Multivariat øving 8

Siawash Naqibi

May 8, 2021

Start with PCA and see if there are any time dependencies

To see if there are time dependencies we start of by choosing a random variable and make a line plot out for this variable. We chose the variable 'wtc_SetAnemo_mean'. We can see in figure **??** that there doesn't seem to be a pattern in the data over time, meaning no time dependencies.

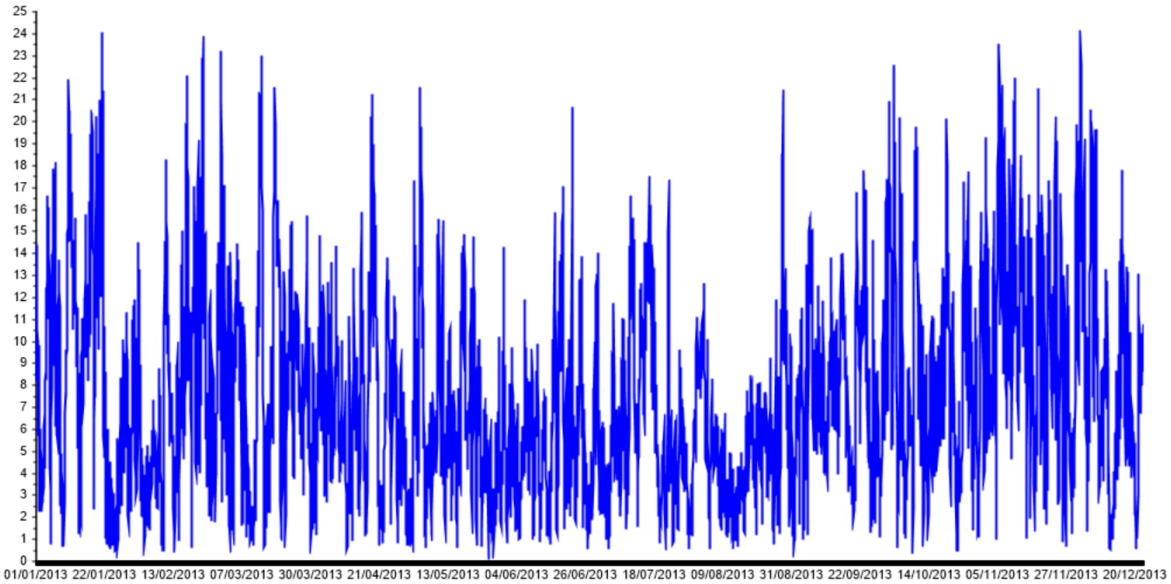

Figure 1: Line plot

We know have made a PCA model and will plot a line plot of one of the scores. We have chosen pc1, which counts for 30% of the total variation. On figure 2 we can see that there seems to be a slight pattern in the data. Therfore it seems like there are time-dependencies in the data.

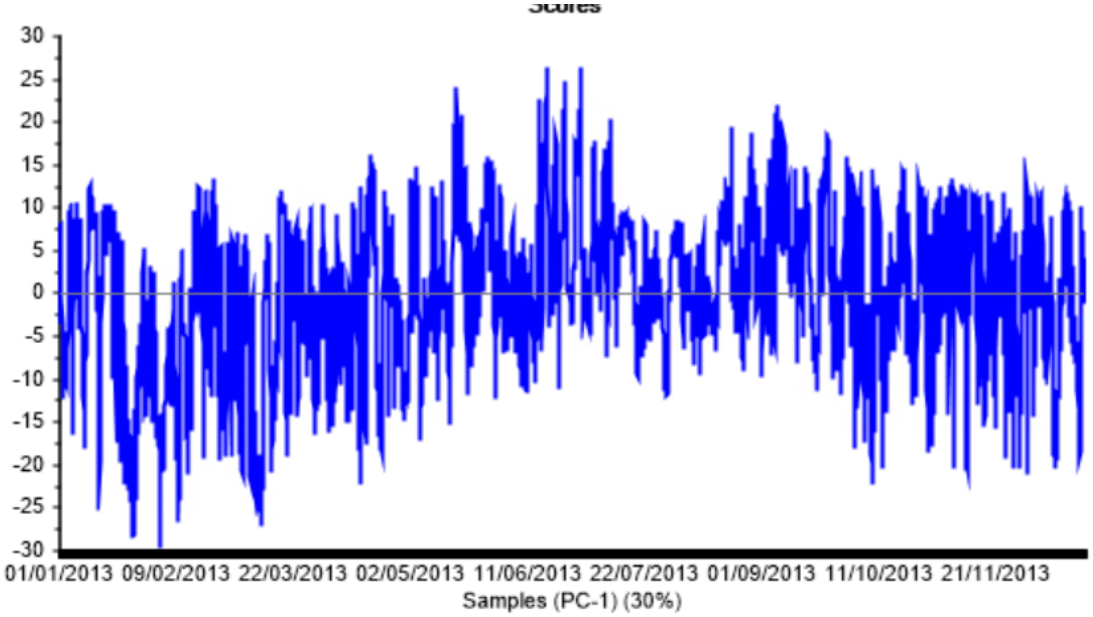

Figure 2: Line plot of pc1

Try both with PCR and PLSR. Are there differences correlation structure in the correlation loadings for the two methods?

The coreelations loading for PCA and PLSR are depicted in figure 3 and figure 4 respectively. We can clearly see that there are differences in the correlation structure in the correlation loadings for the two methods.

The first thing one notices is that the first principal components for PCA and PLSR doesn't explain the same amount of variance. This is because PCA finds the principal components in terms of highest variance in the X data, while PLSR finds the principal components in the data X that best explains the variance in the data Y. In other words, the PCs of PCA captuures the most variation in the X data, while the PCs for PLSR captures the most variation in the Y data.

When there is only one response variable Y, the y-loadings will always appear in the upper right quadrant. This is also what we see.

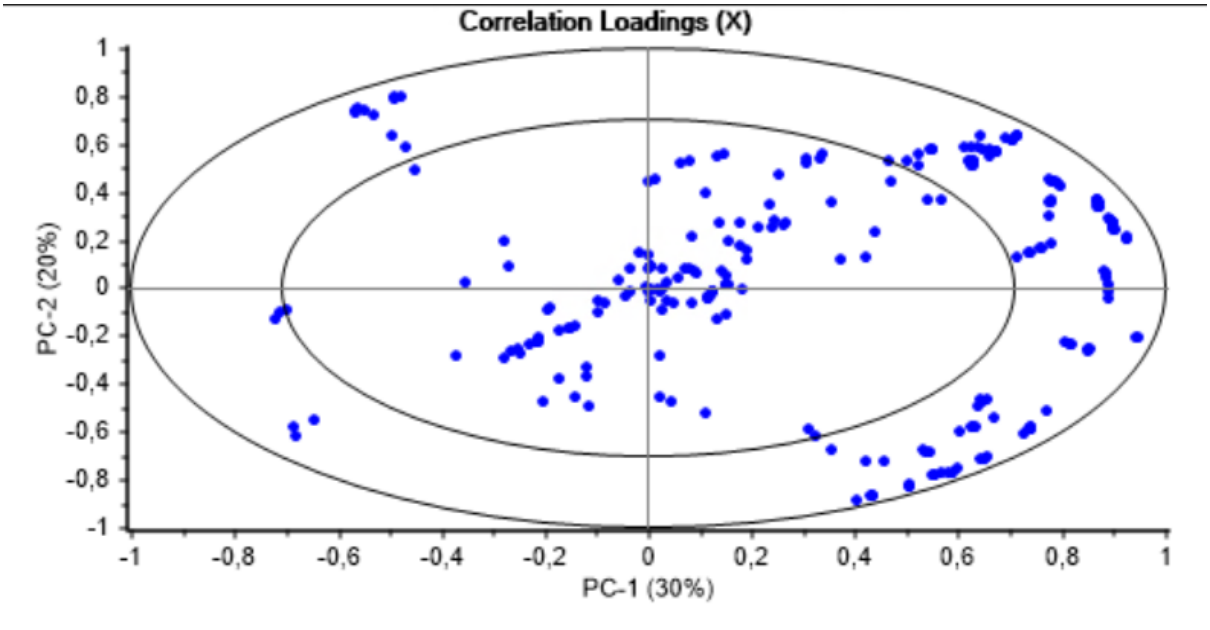

Figure 3: correlation loadings for PCA

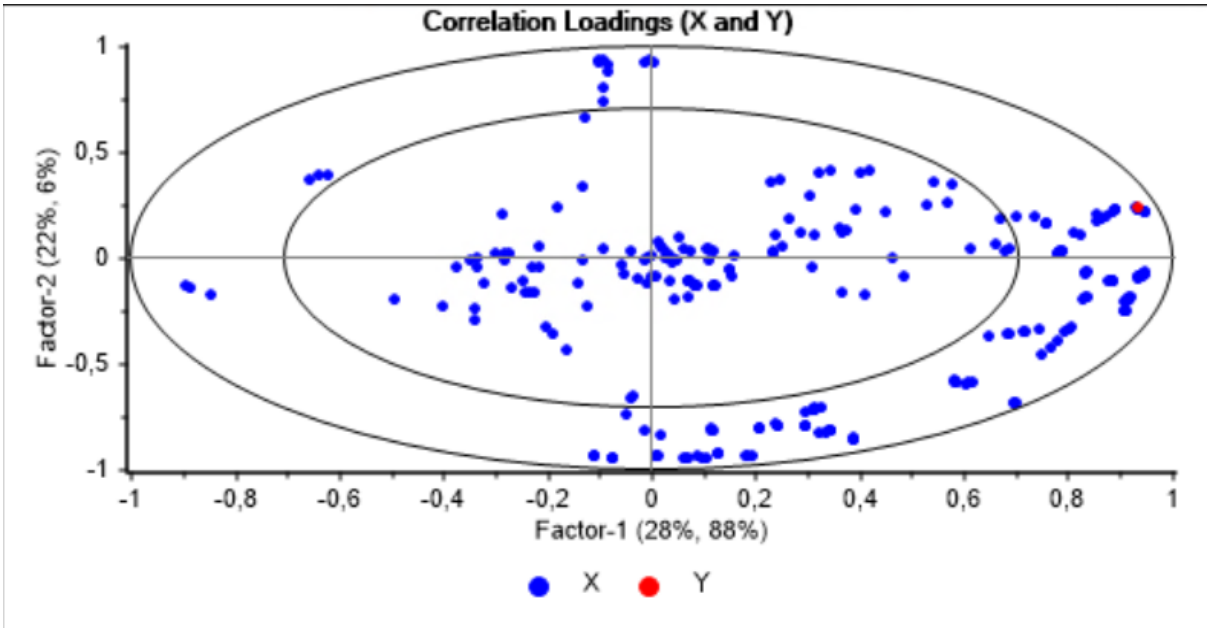

Figure 4: correlation loadings for PLSR

Make models with selected column subsets and the column set “AllX”; compare results

We will now make PLSR models for the column subsets to see if there are any differences with the PLSR model for columnset AllX.

PLSR models for some of the columns subsets are included below. We can see that although

the models look slightly different, there are no major differences. The best PLSR model was the PLSR model for the Grid column subset because it had an RMSE value of 79.6, whereas the PLSR model for the AllX columnset had an RMSE value of 140.8. This is kind of weird because PLSR should produce results as good as or better than its column subsets.

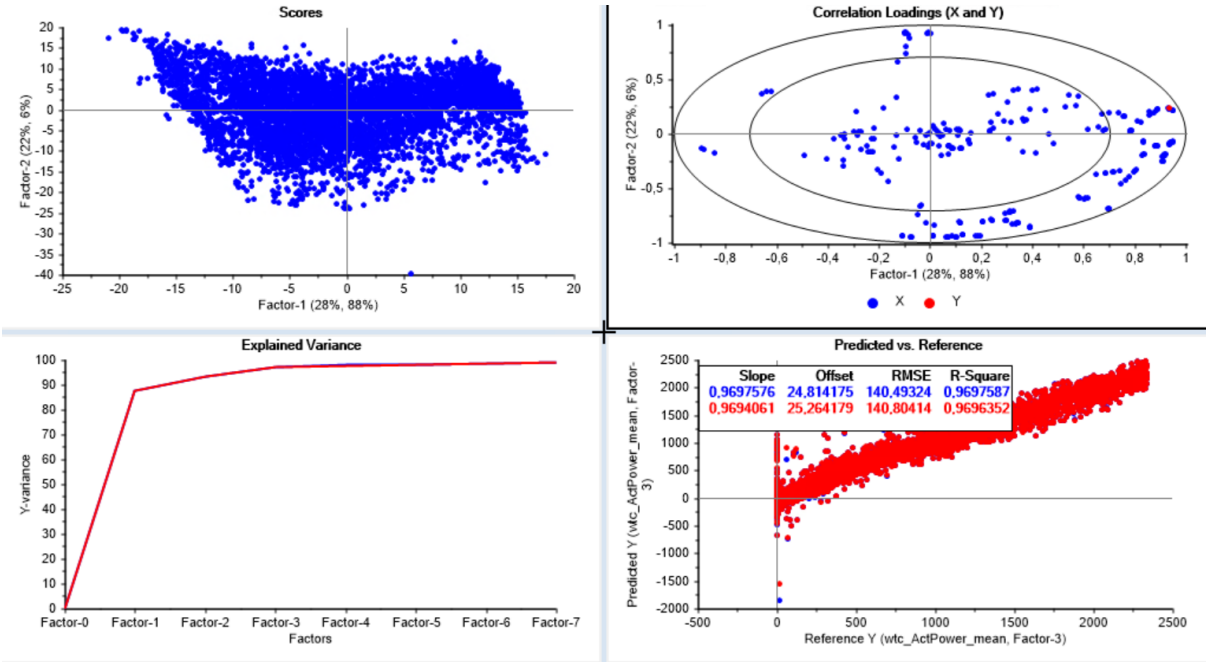

Figure 5: PLSR for columnset AllX

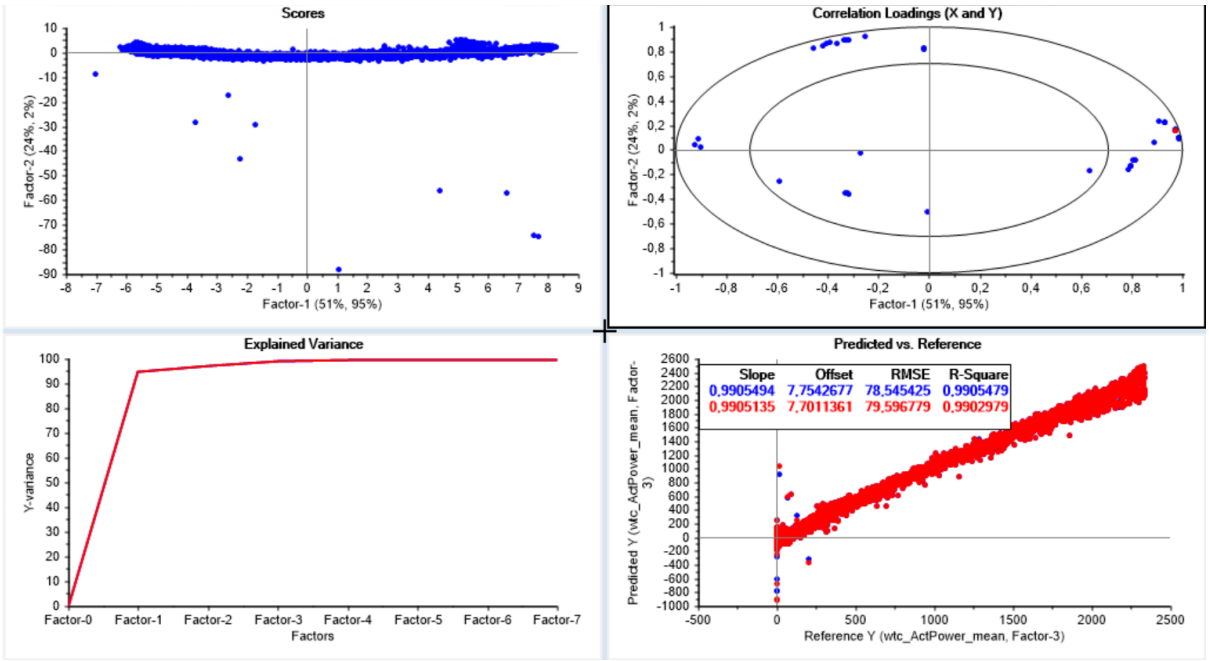

Figure 6: PLSR for columnset Grid

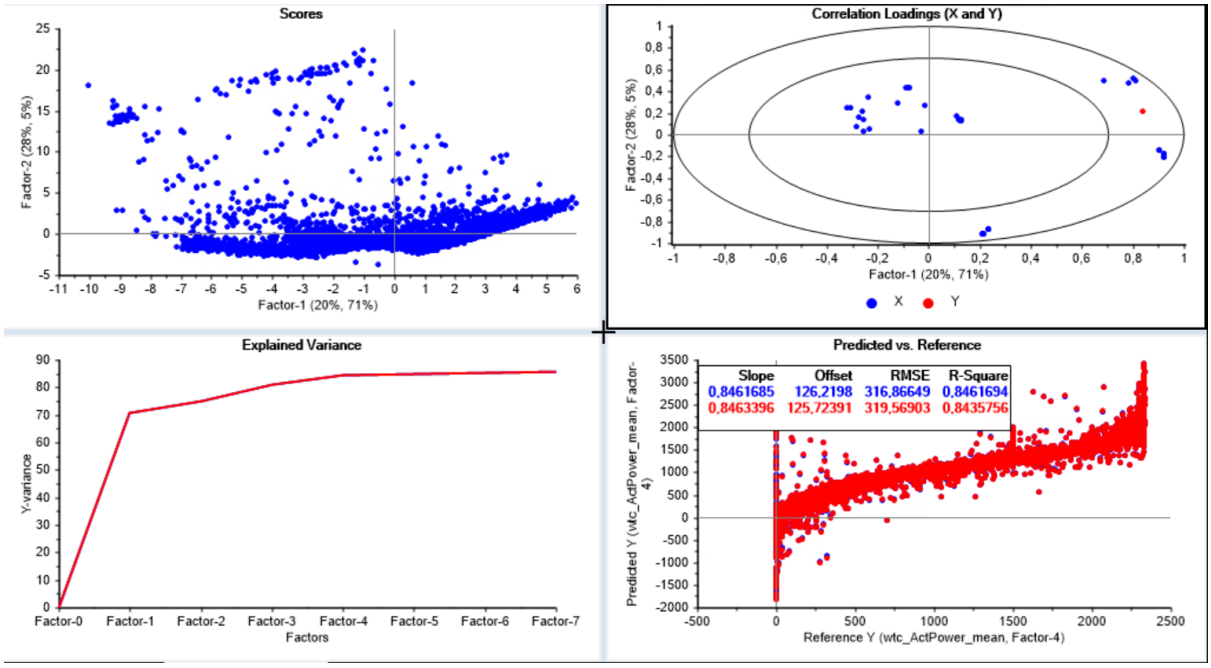

Figure 7: PLSR for columnset Turbine

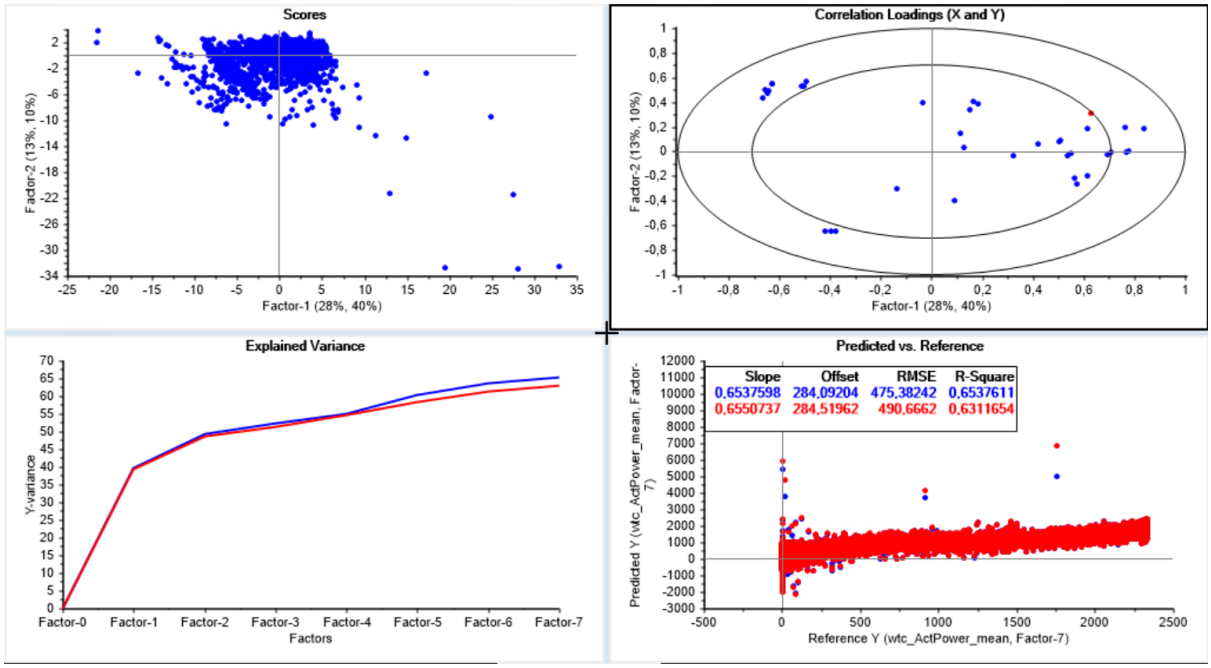

Figure 8: PLSR for columnset Pressure

Find the optimal number of PCs/Factors

For the PLSR model for the columnset AllX, one only need three PCs. Wtih three PCs one can explain 98% of the variation in Y.

Identify outliers and remove them if it can be justified. Does the RMSE improve after removing them?

Looking at the influence plot, the example 3740 should be removed because it has a high hotellings value and a high residual value. This means that it weights highly on the model and its residual is high as well.

Removing the outlier and recalculating reduced the RMSE value from 140.8 to 140.1, with three PCs.

Test different validation schemes, compare RMSE

We will test different validation scheeems for the PLSR model for the columnset AllX, with three PCs.

The first validation scheeeme is to choose 1/3 as Test set and 2/3 as training set. This yielded an RMSE value of 143.0.

The second validation scheme is choosing all data from January to July as training data, and all the data from August to Decembers as test data. This produced an RMSE value of 158.8

The last cross validation scheeme is by cross validating systematically with 112233 with a chosen segment of 20. This gave a validation RMSE of 147.6.

Select a subset of the variables by marking in the correlation loading's plot and/or regression coefficients plot

We have tried to reduce te number of variables, and found out that the model can be reduced to ten variables. The validation RMSE with three PCs came to be 100.5, and the predicted RMSE on the test data came to be 117.6.

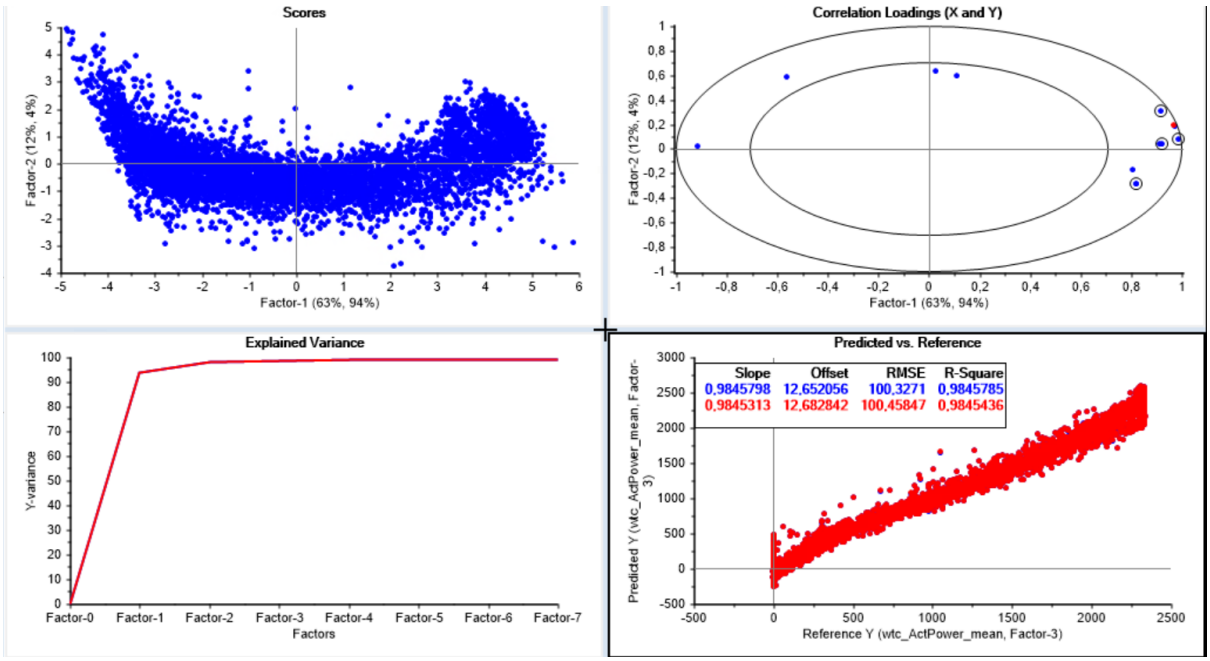

Figure 9: PLSR with feest number of variables