Spatiotemporal Dataset Analysis

MAST30034: Assignment 1

Jason F. Suhartanto
1086250

Semester 2, 2021.

1 Synthetic Data Generation, Preprocessing, and Visualization

Addressing the spatiotemporal data properties, we have these constant values to be used throughout the entire analysis:

1. $N = 240$, the number of variables of the temporal sources.
2. $V = 441$, the number of variables of the spatial sources; where $x1 = 21$ and $x2 = 21$ is the shape of each two-dimensional spatial source.
3. $NSRCS = 6$, the number of spatiotemporal sources.

Additionally, a glossary of abbreviations is provided below for better understanding of the analysis.

1. TC: Time Courses, the temporal source.
2. SM: Spatial Maps, the spatial source.
3. CM: Correlation Matrix.
4. MSE: Mean Squared Error.
5. LSR: Least Square Regression.
6. RR: Ridge Regression
7. LR: LASSO (Least Absolute Shrinkage and Selection Operator) Regression.
8. PCR: Principal Component Regression.

This analysis is performed using Python 3.9.

1.1 Temporal Sources Generation: Time Courses (TC)

Figure 1: Time Courses **TC** generation from six sources.

The temporal sources are generated from the provided onset arrival vector (AV), increment vector (IV), and duration of ones of six sources. **TC** matrix has size 240x6. Figure 1 shows the normalized **TC** accross six sources. Each column is standardized to have a mean of zero and unit variance.
We do not normalize by dividing it by l-2 norm because while the data is scaled to be between zero and one, the contribution of every source will depend on their scales which will cause a potential bias. Additionally, we are unable to infer if the data is already Gaussian, so standardization is preferred.

1.2 Correlation Matrix Between TCs

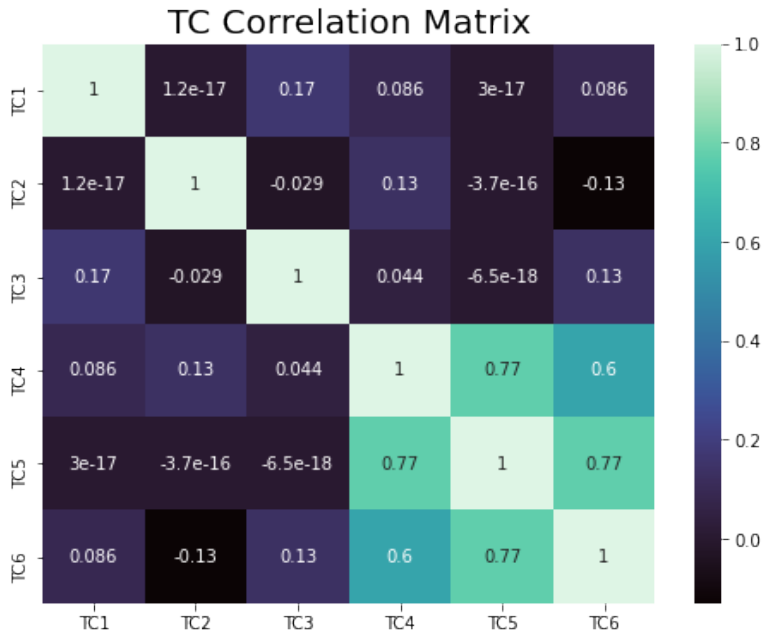

Figure 2: Correlation Matrix (Pearson) between six Time Course (TC) sources.

As of Figure 2, there are two pairs of TCs that have the highest correlation: TC4 and TC5 (0.77), and TC5 and TC6 (0.77). Additionally TC4 and TC6 have a correlation score of 0.6. The other TCs do not show a significant correlation with any of the other TCs.

1.3 Spatial Sources Generation: Spatial Maps (SM)

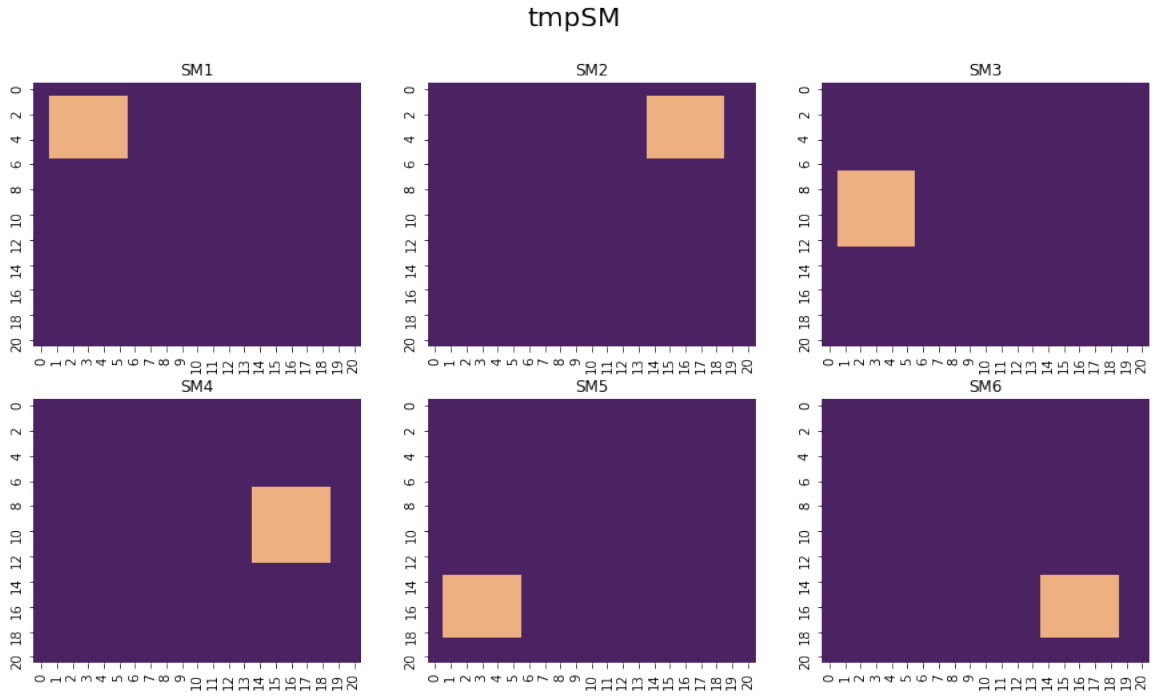

Figure 3: Spatial Maps (SM) Generation from six sources.

For the spatial sources, the ones are visualized as the orange squares on the heatmaps, as shown in Figure 3. The ticks use Python indexing. To match the indexing in R in the assignment questions, the numbers are adjusted by one. E.g. for source 1, indices [2:6, 2:6] inclusive becomes indices [1:5, 1:5] inclusive in Python. The resulting **SM** matrix has size 6x441.

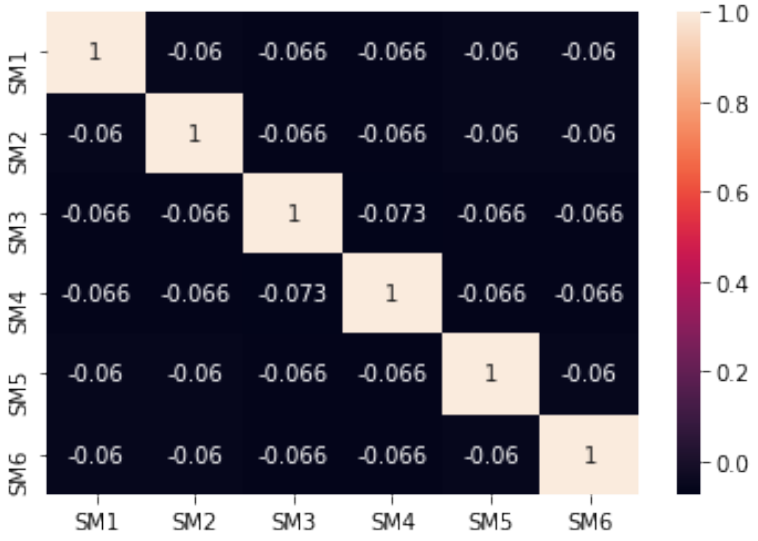

Figure 4: Correlation matrix between six generated spatial sources.

As in Figure 4, the generated Spatial Maps (SM) sources do not have significant correlation between each other. However, correlation does not imply independence so we cannot say for sure that they are independent. Standardization for these spatial sources are not needed because each source already have a similar scale and mean.

1.4 Noise Generation ($\mathbf{\Gamma_t}$ and $\mathbf{\Gamma_s}$)

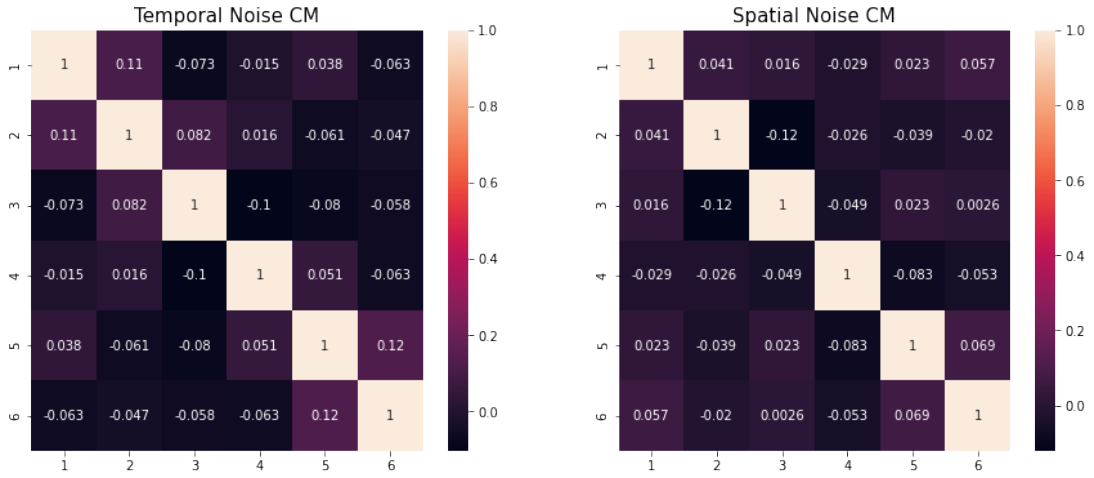

Figure 5: Correlation matrix of generated temporal noises (left) and spatial noises (right).

The generated noise for temporal ($\mathbf{\Gamma_t}$) and spatial ($\mathbf{\Gamma_s}$) do not show a significant correlation with each other, as show in Figure 5. This is expected since they are generated randomly, and desirable because we do not want the noise to change the actual correlation information of the original data.

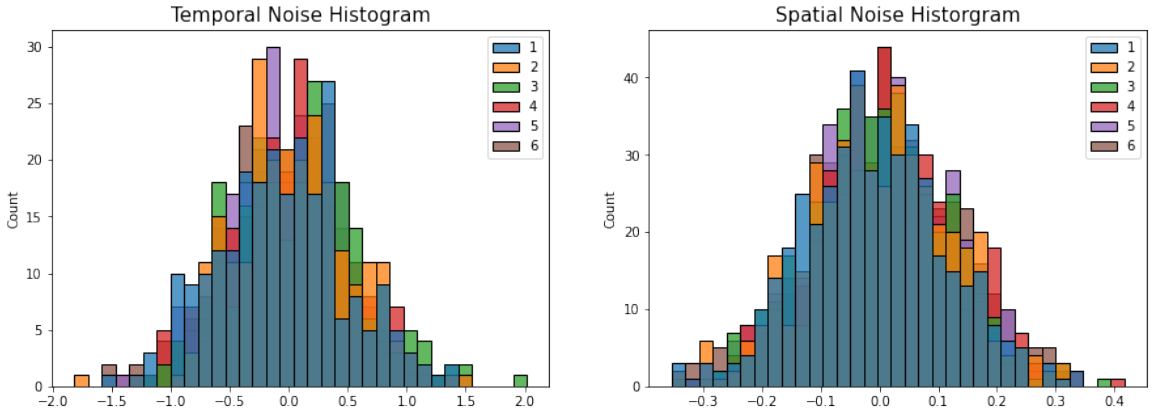

Figure 6: Histograms showing distribution of values of temporal noises (left) and spatial noises (right).

The noise for each temporal and spatial source are shown to have a normal distribution, as in Figure 6. It can also be observed that they are both centered at the zero mean, and the distribution of the temporal noise is wider than the spatial noise because it has a higher standard deviation ($\sigma_t^2 = 0.25$; $\sigma_s^2 = 0.015$). Both distributions fulfill the mean and variance criteria.

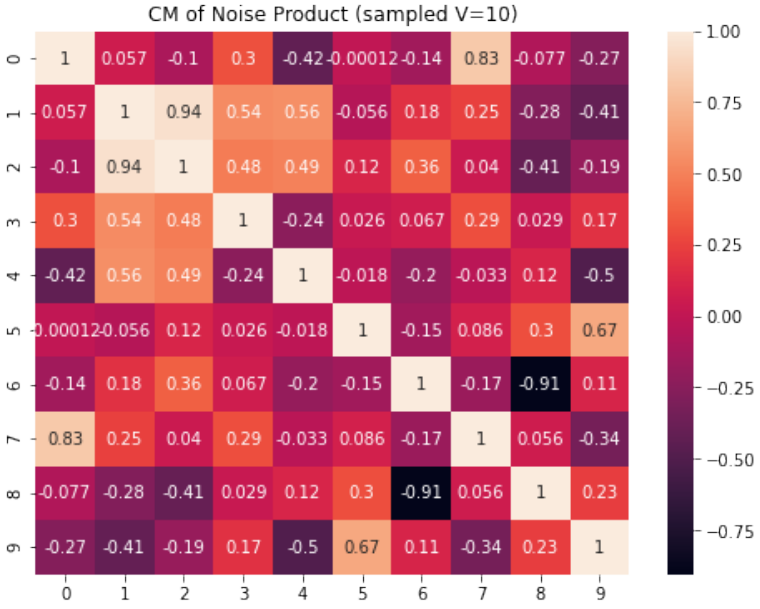

Figure 7: Correlation matrix of noise product.

Figure 7 shows the correlation of a sampled number of V (out of 441 variables) of the Noise Product $\mathbf{\Gamma_t \Gamma_s}$. It is observed that some variables are highly correlated, both positively and negatively.

1.5 Synthetic Dataset Generation

The synthetic spatiotemporal dataset $\mathbf{X}$ is created from the equation $\mathbf{X} = (\mathbf{TC} + \mathbf{\Gamma_t}) \times (\mathbf{SM} + \mathbf{\Gamma_s})$, and therefore has size 240x441. The matching matrix sizes allow $\mathbf{TC} \times \mathbf{\Gamma_s}$ and $\mathbf{\Gamma_t} \times \mathbf{SM}$ to exist. However if we use these products instead of the original equation, the information will be meaningless because of the product result of noise that is mostly centered at zero.

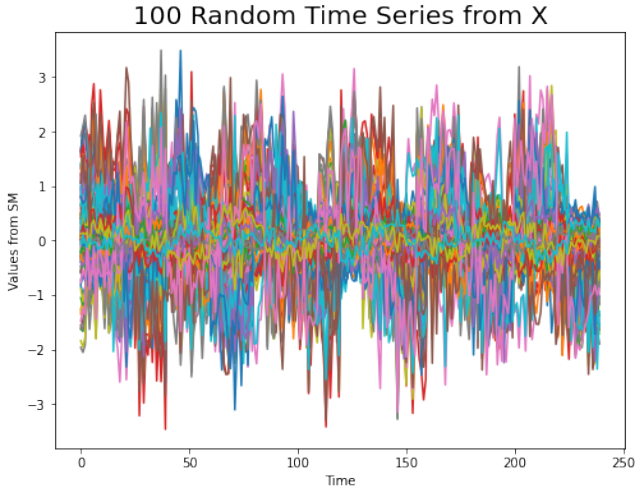

Figure 8: 100 randomly selected time series from the spatiotemporal synthetic dataset ($\mathbf{X}$) before standardization.

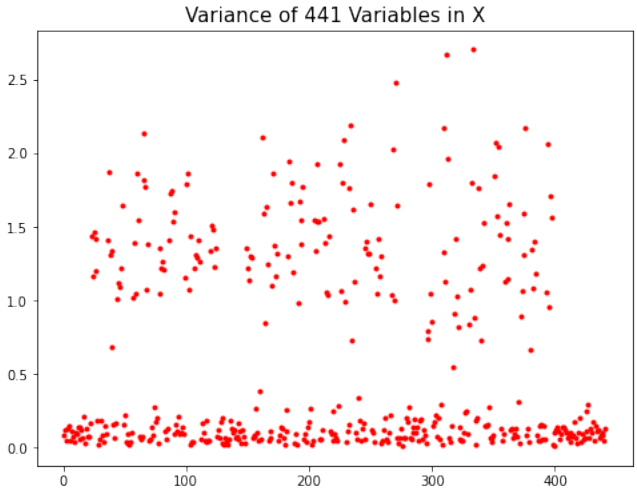

Figure 9: Variance distribution of the 441 variables from the spatiotemporal synthetic dataset $\mathbf{X}$ before standardization.

Randomly selected samples of time-series from a non-standardized $\mathbf{X}$ are shown in Figure 8, which shows behaviour of the spatial data across 241 time units. We can also observe inconsistent variances across 441 variables, as shown in Figure 9, as some variables have variance close to zero, but also many have high variances of more than 2.0. Observing these two plots, $\mathbf{X}$ needs to be standardized.

2 Data Analysis, Results Visualization, Performance Metrics

2.1 Least Square Regression (LSR)

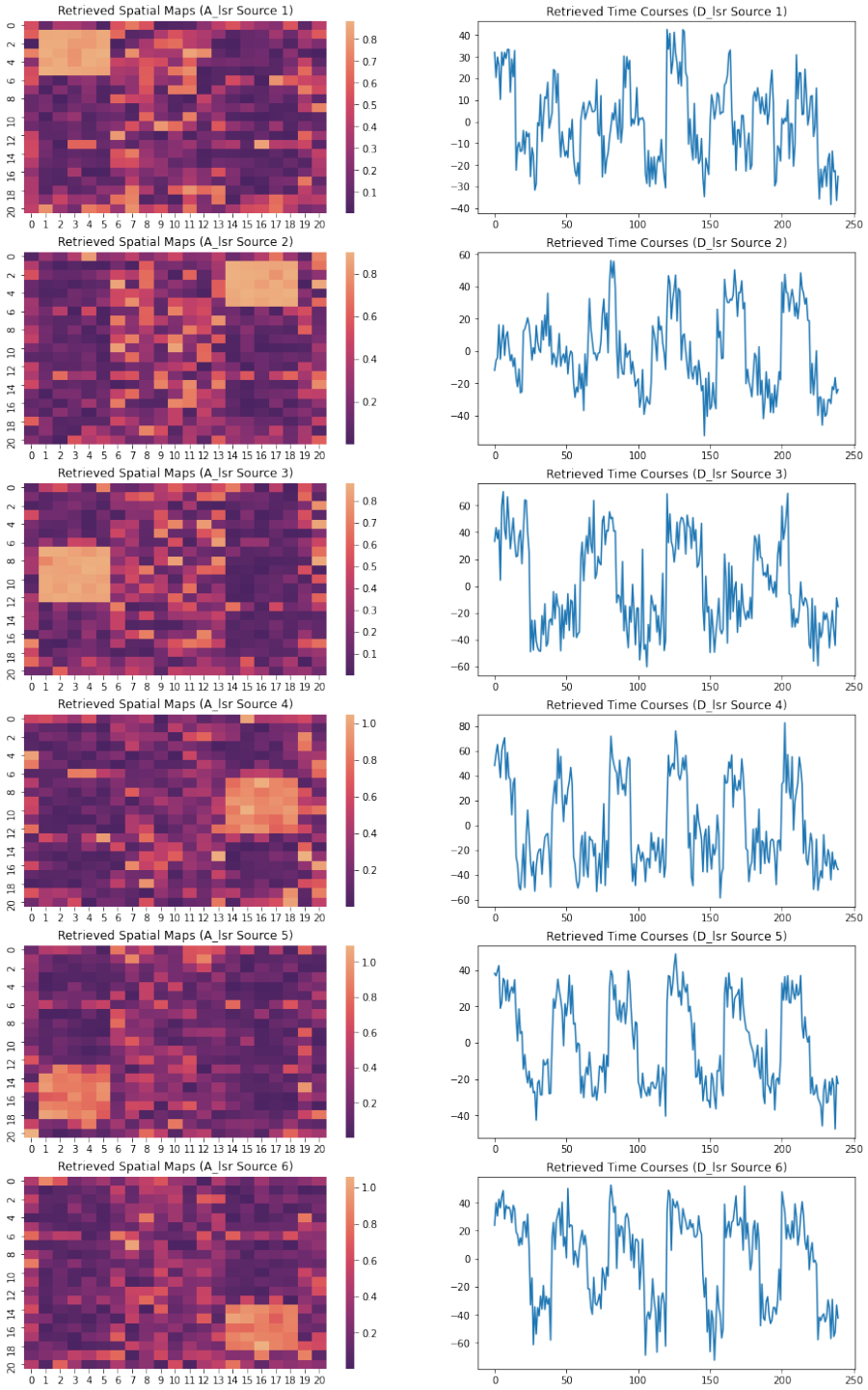

Figure 10: Retrieved spatial sources (left column) and temporal sources (right column) by least square regression (LSR).

The synthetic spatiotemporal dataset $\mathbf{X}$ follows a linear regression model $\mathbf{X} = \mathbf{DA} + \mathbf{E}$. Using the least squares method:

1. $\mathbf{A}_{LSR} = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{X}$, where $\mathbf{D} = \mathbf{TC}$.
2. $\mathbf{D}_{LSR} = \mathbf{X}\mathbf{A}_{LSR}^\top$.

From this LSR parameters estimation, these correlation vectors are obtained:

1. $\mathbf{c}_{TLSR}$, the absolute correlation vector between $\mathbf{TC}$ and $\mathbf{D}_{LSR}$.
2. $\mathbf{c}_{SLSR}$, the absolute correlation vector between $\mathbf{SM}$ and $\mathbf{D}_{LSR}$.

The retrieved spatial and temporal sources across six sources are shown in Figure 10.

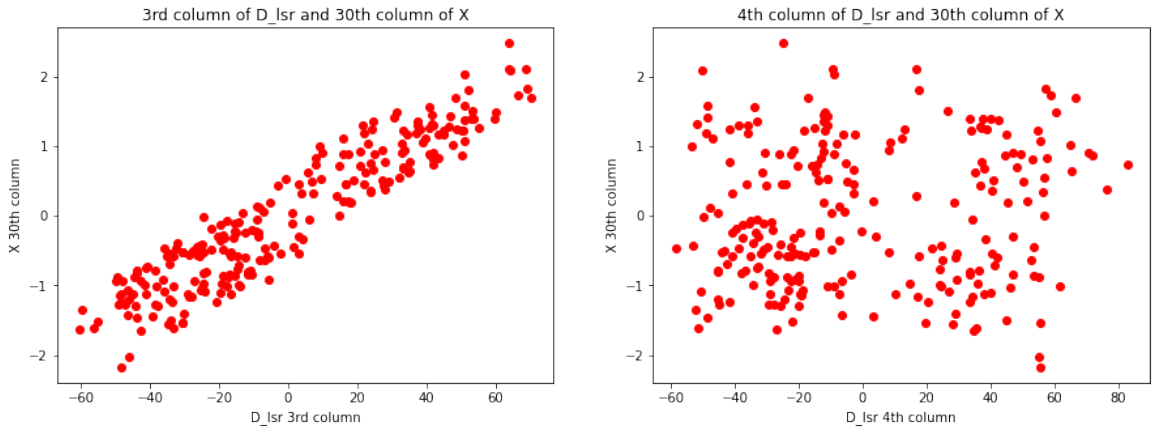

Figure 11: These scatter plots investigates linear relationships between slices of the D_{LSR} estimates and X.

In Figure 11, we can see that some columns in $\mathbf{X}$ and columns in $\mathbf{D}_{LSR}$ have a linear relationship. E.g., the 3rd column of $\mathbf{D}_{LSR}$ is constructed by the 30th column of $\mathbf{X}$ when estimating with $\mathbf{X}\mathbf{A}_{LSR}^\top$. However, this is not the case of the following 4th column in $\mathbf{D}_{LSR}$ because they are not estimated using the same 30th column variable in $\mathbf{X}$.

2.2 Ridge Regression (RR)

The Ridge Regression (RR) parameters are estimated by:

1. $\mathbf{A}_{RR} = (\mathbf{D}^\top \mathbf{D} + \tilde{\lambda}\mathbf{I})^{-1} \mathbf{D}^\top \mathbf{X}$, where $\mathbf{D} = \mathbf{TC}$ and $\tilde{\lambda} = \lambda V$.
2. $\mathbf{D}_{RR} = \mathbf{X}\mathbf{A}_{RR}^\top$.

We set $\lambda = 0.2$ for this RR estimation using a check and guess method. From this RR parameters estimation, these correlation vectors are obtained:

1. $\mathbf{c}_{TRR}$, the absolute correlation vector between $\mathbf{TC}$ and $\mathbf{D}_{RR}$.
2. $\mathbf{c}_{SRR}$, the absolute correlation vector between $\mathbf{SM}$ and $\mathbf{D}_{RR}$.

```
Correlation Vector between TC and D_lsr
c_tlsr = [0.768966477548226, 0.7765175930943867, 0.847807772427148, 0.8891160218704824, 0.8942447298432019, 0.8850693423700485]
Sum of c_tlsr = 5.061721937153493

Correlation Vector between TC and D_rr
c_trr = [0.7700344220360344, 0.7741357952352182, 0.847579321296557, 0.9134178375285964, 0.8853721069867683, 0.8802643283482977]
Sum of c_trr = 5.070803811431472
```

Figure 12: Comparison of sums of correlation vectors from Least Square Regression (LSR) and Ridge Regression (RR).

As in Figure 12, we found that the $\sum \mathbf{c}_{TRR}$ (5.0708) to be higher than $\sum \mathbf{c}_{TLSR}$ (5.0617). Therefore we have selected a good value of λ for the RR estimation, which is 0.2.

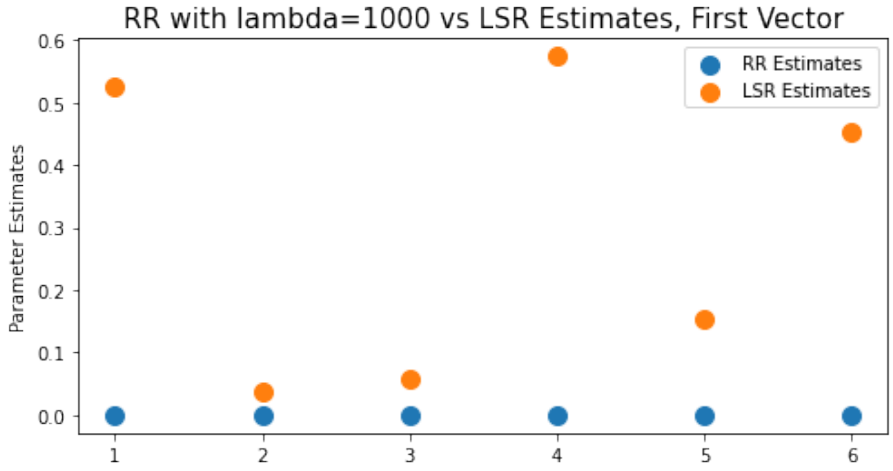

Figure 13: First vector from $\mathbf{A_{RR}}$ when $\lambda = 1000$, compared with the first vector from $\mathbf{A_{LSR}}$.

When λ is set to be 1000, we found all estimated RR parameters to shrink towards zero, as expected. This is shown in Figure 13, and compared with the estimated parameters from LSR in Section 2.1.

2.3 Finding ρ for LASSO Regression (LR)

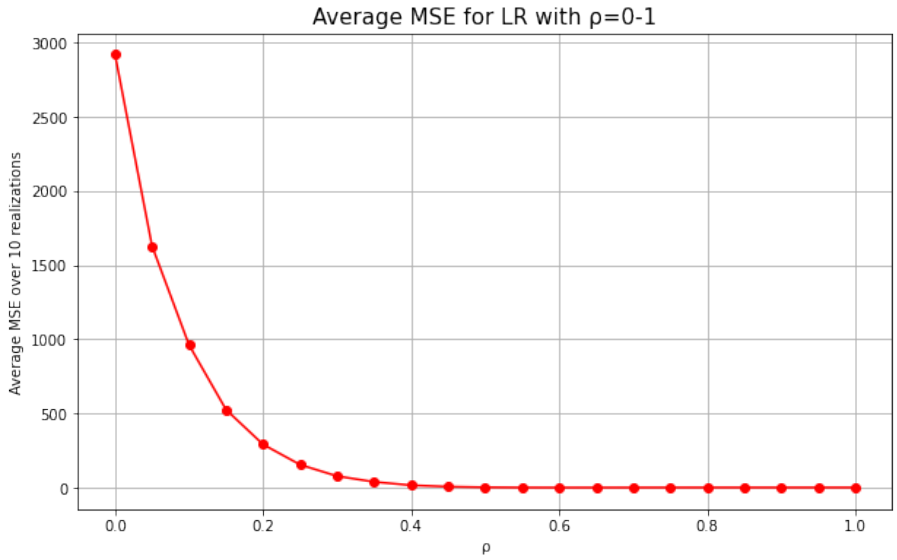

Figure 14: Average Mean Square Error (MSE) of LR estimates for different ρ values.

Testing different ρ values for LR estimation on different realizations, we found the optimal ρ value to be 0.60, as seen in Figure 14. Setting $\rho = 0.60$ gives the lowest average Mean Squared Error (MSE) of 0.4575, and therefore we select this value for the regression in the following section. The MSE is also observed to start increasing again for ρ greater than 0.60.

2.4 LASSO Regression (LR) Estimation

LASSO Regression is performed following the provided pseudocode and setting $\rho = 0.60$ as in Section 2.3. From this LR parameters estimation, these correlation vectors are obtained:

1. $\mathbf{c}_{TLR}$, the absolute correlation vector between **TC** and $\mathbf{D}_{LR}$.
2. $\mathbf{c}_{SLR}$, the absolute correlation vector between **SM** and $\mathbf{D}_{LR}$.

```
(i) Correlation Vector between TC and D_rr
c_trr = [0.7700344220360344, 0.7741357952352182, 0.847579321296557, 0.9134178375285964, 0.8853721069867683, 0.8802643283482977]
Sum of c_trr = 5.070803811431472

(ii) Correlation Vector between SM and A_rr
c_srr = [0.5496277355286115, 0.5769846642681355, 0.6199181847909212, 0.6028053485980495, 0.599095694089801, 0.5811655730447105]
Sum of c_srr = 3.5295972003202287

(iii) Correlation Vector between TC and D_lr
c_tlr = [0.878260340675357, 0.897685515141408, 0.8673621453177598, 0.9122135023428125, 0.9143097770013996, 0.9098138890295694]
Sum of c_tlr = 5.379645169508307

(iv) Correlation Vector between SM and A_lr
c_slr = [0.8429287296502073, 0.8421523083350385, 0.8480771857760774, 0.8353801739581002, 0.820297114537954, 0.7925484471472503]
Sum of c_slr = 4.981383959404628

==================================

sum(c_tlr): 5.379645169508307    >    sum(c_trr): 5.070803811431472
sum(c_slr): 4.981383959404628    >    sum(c_srr): 3.5295972003202287
```

Figure 15: Correlation vectors comparison between estimates of Ridge Regression (RR) and LASSO Regression (LR).

We found that $\sum \mathbf{c}_{TLR}$ (5.3796) to be higher than $\sum \mathbf{c}_{TRR}$ (5.0708), and $\sum \mathbf{c}_{SLR}$ (4.9814) to be higher than $\sum \mathbf{c}_{SRR}$ (3.5296). Therefore we have selected a good value of ρ for the LR estimation, which is 0.6.

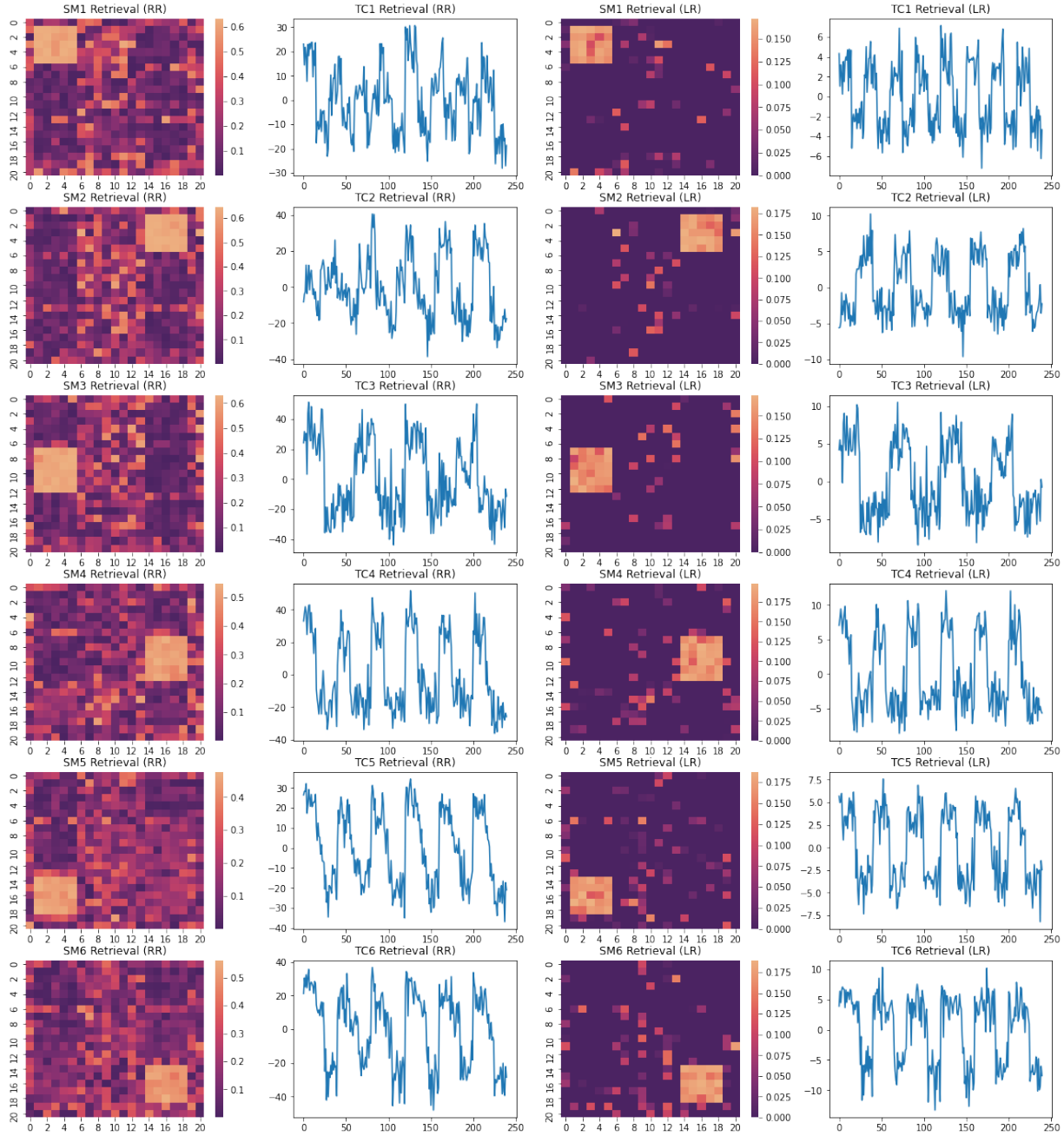

Figure 16: Retrievals comparison of Ridge Regression (left half) and LASSO Regression (right half).

The comparison of **D** and **A** estimates for LR and RR are shown in Figure 16. We can see that LR performs exceptionally better than RR, as RR gives more false positives. One reason is because LR works better for multicollinearity (MC). Another reason is because LR is able to remove variables that have very low relevance, rather than just shrinking it towards zero, and therefore reduces overfitting better. In this case, we can infer that the data carries noise.

2.5 Principal Component Regression (PCR)

Estimating the Principal Components (PC) of the TCs is performed by calling the Singular Value Decomposition[1] function from NumPy in Python. The regressors are denoted as **Z**.

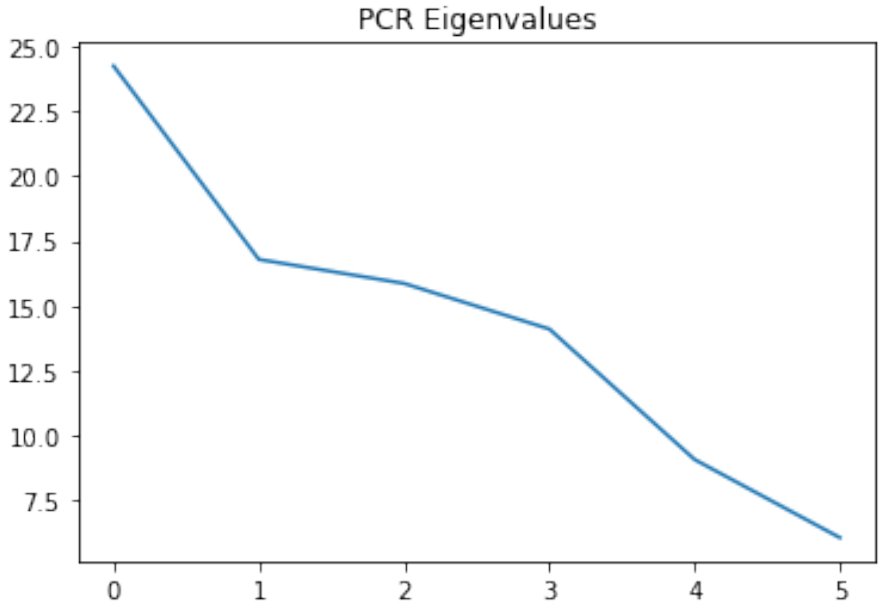

Figure 17: Eigenvalues from the Principal Component Regression.

From Figure 17, we see that the eigenvalue is smallest for the sixth Principal Component (PC).

[1]Retrieved from https://numpy.org/doc/stable/reference/generated/numpy.linalg.svd.html, accessed on September 5, 2021.

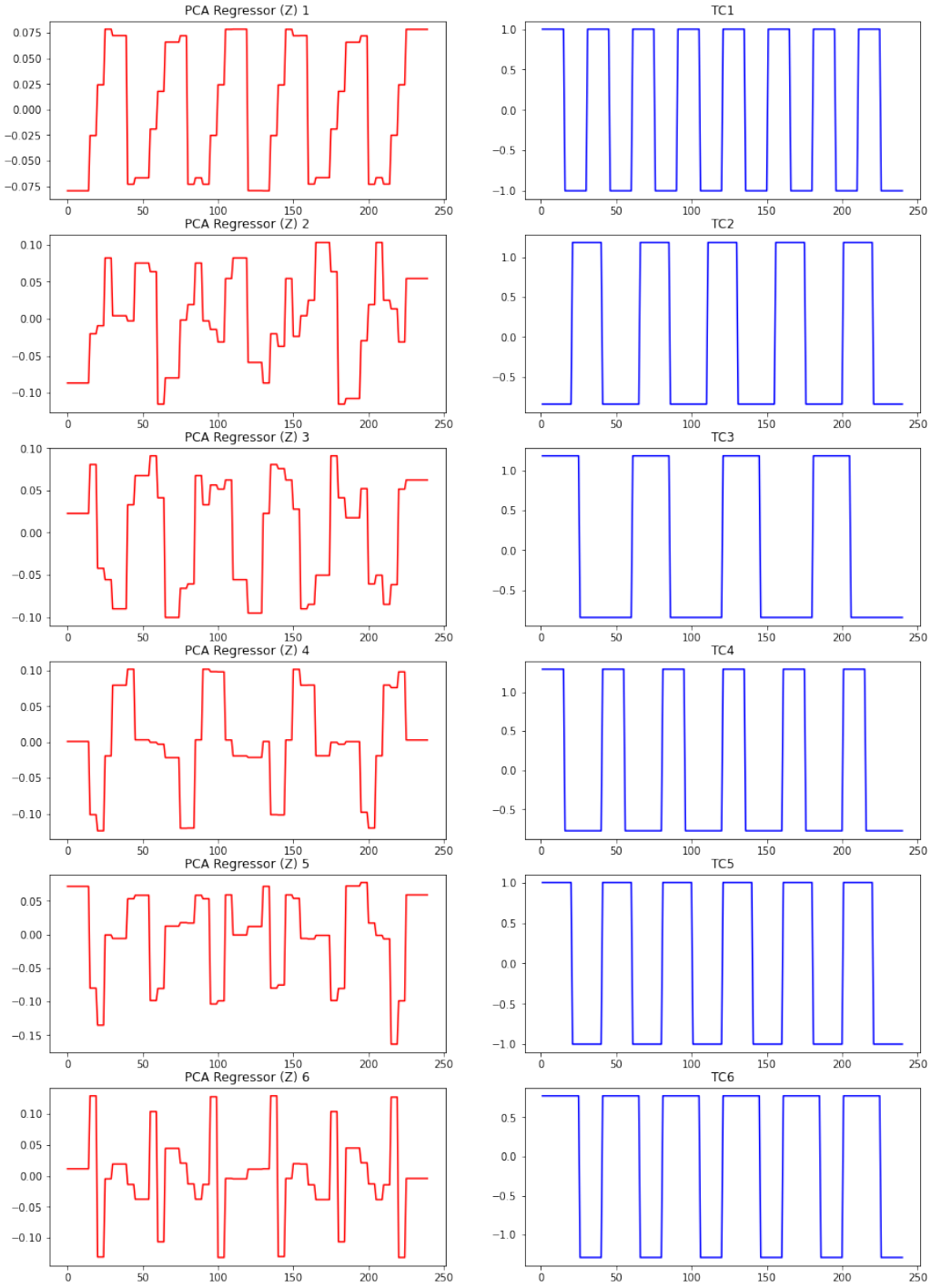

Figure 18: Regressors from PCR (left column) and the actual temporal sources (right column).

The regressors **Z** are compared with its actual TC source in Figure 18. We can also see that the PCs have a deteriorated shape and the shape of the TCs are lost. This is because the Principal Components are linear combinations of the TCs. Applying PCA before the regression will remove variability of TC without knowledge which are the most significant.

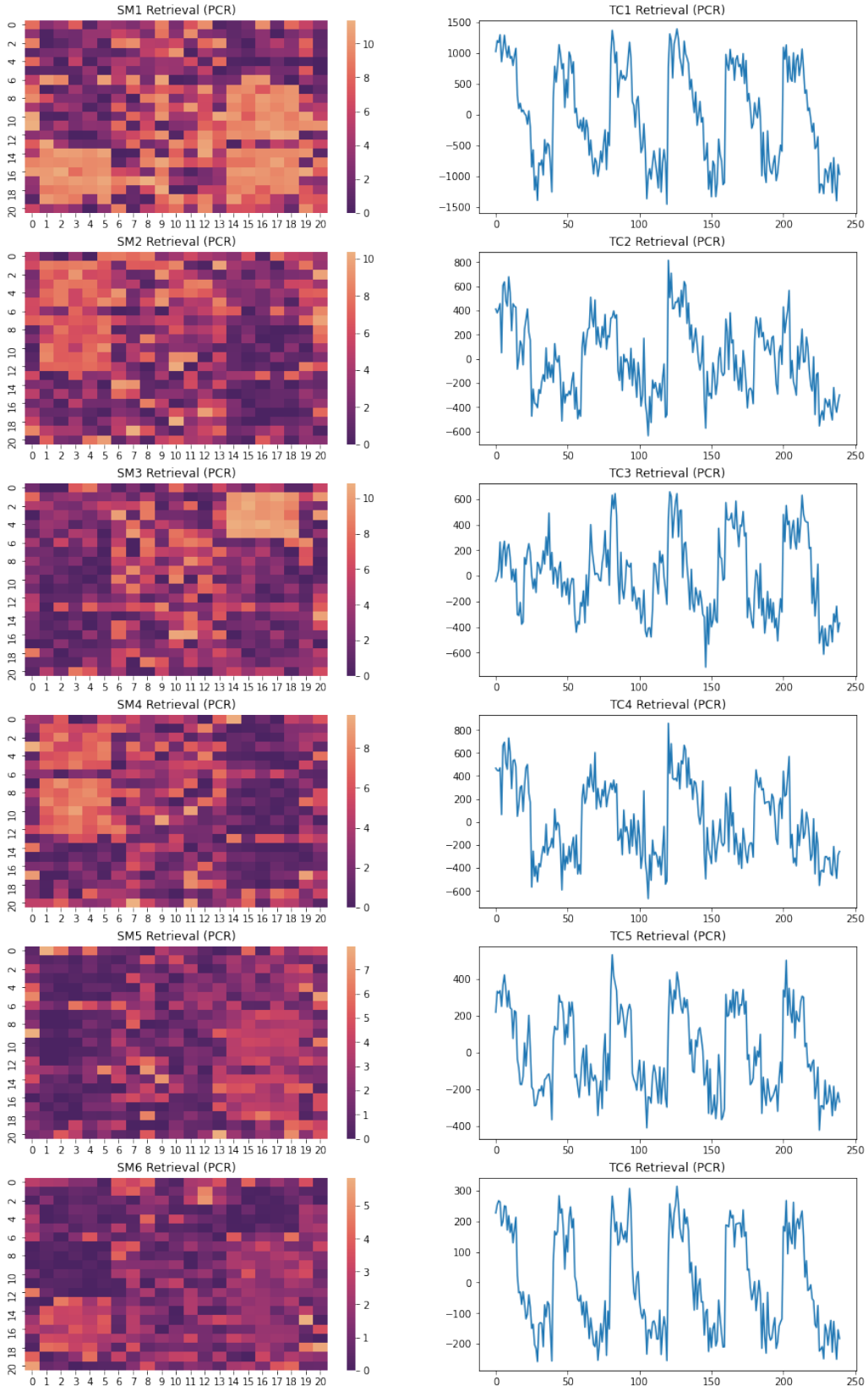

Figure 19: Spatial (left column) and Temporal (right column) retrievals of Principal Component Regression.

For the Principal Component Regression (PCR), a LASSO Regression is performed on $\mathbf{X}$ using $\mathbf{Z}$ as the regressors, instead of regressing the dependent variables directly on the explanatory variables. A ρ value of 0.001 is used. The retrievals, $\mathbf{D}_{PCR}$ and $\mathbf{A}_{PCR}$, are shown in Figure 19. This result is inferior and shows a poorer estimate compared to LSR, RR, and LR. This is expected because on PCR we performed dimensionality reduction on TC to get $\mathbf{Z}$ as the regressors $\mathbf{D}$, whereas on LSR, RR, and LR, the regressors are assumed to be known as we set $\mathbf{D} = \mathbf{TC}$, and therefore keeping a complete information. Applying PCA only makes sense if we want to simplify the data's dimensionality, not to obtain better accuracy.