963632 MAST30034
Kai Stevens-Noguchi **Assignment 1** September 9, 2021

Question 1

Synthetic dataset generation, data preporcessing, & data visualization

1. Generated time courses, **TC**

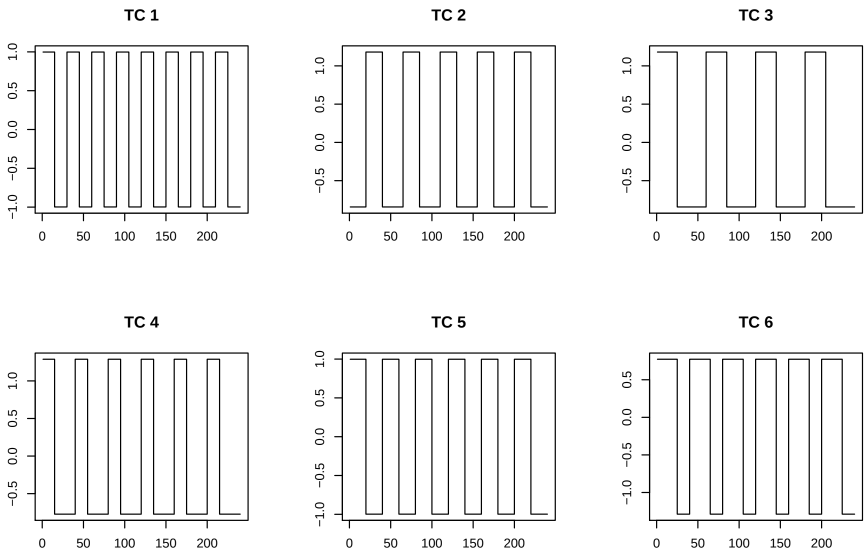


Figure 1: Generated TC variables

When scaling the data, we standardize the dataset and not normalize. This is because the original dataset is consist of 0's and 1's and if normalization (min-max scaling) is done to this, it won't affect the data at all.

2. **TC** correlation

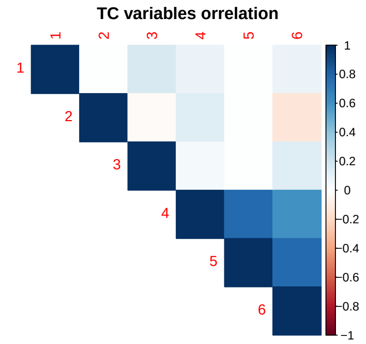
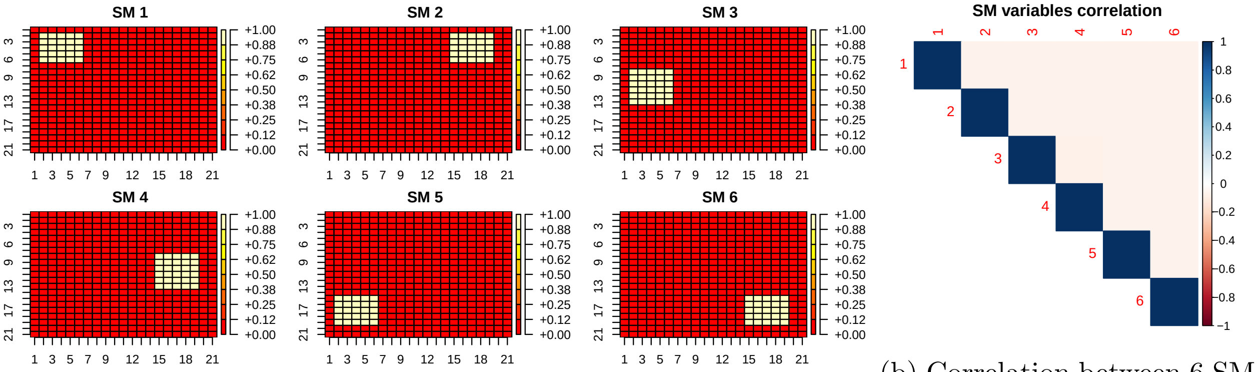


Figure 2: Correlation between 6 TC variables

Fig 2 shows that variable 4 & 5 and 5 & 6 are highly correlated and 4 & 6 with medium correlation.

3. Generate spatial sources, **SM**



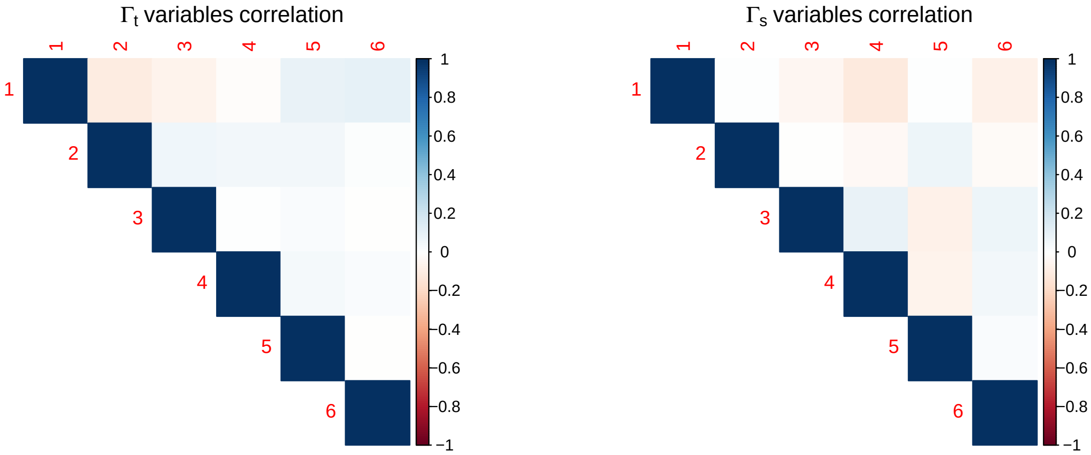
(a) Generated SM variables

(b) Correlation between 6 SM variables

Figure 3: Generated SM and correlation matrix

Fig 3b shows all six variables have no correlation among them, meaning that they are independent. Here, standardization is not important as all the spatial sources have the boolean value which are consistent throughout.

4. Generate Gaussian noise for temporal and spatial sources

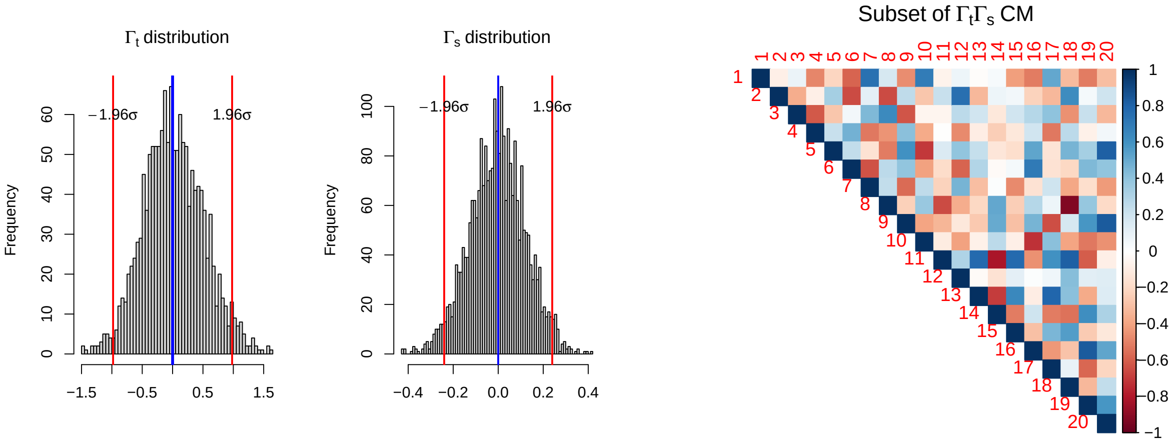


(a) $\mathbf{\Gamma}_t$ correlation matrix

(b) $\mathbf{\Gamma}_s$ correlation matrix

Figure 4: Correlation matrix for $\mathbf{\Gamma}_t$ and $\mathbf{\Gamma}_s$

From Fig 4, it can be observed that both $\mathbf{\Gamma}_t$ and $\mathbf{\Gamma}_s$ are not correlated across the sources.



(a) Distribution of $\mathbf{\Gamma}_t$ and $\mathbf{\Gamma}_s$

(b) Subset of $\mathbf{\Gamma}_t\mathbf{\Gamma}_s$ correlation matrix

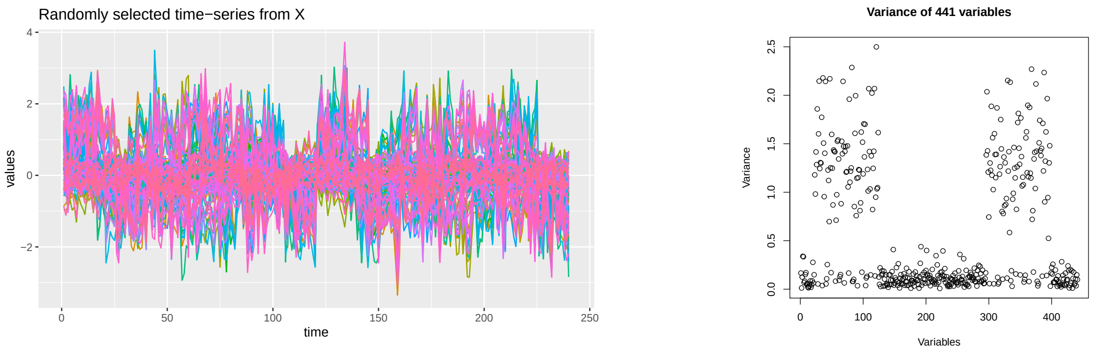
Figure 5: Distribution and correlation matrix

Two distribution (Fig 5a) exhibits shape of the normal distribution with all noises centred around $\mu = 0$ and seems to fullfills the 1.96σ criteria relating to 0.25, 0.015 (shown by red lines). For the correlation of $\mathbf{\Gamma}_t\mathbf{\Gamma}_s$, Fig 5b suggests that there are no particular pattern to the correlations and rather a noise.

5. Generate synthetic standardized dataset $\mathbf{X}$

$$\mathbf{X} = (\mathbf{TC} + \mathbf{\Gamma}_t) \times (\mathbf{SM} + \mathbf{\Gamma}_s) = \mathbf{TC} \times \mathbf{SM} + \mathbf{TC} \times \mathbf{\Gamma}_s + \mathbf{\Gamma}_t \times \mathbf{SM} + \mathbf{\Gamma}_t \times \mathbf{\Gamma}_s$$

Both $\mathbf{TC} \times \mathbf{\Gamma}_s$ and $\mathbf{\Gamma}_t \times \mathbf{SM}$ exists as the dimensions are both $(240 \times 6) \cdot (6 \times 441)$ and these terms are accounted in the model error $\mathbf{E}$.



(a) 100 random sample of X

(b) Variance of 441 variables from X

Figure 6: Random sample of X and variance of X

The two peak in Fig 6b suggests that the variance of 441 variables from X is a bimodel. This due to two different variance used when constructing the Gaussian noises.

Question 2

Data analysis, results visualization, & performance metrics

1. Retrieving SM and TC with the Least Square Regression (LSR)

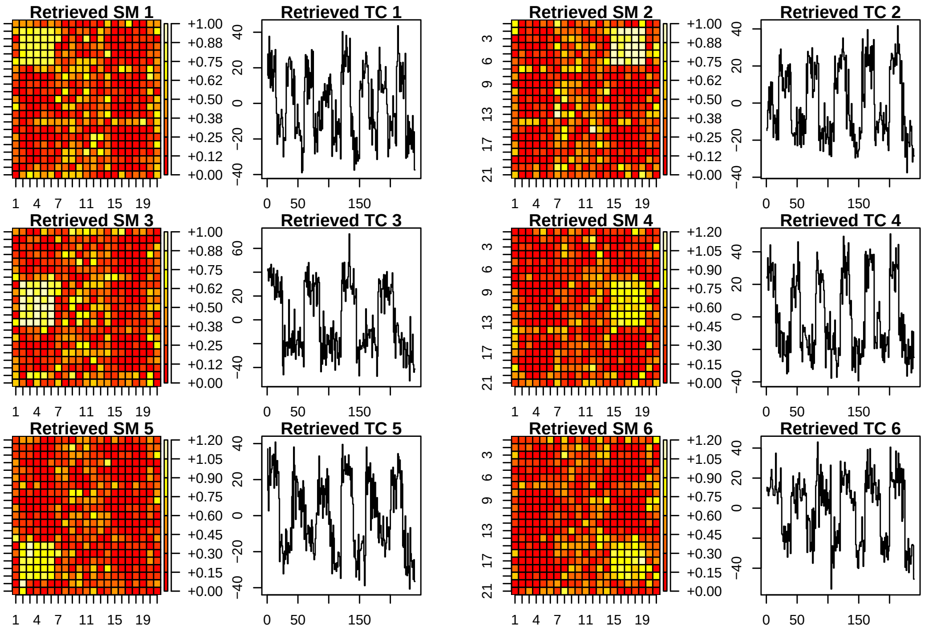
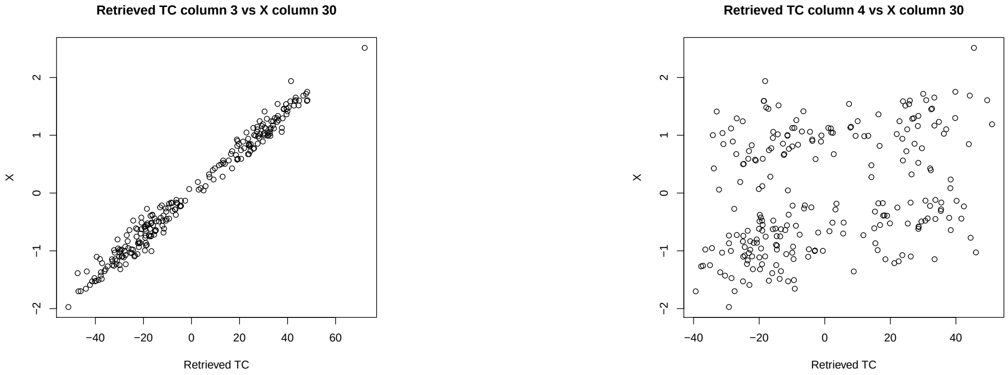


Figure 7: Retrieved SM and TC using LSR



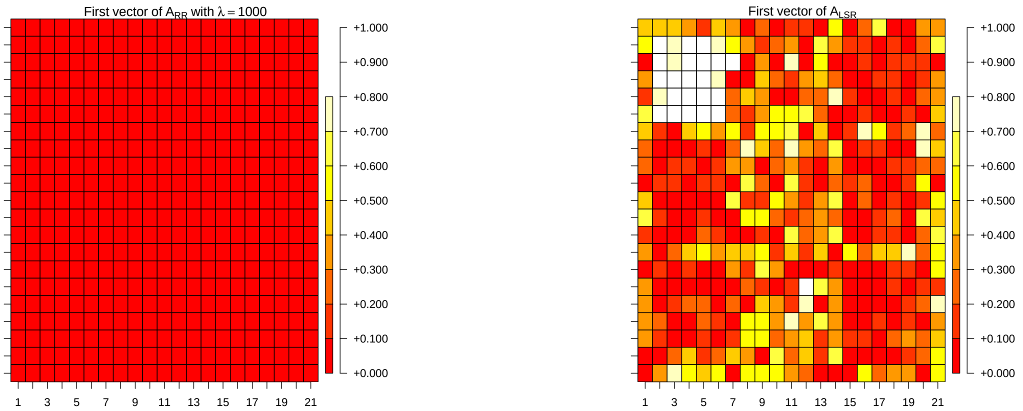
(a) Retrieved TC column 3 vs X column 30 (b) Retrieved TC column 4 vs X column 30

Figure 8: Retrieved TC vs SM

Fig 8a exhibits a linear relationship while Fig 8b does not because when retrieving the third column of $\mathbf{D}_{LSR}$, the third column of $\mathbf{A}_{LSR}$ and 30th column of $\mathbf{X}$ is used and rest of the data become a noise, hence only Fig 8a shows relationship.

2. Estimate the Ridge Regression parameters

   For $\lambda = 1$, we find that $\sum \mathbf{c}_{TRR} = 5.2605$ and $\sum \mathbf{c}_{TLSR} = 5.2711$ and here $\sum \mathbf{C}_{TRR} > \sum \mathbf{c}_{TLSR}$.

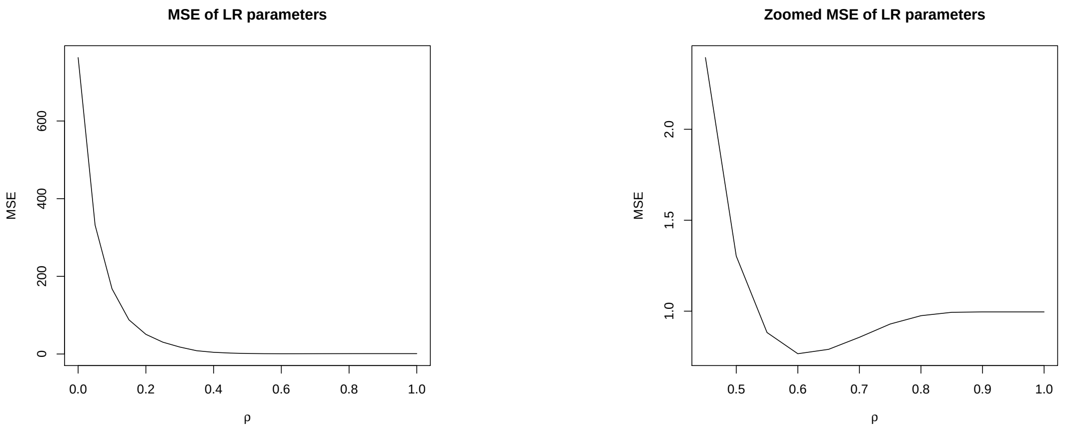


(a) $\mathbf{a}^1_{RR}$ with $\lambda = 1000$ (b) $\mathbf{a}^1_{LSR}$

Figure 9: $\mathbf{a}^1$ comparison for RR and LSR

Fig 9 shows that the all values in $\mathbf{a}^1_{RR}$ has shrunk towards zero, even the area where we expect to see the lice, compared to $\mathbf{a}^1_{LSR}$.

3. Retrieving SM and TC with the Lasso Regression



(a) LR parameters MSE (b) LR parameters MSE zoomed

Figure 10: MSE of LR parameters

Fig 10b shows that MSE is minimum at $\rho = 0.6$ and from here the MSE started to increase again. This value of ρ is reasonable as this controls the penalty for the Lasso Regression.

4. Estimate the Lasso Regression parameters

For the Ridge Regression, $\lambda = 1$ and for the Lasso Regression, $\rho = 0.6$ are used. Calculating the correlation vectors yields; $\sum \mathbf{c}_{TLR} = 5.3942$, $\sum \mathbf{c}_{TRR} = 5.2711$, $\sum \mathbf{c}_{SLR} = 362.7012$ and $\sum \mathbf{c}_{SRR} = 356.1611$. Hence $\sum \mathbf{c}_{TLR} > \sum \mathbf{c}_{TRR}$ and $\sum \mathbf{c}_{SLR} > \sum \mathbf{c}_{SRR}$.

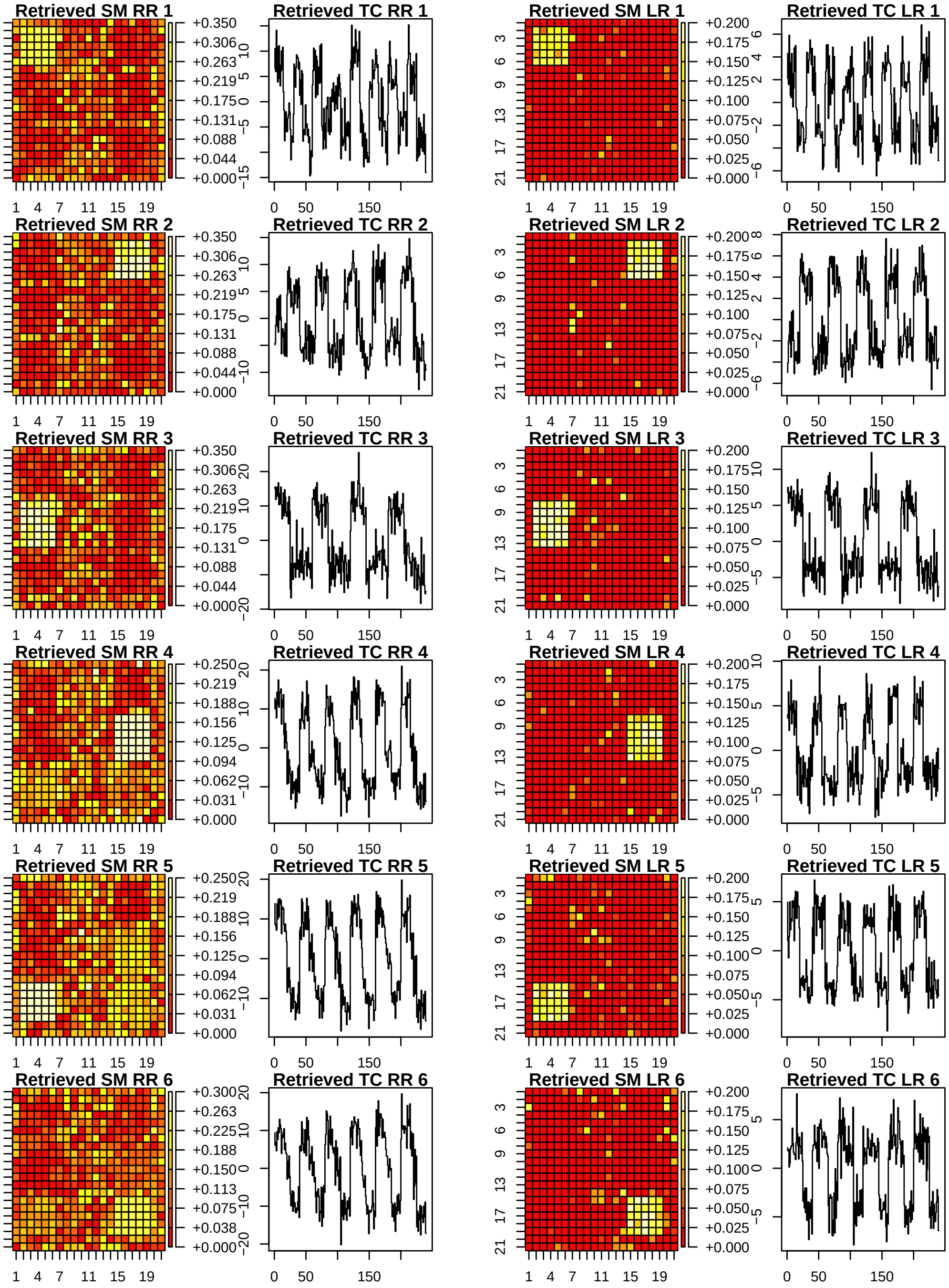


Figure 11: Retrieved SM and TC with RR and LR

While the Lasso Regression was able to retrieve the SM accurately, the Ridge Regression has higher false positives especially for variables 4 to 6 in the SM. One reasoning is due to the high correlation in these variables (Fig 3b). Furthermore, the Ridge Regression penalises predictor to have higher weight than others while the Lasso Regression assigns 0 to non-important predictors. In this case, the Lasso Regression successfully extracted relevant features while Ridge Regression struggles yields useful features.

5. Estimate PCs of the TCs

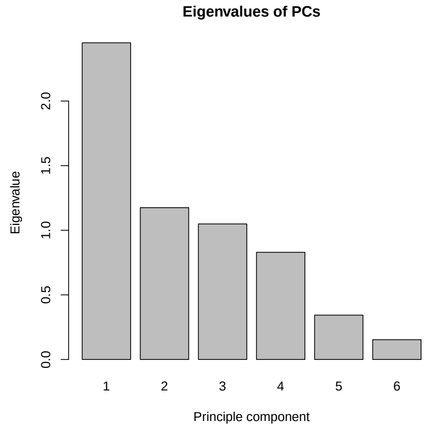


Figure 12: Eigenvalues of the Principal Components

Fig 12 shows that the sixth principal component has the smallest eigenvalue.

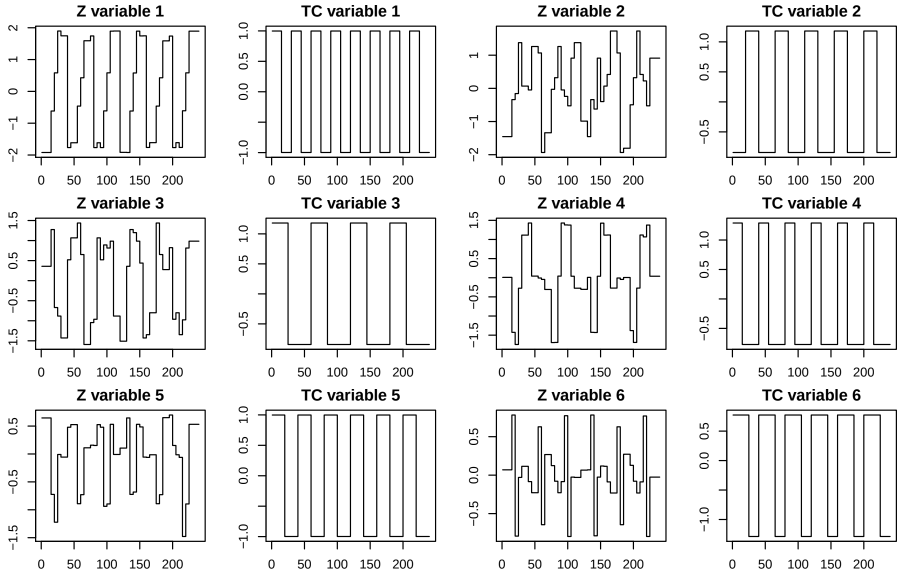


Figure 13: Regressors **Z** and source TCs

In Fig 13, the distortion is obvious for the latter variables, **Z** variable 2 and onward. This can be explained by the eigenvalues of the principal components (Fig 12). The PCR has placed more importance on the principal component 1 (higher eigenvalue) while other components have lower eigenvalue and hence the distortion in the shape of TCs.

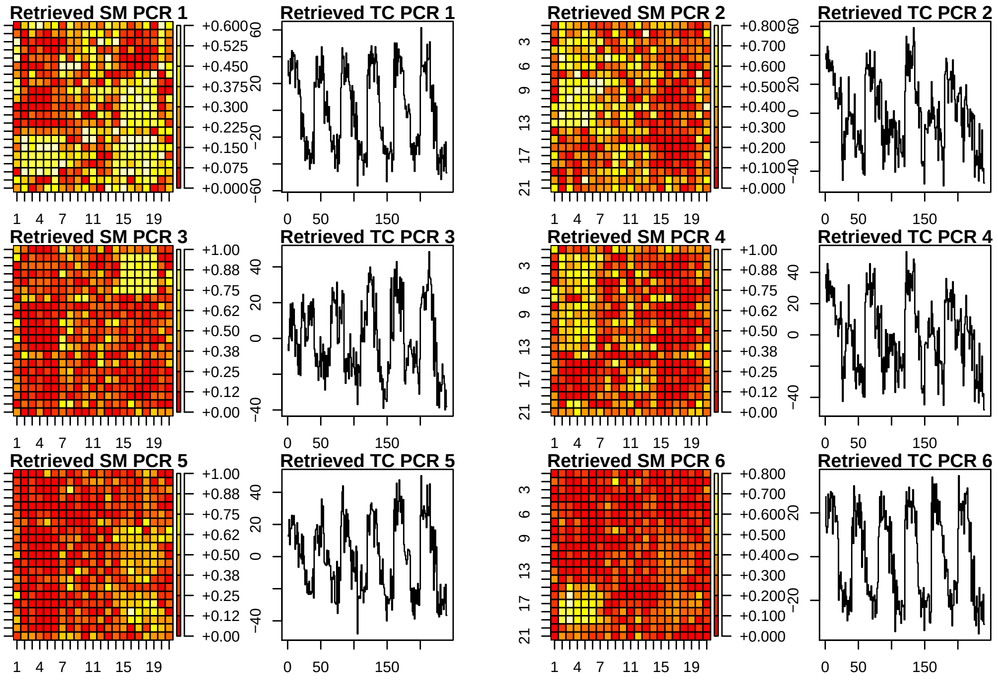


Figure 14: $\mathbf{D}_{PCR}$ and $\mathbf{A}_{PCR}$

The Principle Component Regression has inferior performance compared to the other three regression models. This is can be seen by comparing retrieved SM in Fig 14 to Fig 7, 11. One possible explanation to this is due to the loss of information as PCR only uses a subset of all the principal components.