Défi Big Data

Data Science Mines St Étienne 2017/2018

This work will be assigned. You must provide a **report (pdf)** and your **R scripts** on Campus before Friday, 4^{th} of May at 6 p.m. You will work in groups of four. Your report should not only comment the results, but must give intuition and analyze them.

Dataset

This data set contains data from the bike sharing system of Paris, called *Vélib*. The functional data are loading profiles of the bike stations over one week. The data were collected in 2014 every hour from Sunday 1^{st} of September at 11 a.m. to Sunday 7^{th} of September at 11 p.m. These data can be found in the package 'funFEM' with the following command `data(velib)`.

The purpose of the lab is to identify some features characterizing typical velib stations and then to predict the loading profiles of the stations.

Questions

1. Visualize a few load curves over the 1189 curves availables. Why is the Fourier basis suitable to fit the data?

2. Consider the load profiles of the first week: from Sunday 1^{st} of September, 11 a.m. to Sunday 7^{th} of September, 10 a.m. **It consists of the first 168 records for each station.** Adjust these data using Fourier basis. How many number of basis did you choose? Display a few raw load curves and their decomposition in the Fourier basis.

3. Carry out a principal component analysis (PCA) on the adjusted data with the vanilla PCA and with the VARIMAX criterion. Interpret the results.

4. Using the kmeans method, propose a clustering of the adjusted data with 6 clusters over the Fourier basis and over the PCA decomposition.

   Visualize and interprete the resulting partitions and the group means. Visualize the group means obtained with 4 clusters. What information do you lose?

5. Using the previous clusterings, we want to predict the load profile of the data that have not been used yet: load curves of Sunday 7^{st} of September, from 11 a.m. to 11 p.m. **It consists of the last 13 records for each station, from the 169^{th} to the 181^{th} value.** Test the methods described thereafter.

   (a) ***Naive prediction method.*** The prediction $pred_i^i$ is given by the load of station i at $t-\omega$, where ω is the periodicity of the data.

(b) *Prediction methods using the clusterings.* The prediction $pred_t^i$ is given by the center of the cluster to which i belongs, at time $t - \omega$, where ω is the periodicity of the data. Test this method with the 2 clusterings obtained at question 4.

For each method, compute:

- the Root Mean Square Error (RMSE) at time t. RMSE at time t is given by

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(pred_t^i - real_t^i)^2}.$$

  where $pred_t^i$ is the load prediction for station i at time t and $real_t^i$ is the realization of the load for station i at time t and N is the total number of stations.
- the averaged RMSE over the prediction period;
- the standard error of the RMSE over the prediction period;
- the averaged RMSE over the prediction period for each cluster.

Discuss the performance of the methods, globally and clusters by clusters.

Bonus. Propose another prediction method and compare it to the suggested methods.