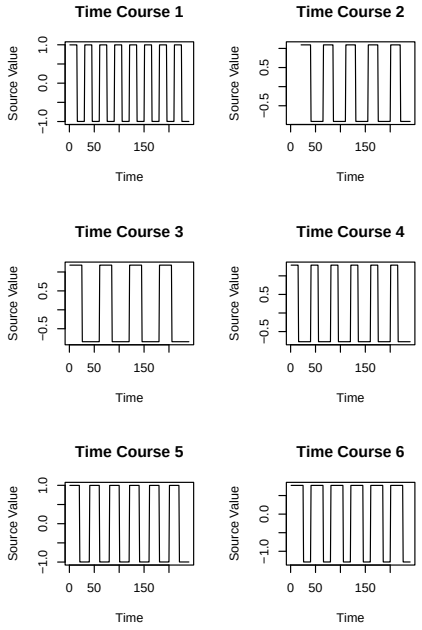
Assignment 1 Applied Data Science

By Audrey Thompson

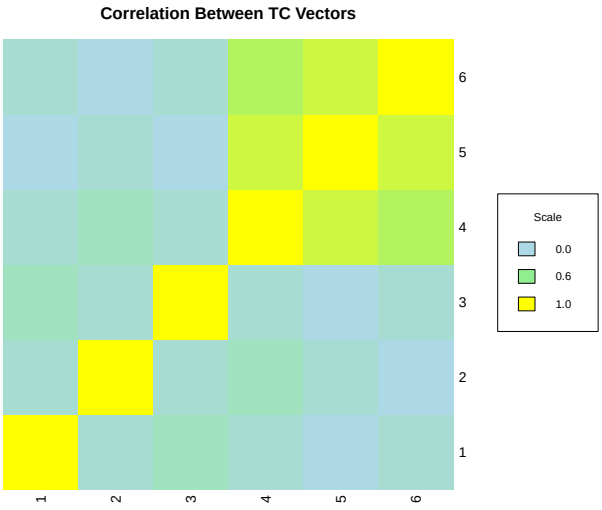
Question 1.

1.1



Normalising, while often confused with standardising, is very different. Standardising ensures the data has a unit variance and a mean of 0. Normalising the data rescales the values to $[0, 1]$. We do not want to do this as it will not make the data bias free and we want to maintain the current data scale.

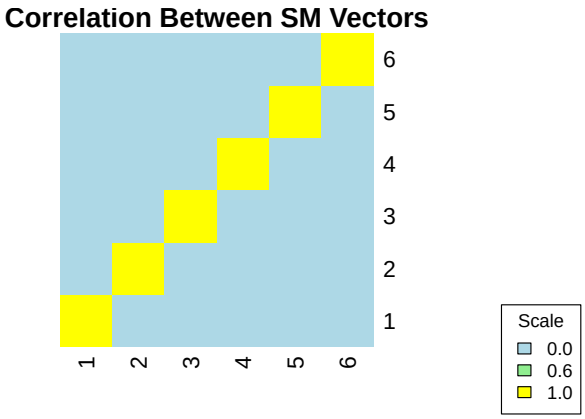
1.2



TCs 5 and 6 are highly correlated, and TCs 5 and 4 are verging on being correlated.

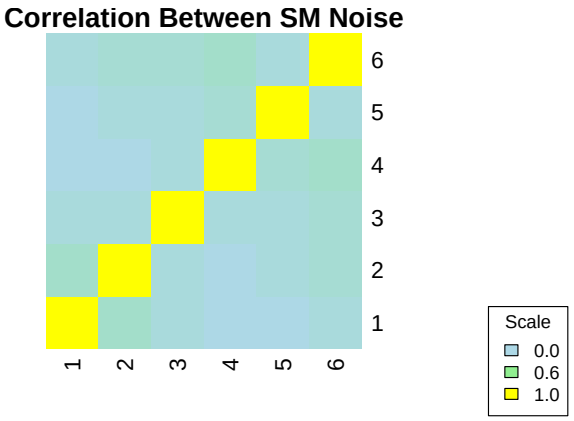
1.3

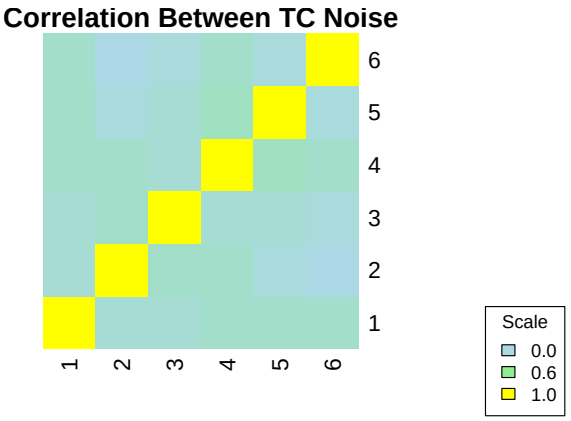
Showing if these six SMs are independent:



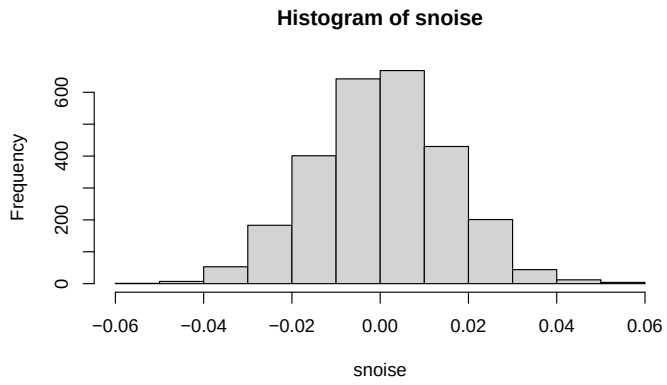
There appears to be no correlation between SM vectors.
Standardisation is not important as the SM vectors all have the same distribution of values, it is just their position in the vector that differ. Hence all of their means and variances are very similar.

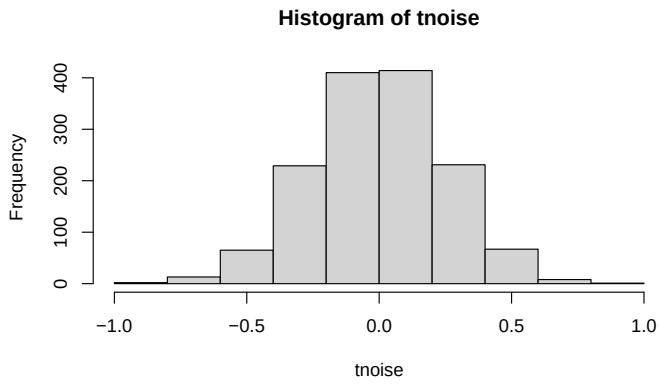
1.4



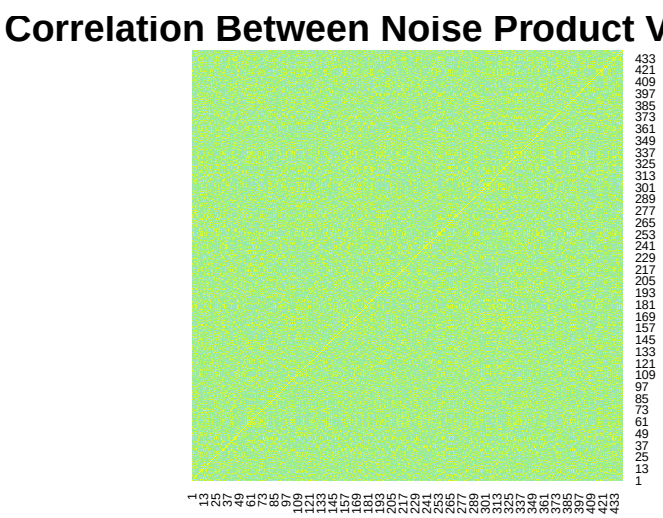


The noise does not look correlated across sources.





These distributions appear normal and reflect the distributions of $\sim$N(0, 0.015) and $\sim$N(0, 0.25) respectively.



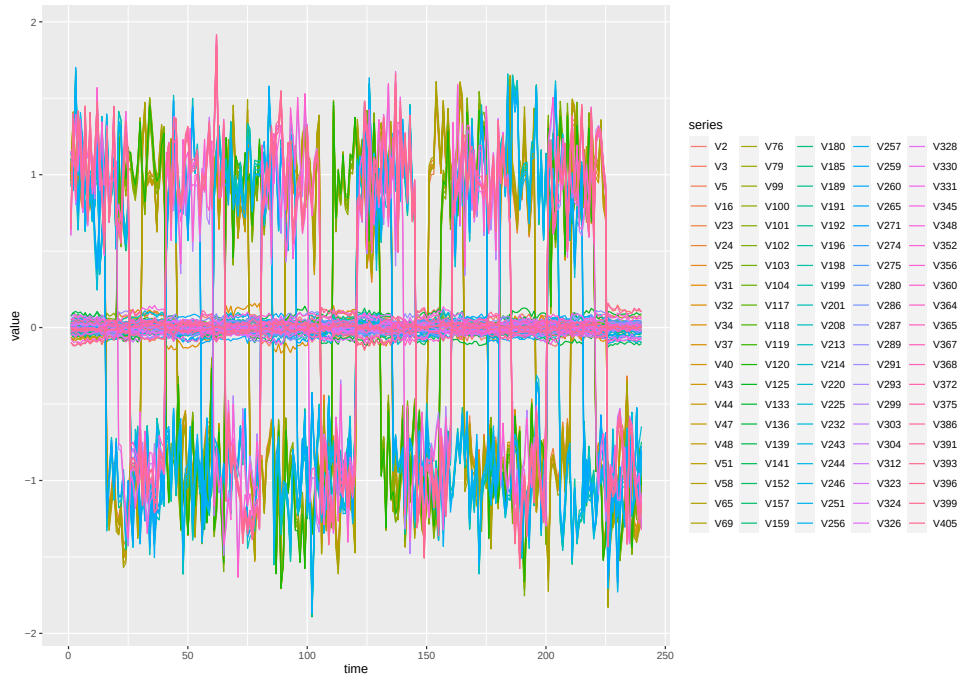
While this is a very granular heatmap, there is evidence of some correlation across the V number of variables with the product of the noise types.

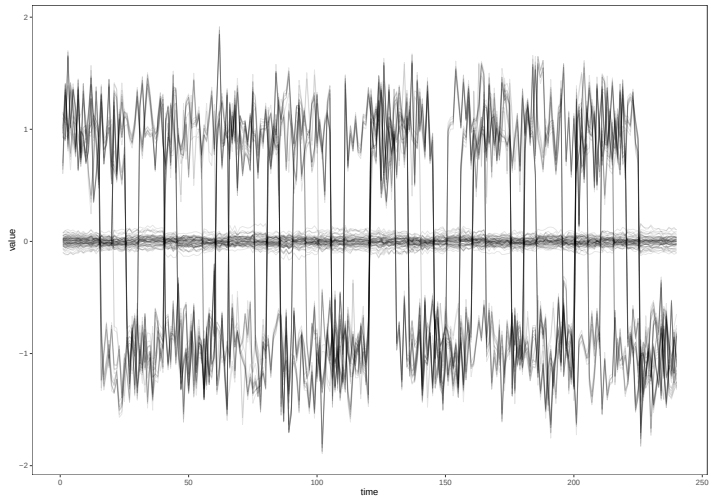
1.5

These products could exist if the noise matricies were transposed. In which case the resulting calculation

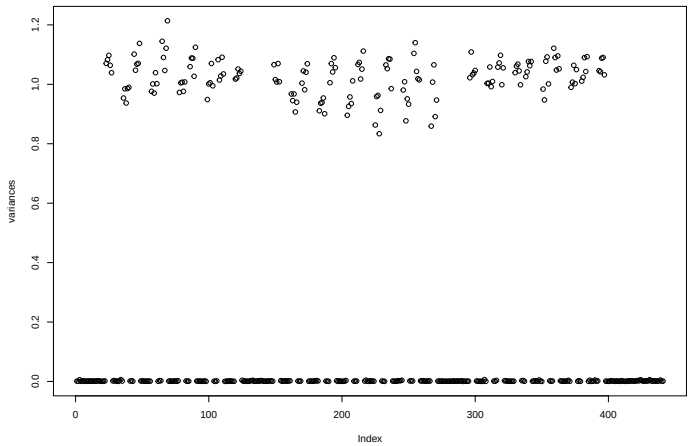
```
X = (TC %*% t(TC_noise)) %*% (SM %*% t(SM_noise))
```

would have non-conforming dimensions. Typically, noise is only multiplicative if the noise is correlated with the data; in this case, it is not. Hence additive noise is okay.
I have included two plots of 100 randomly selected time-series from $\mathbf{X}$:





A plot of the variance of the 411 variables:

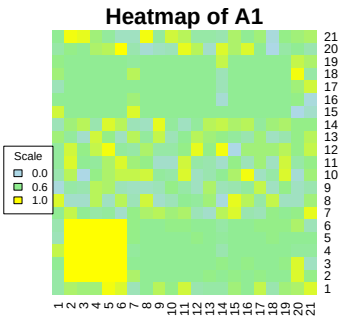


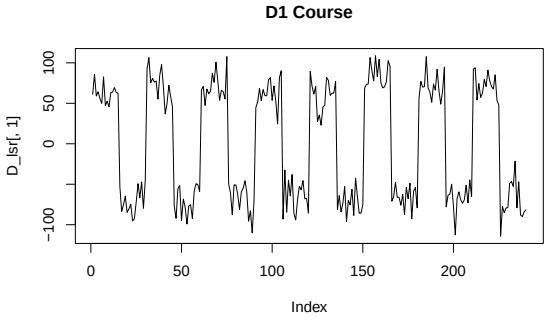
The variance plot displays some interesting information. There are two main clusters, variances of value $\sim$1, and of value 0.

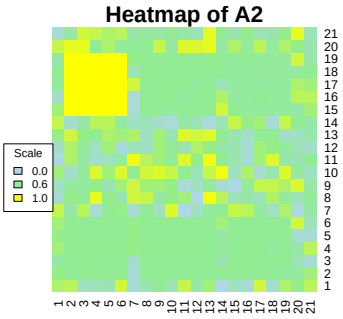
Question 2.

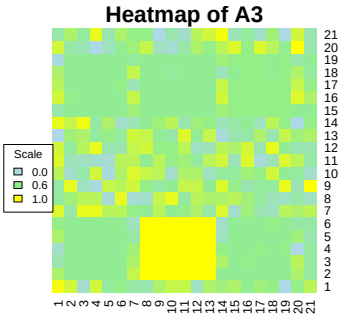
2.1

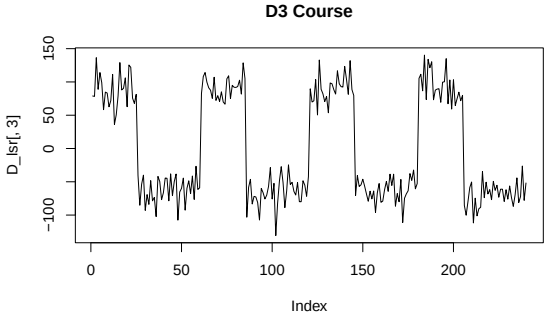
$\mathbf{A}_{LSR}$ and $\mathbf{D}_{LSR}$ retrieved sources plotted side by side:

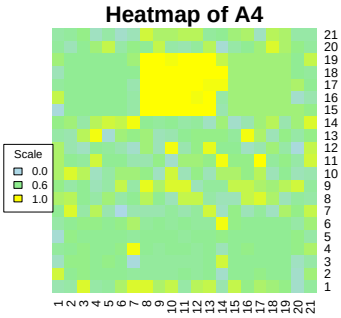


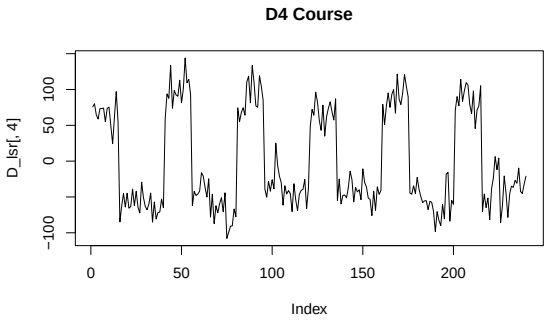


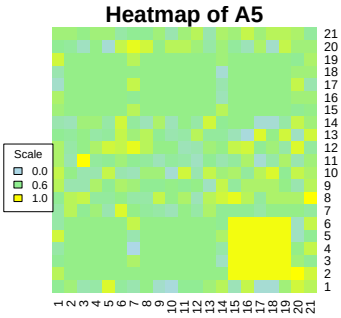


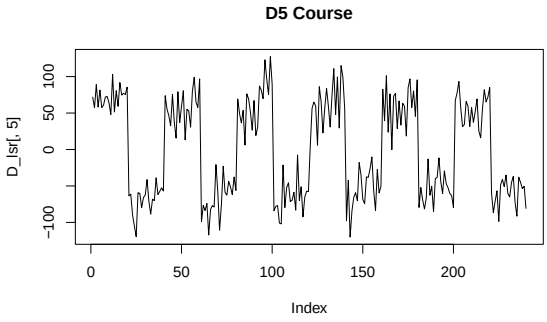


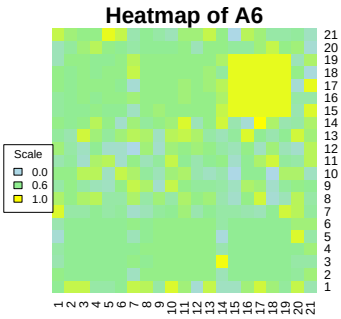


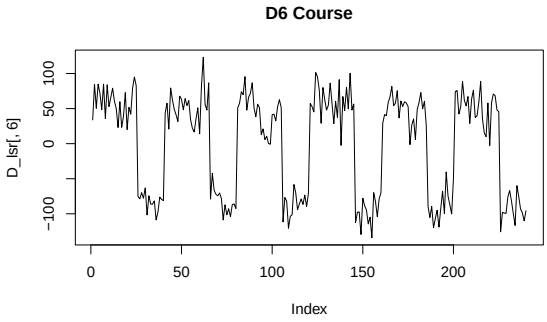




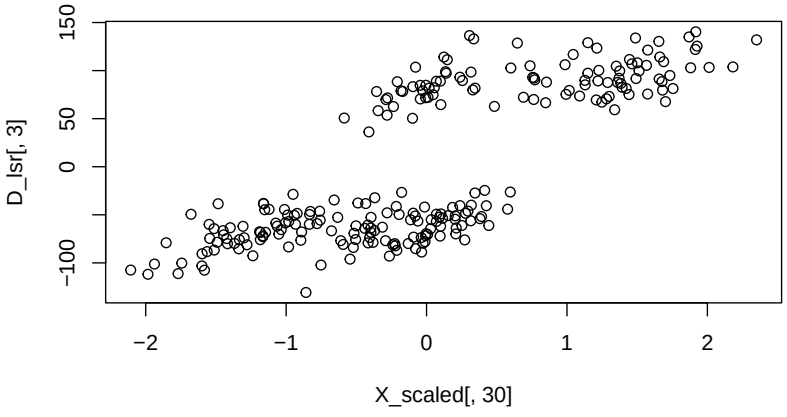




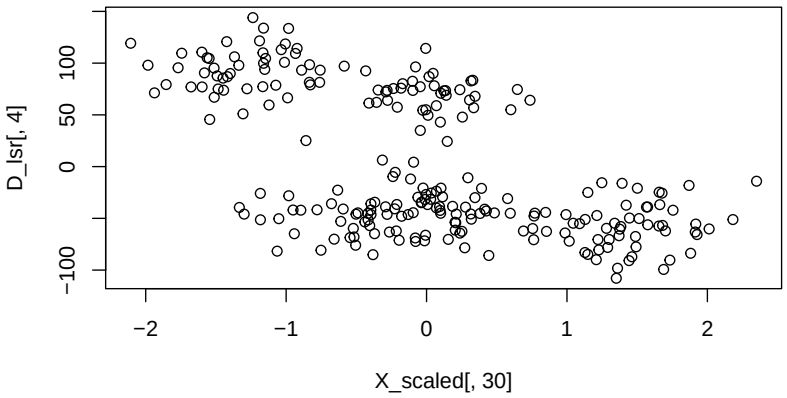




Scatter plot between the $3rd$ column of $\mathbf{D}_{LSR}$ and the $30th$ column of standardized $\mathbf{X}$:



When compared to the relationship the $4th$ column of $\mathbf{D}_{LSR}$ has with $\mathbf{X}$:



You can see there appears to be a more linear relationship in the first plot. This is due to the $30th$ column of $\mathbf{X}$ being constructed by the third $\mathbf{TC}$. As $\mathbf{D}_{LSR}$ is an estimate of $\mathbf{TC}$, this relationship is carried through and hence a more linear relationship is shown between $\mathbf{X}$ and $\mathbf{D}_{LSR}$'s third column than fourth.

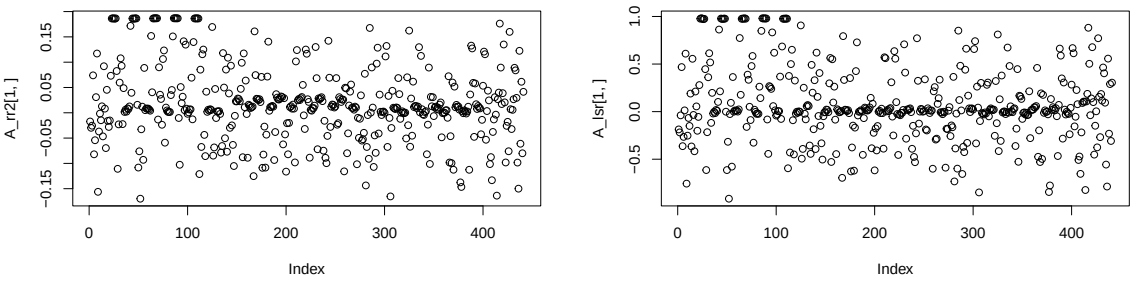
2.2

The sum of the C_{TRR} is greater than the sum of C_{TLSR}:

```
c(sum(c_trr), sum(c_tlsr))
```

```
[1] 12.36234 10.62412
```

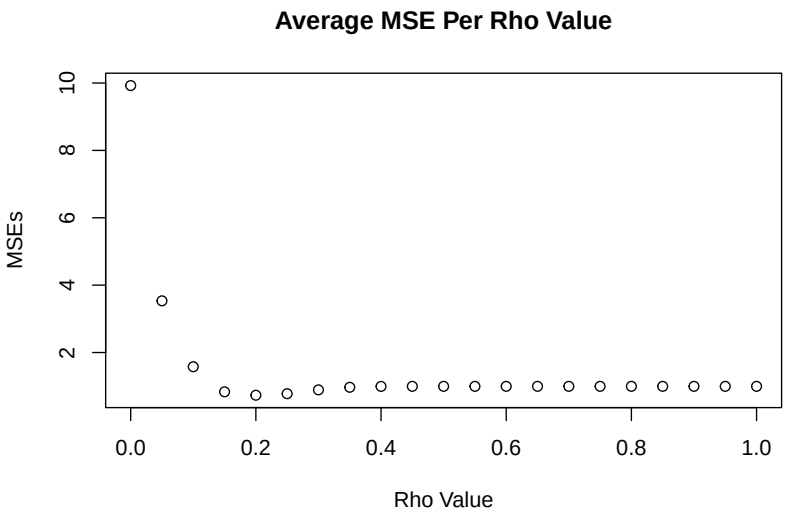
Now for $\lambda = 1000$:



These graphs show there is a majority of datapoints at 0.

2.3

The code to produce the MSEs through Lasso Regression that was used is from the Assignment sheet and translated into R.



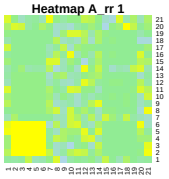
The minimum ρ is 0.2. LR diverged at ρ value of 0.4. Hence I will be choosing $\rho = 0.2$. Whether or not this was an appropriate decision is addressed in question **2.4**.

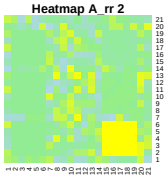
2.4

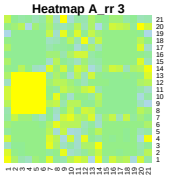
```
#i TC and DRR
C_trr = cor(TC, D_rr)
#i SM and Arr
C_srr = cor(SM, t(A_rr))
#i TC Dlr
C_tlr = cor(TC, D_lr)
#i TC and DRR
C_slr = cor(SM, t(A_lr))

c(sum(C_tlr), sum(C_trr))
c(sum(C_slr), sum(C_srr))
```

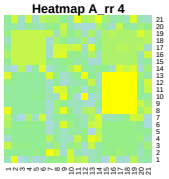
```
[1] 11.03923 12.36234
[1] 3.826995 3.346365
```

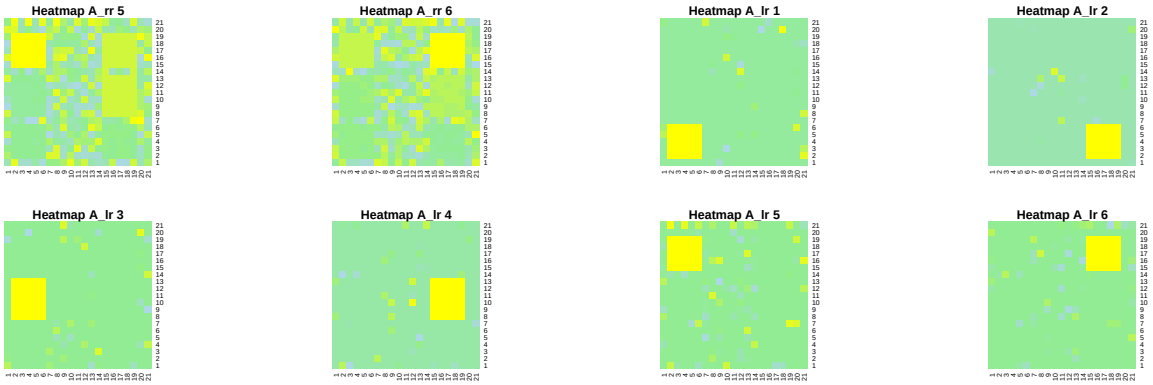
My $C_{TLR} < C_{TRR}$ which is not supposed to happen, however my $C_{SLR} > C_{SRR}$ which is in-line with the expected outcome. Hence my ρ value, while producing the minimum MSE, may not have been the best choice of value.



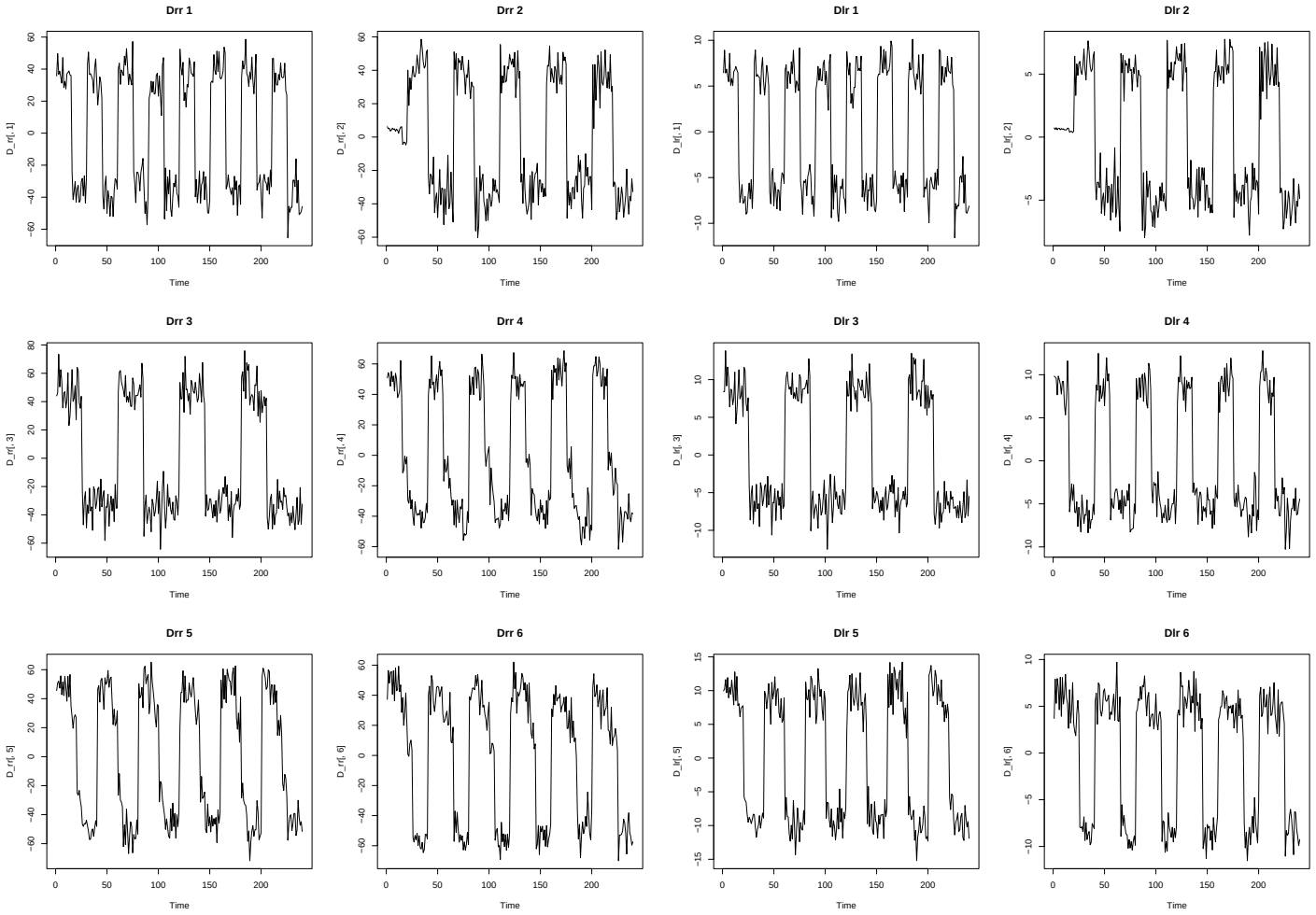








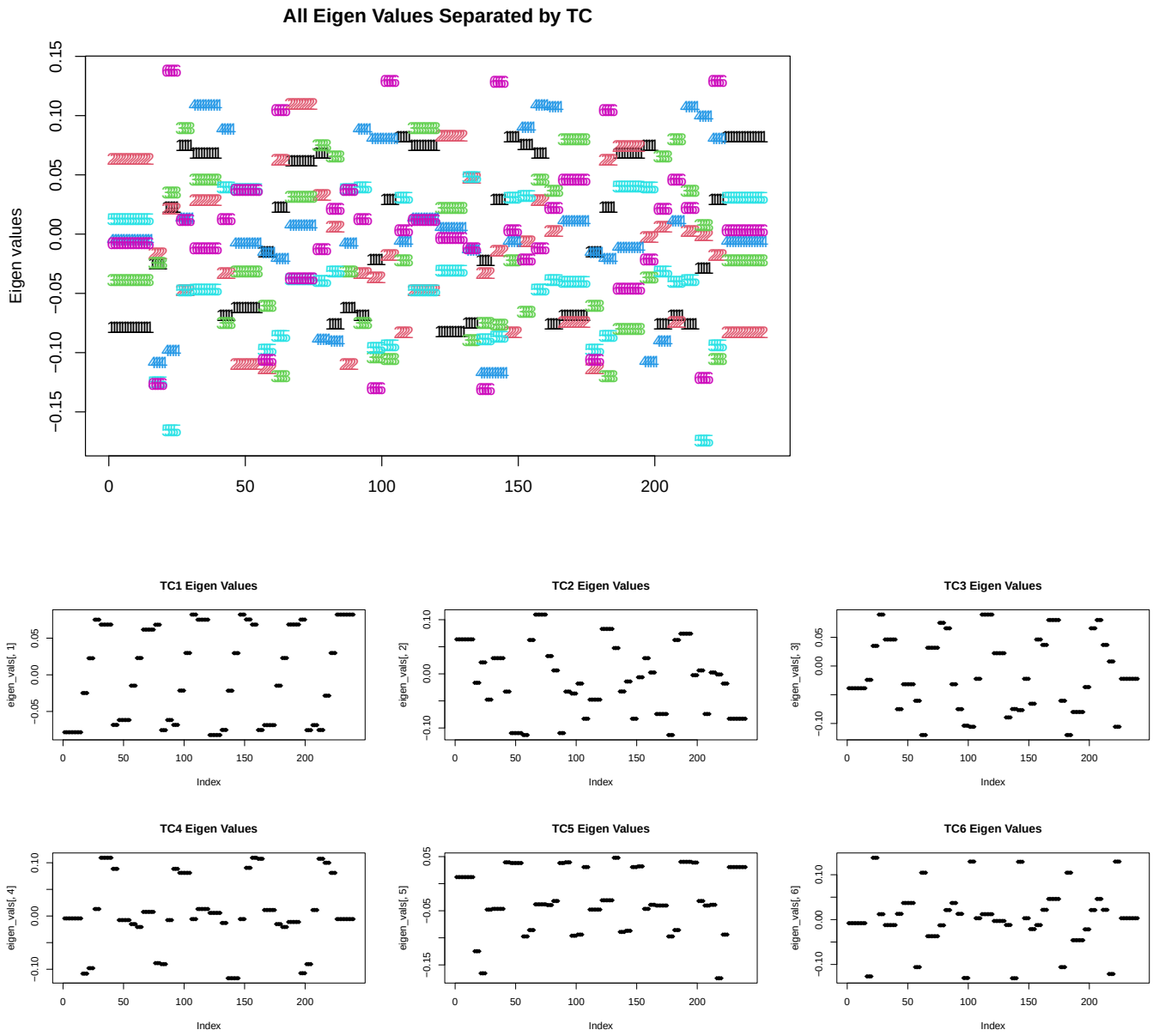
$\mathbf{D}_{rr}$ and $\mathbf{D}_{lr}$ are graphed respectively below.



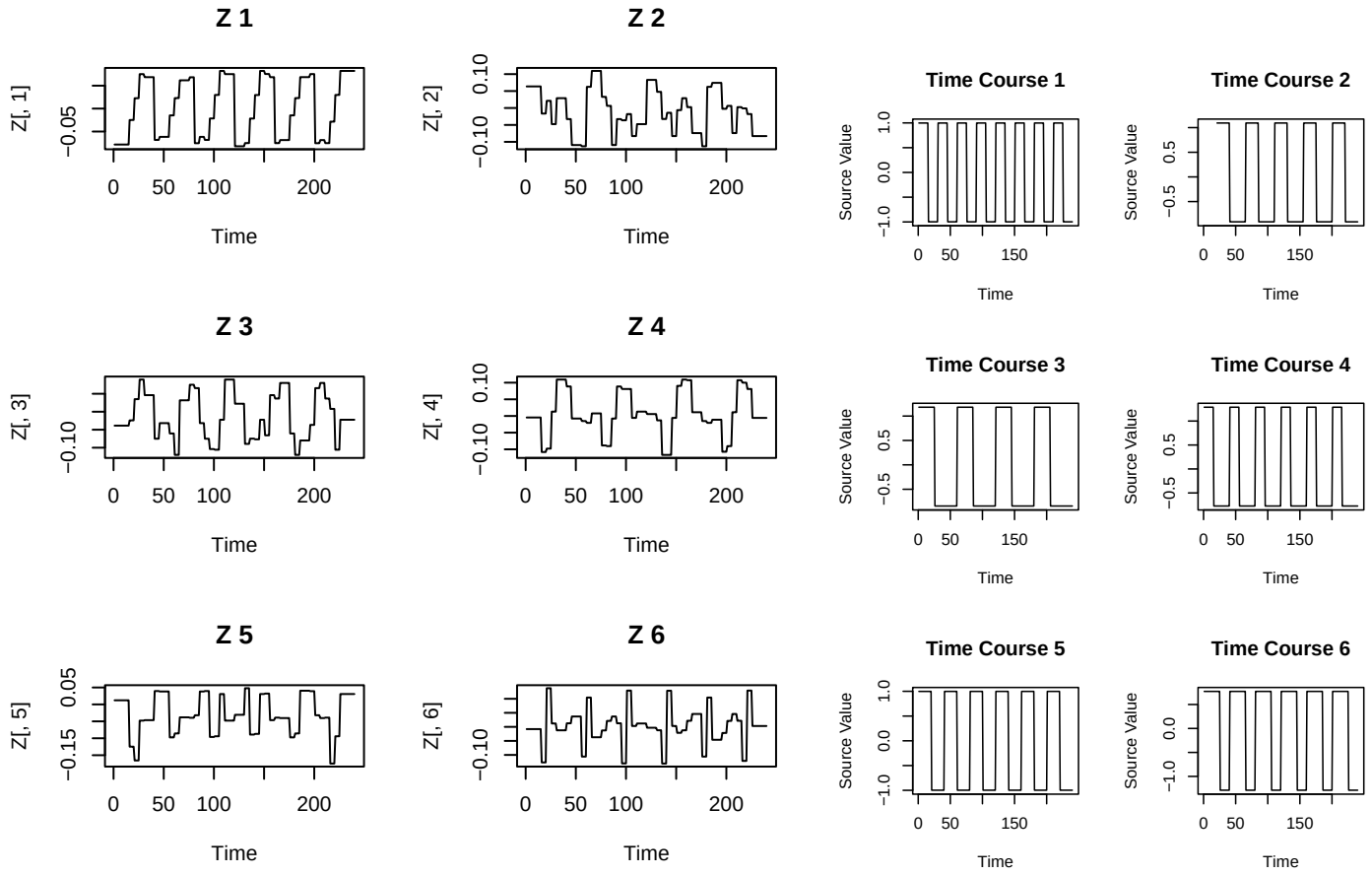
The reason behind the difference in the estimates of $\mathbf{A}$ is that LR has an optimised regularization term, ρ, which penalises the $l-1$ norm used to estimate $\mathbf{A}$.

2.5

Below are plots of the eigen values of the PCs. The first graph contains all 6 TCs' eigen values, and then they are plotted individually.

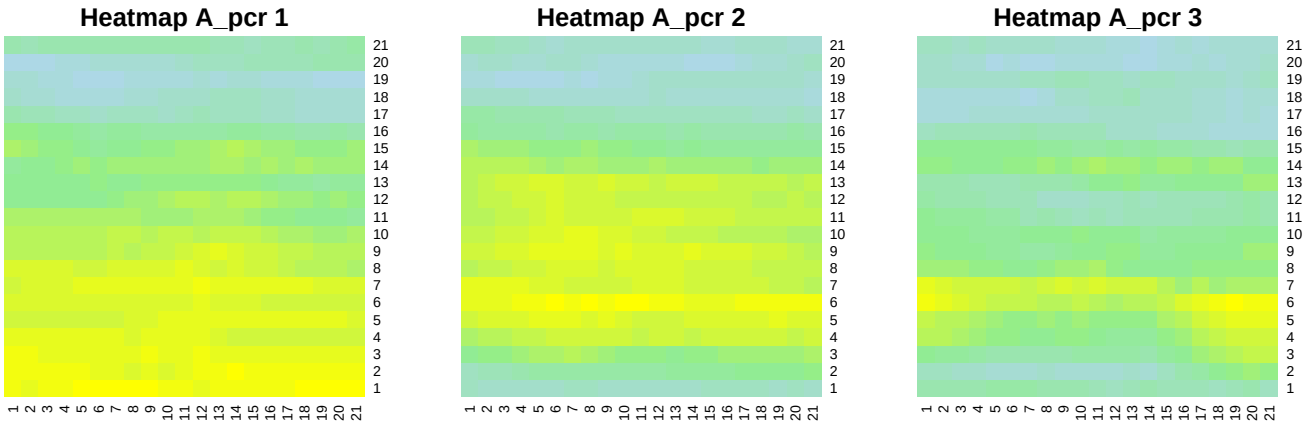


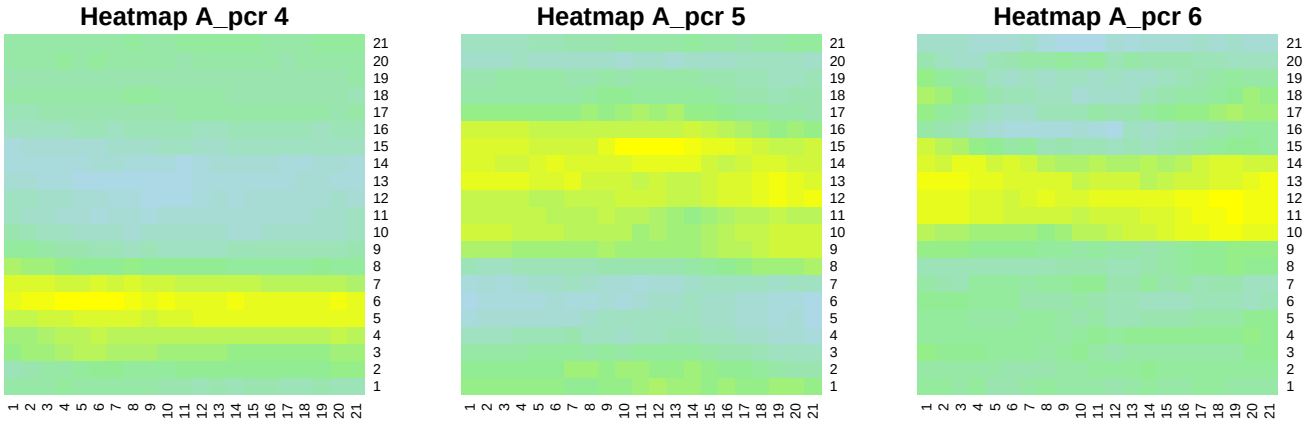
The fifth PC has the smallest eigen values.

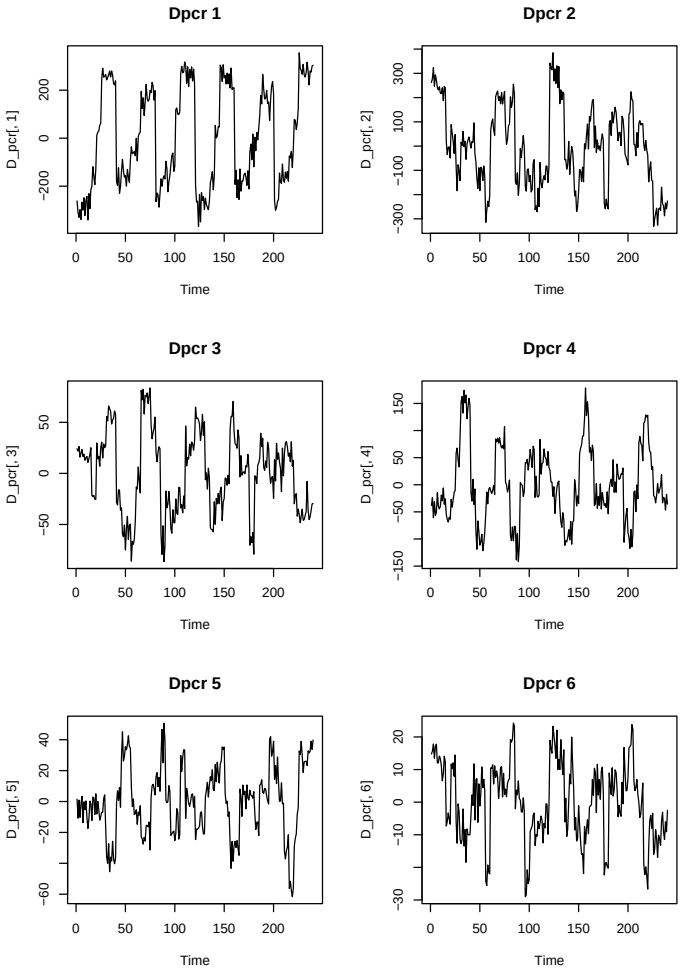


The deteriorated shape is due to the dimensionality reduction that is inherent to PCA.

Now after applying Lasso Regression on **Z**:




Heatmap A_pcr 4
Heatmap A_pcr 5
Heatmap A_pcr 6


Dpcr 1
Dpcr 2
Dpcr 3
Dpcr 4
Dpcr 5
Dpcr 6
Time

The PCR performed worse than the other three regression models largely to its focus on dimensionality reduction, which in the case of this dataset is not needed or used. Hence when performing PCR here, not much benefit is gained as we are not utilising the primary reason to use PCA, and hence it is distorting the shape of the TCs.