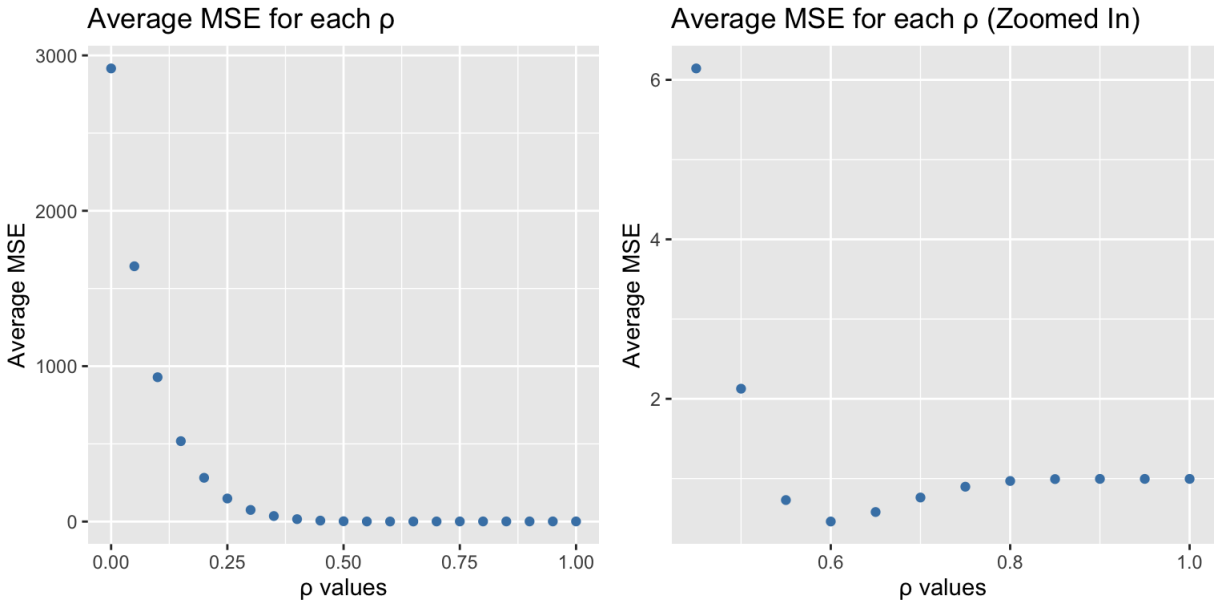

Figure 13: Average MSE for each ρ

Figure 13 shows the average MSE for each ρ, as we can see from the zoomed in plot on the right, the average MSE are minimised at $\rho = 0.6$. It is okay to select this ρ value, since we are aiming the minimise the MSE. And the average MSE starts to increase again when ρ value is greater than 0.6.

Question 2.4

After estimating the LR parameter using $\rho = 0.6$, four correlation vectors are estimated by retaining only maximum absolute correlations.

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

And we've obtained that,

$$\Sigma \mathbf{c}_{TLR} = 5.294203$$
$$\Sigma \mathbf{c}_{TRR} = 5.162858$$
$$\Sigma \mathbf{c}_{SLR} = 5.043101$$
$$\Sigma \mathbf{c}_{SRR} = 3.388821$$

We can see that $\Sigma \mathbf{c}_{TLR} > \Sigma \mathbf{c}_{TRR}$ and $\Sigma \mathbf{c}_{SLR} > \Sigma \mathbf{c}_{SRR}$

Plot side by side in form of 4 columns estimates of D and A for both RR and LR to know the difference visually.

Plots are located next page.

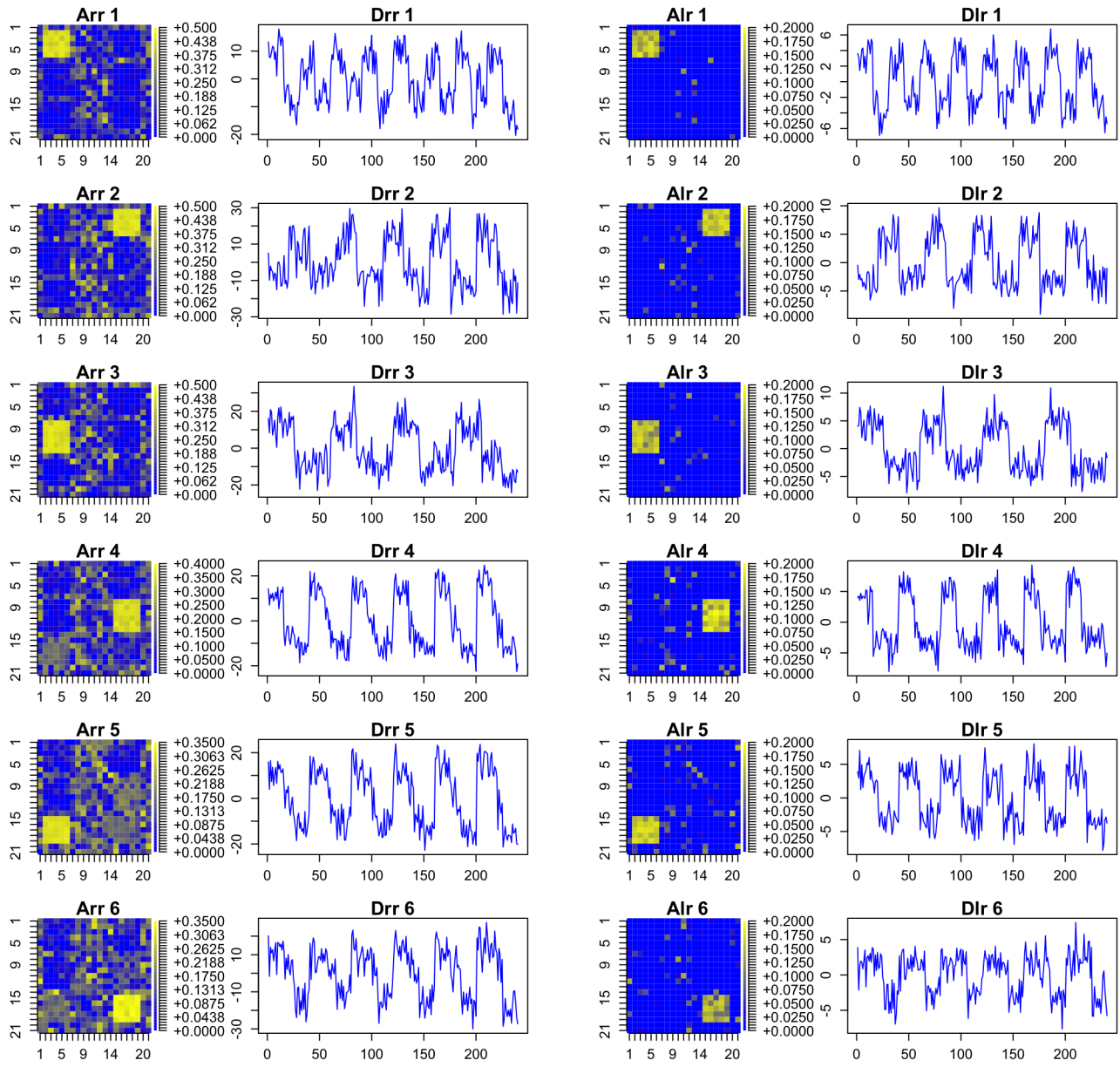

Figure 14: Estimates of $\mathbf{D}$ and $\mathbf{A}$ for both RR and LR

As we can see from Figure 14, Lasso Regression eliminate less significant variable by shrinking their coefficient to absolute zero, hence we can see that false positive in the plots of $\mathbf{A}_{LR}$ are significantly less than the one of $\mathbf{A}_{RR}$. On the other hand, Ridge Regression never leads to a coefficient of zero rather only minimises it. And Ridge Regression's bad performance is mainly because it producing many false positives while recovering coefficients; it incorporate the noise carrying pixels into the estimate of $\mathbf{A}$.

Question 2.5

Plot their eigenvalues

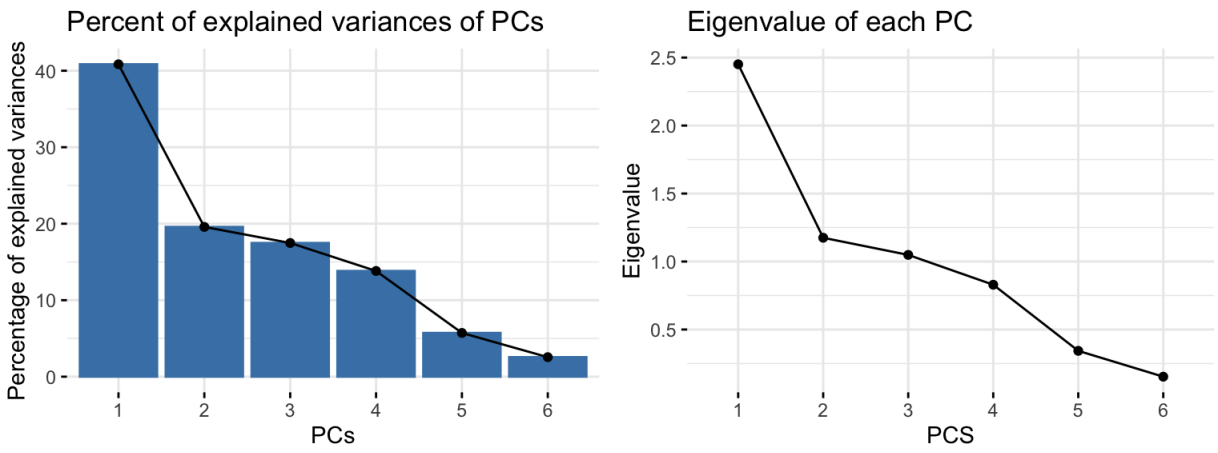

Figure 15: Eigenvalues of each PCs

As we can see from Figure 15, the plot on the left shows how much variation each PC captures from the data. And the plot on the right shows the eigenvalue of each PC. Thus, we can see that principle component 6 has the smallest eigenvalue. And based on the Kaiser Criterion, the first three components seems have capture most of the information.

	PC1 <dbl>	PC2 <dbl>	PC3 <dbl>	PC4 <dbl>	PC5 <dbl>	PC6 <dbl>
X1	-0.07547725	-0.65622983	-0.27383063	0.694910226	-0.02352637	0.072309562
X2	0.00182036	0.23060954	-0.92286110	-0.154524689	-0.26694917	-0.001893751
X3	-0.07644906	-0.69510789	-0.08033206	-0.701461272	0.08221610	0.075630108
X4	-0.55928169	0.09419415	-0.17961053	0.023956167	0.68756474	-0.415634868
X5	-0.59912548	0.14242208	0.03861215	0.002059207	-0.02130854	0.786648215
X6	-0.56276821	-0.06204901	0.18204424	-0.024409908	-0.66949671	-0.444387298

Figure 16: Coefficients of the linear combination of the TCs

Plot the regressors in Z and source TCs side by side

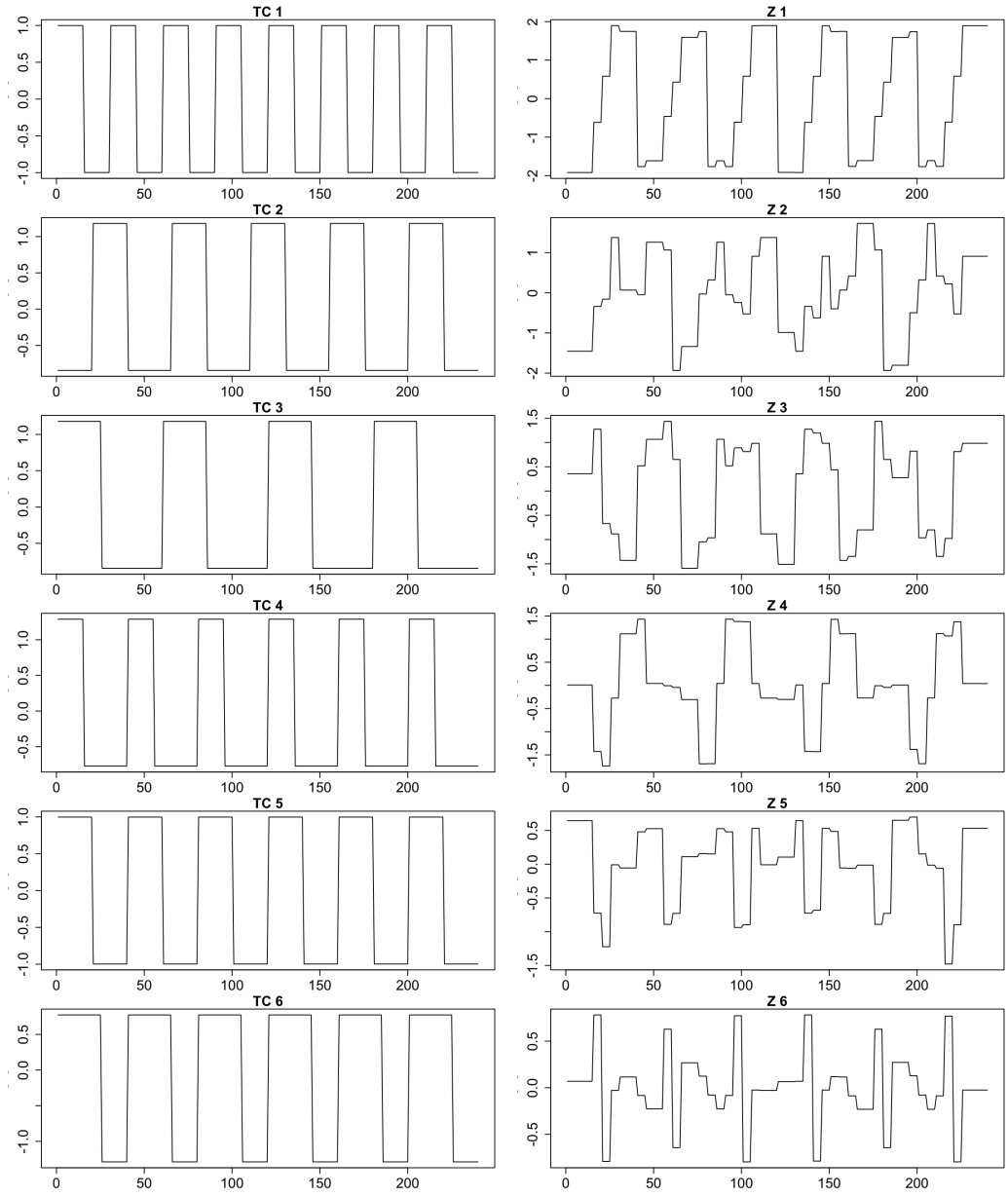

Figure 17: TCs and **Z**

It is clear when comparing **Z** to TCs in Figure 17, the shape of the PCs are deteriorated. PCs are a linear combination of the TCs. The shapes of the TCs are lost because it has been projected to the direction of the loading vectors. Therefore, not all the variances of the ground truth are kept.

Plot the results of $\mathbf{D}_{PCR}$ and $\mathbf{A}_{PCR}$ side by side. Did you notice the inferior performance of PCR compared to the other three regression models? Why is that so?

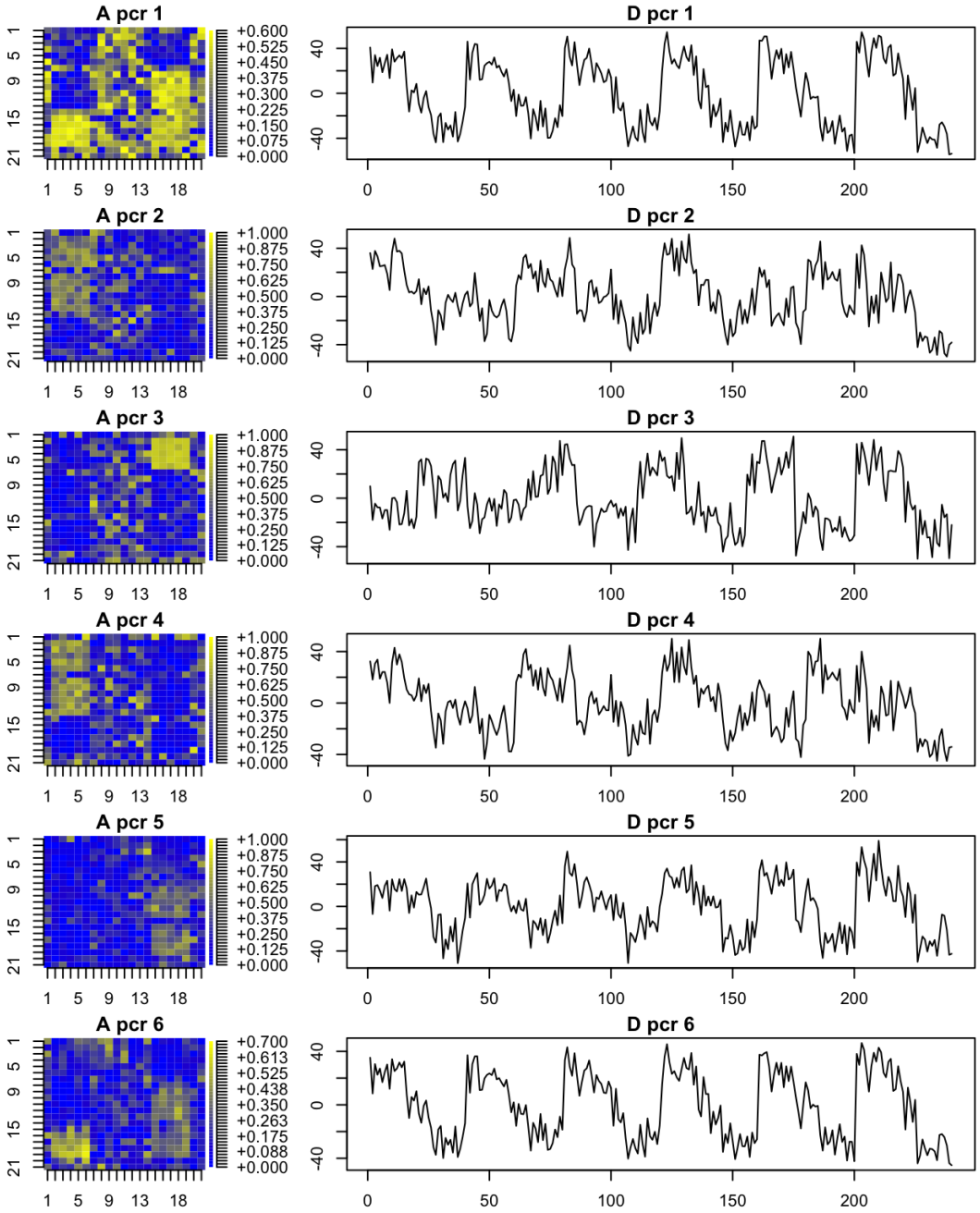

Figure 18: Plots of $\mathbf{A}_{PCR}$ and $\mathbf{D}_{PCR}$

As Figure 18 shown, we can see that the **PCR** is inferior compared to the other three regression models. The **PCR** has created more noise in areas that were previously 0, and even failed to maintain some of the slices; which might be because of the **PCR** is constructing an estimate of the data from the linear combination of the TCs. And as we can see from the first spatial map of $\mathbf{A}_{PCR}$, the yellow slices can be seen to come from **SM 4**, **SM 5**, and **SM 6** in Figure 3. And it is mainly because that the first PC is used as the regressor, and we can notice from Figure 16 that the last three coefficients in the first principle component are higher when absolute. And the coefficients of the first three **SM** is close to zero.