CE263N/CYP 257: Assignment 4: PCA/Eigenbehaviors: identifying structure in routine

DUE ON 11/09/2020

PROF. M.C. GONZALEZ

Problem 1: Eigenbehaviors (100 pts)

Use the activity matrix of subject 4, and answer the following questions:

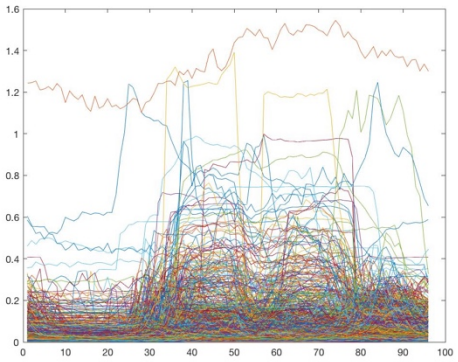
- How the first 3 eigenvectors for the chosen subject relate to the behaviors seen in days 10, 15 and 20 of this subject. Do the projections to answer this question. (10 pts)
- Draw the reconstruction of these three sample days with the first three eigenvectors. (10 pts)
- What percentage of the variance of the entire data the first 3 eigenvectors account for? How many eigenvectors do you need to reconstruct each of the 3 sample days with more than 75% accuracy? (10 pts)
- Can you identify a day that is the worst reconstructed by the first 3 eigenbehaviors? justify your answer (10 pts)
- Plot the data projected in the first vs. the Second PCA, coloring the scatter plot according to the Euclidean distance to the mean. Present two scatter plots, using different number of components for the PCA, one scatter plot with number_c= 2 and another using number_c=5, what is the difference between the distances? (20 pts)
- Again use number_c= 2 and number_c=5 for the PCA decomposition, project the data in number_c PCs and calculate Kmeans clustering in each case. Plot the square distance vs. numbers of clusters K for each case (20 pts)
- With the two plots of the previous question justify one selection num_c to decompose the data and K to cluster the data based on a small squate distance and a small number of clusters. With that section, plot the average day of the members of each cluster (K plots of mean(Mbw) vs. time) (20 pts)

Note: The data is in realitymining (and realitymining.mat)
Upload the ipynb script that produces each answer

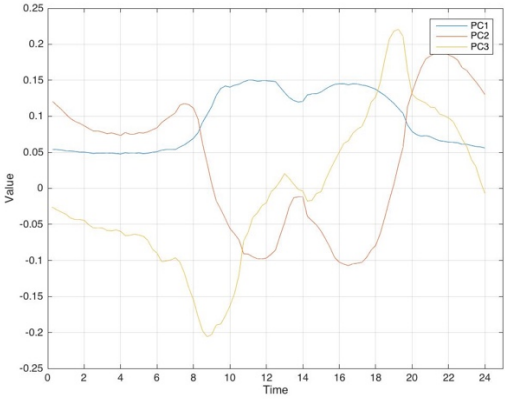
Upload the the answers to each question in a pdf file: I will evaluate the written solutions and check the code for reference.

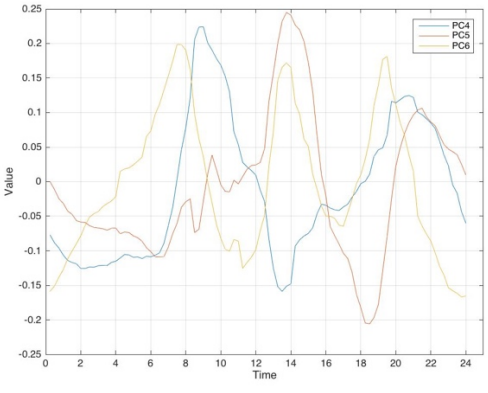
Clustering Electric Consumption with PCA (50 pts)

1) Load and plot data provided in `TypicalWeekdayProfile.txt`, it has the average reads of energy consumption in kilowatts [KW]each 15 minutes interval during one day for several accounts. How many accounts are given? and What is the dimension of the data? (10 pts)



2) Plot the first 6 eigenvectors, convert the x axis in a range from 1 to 24hs (10 pts)





3) How many K eigenvectors are needed to explain at least 92% of the variance? (10 pts)

4)Use that number to apply Kmeans with K equal to the answer above, show the clusters given by the method coloring each cluster differently in the scores of PC2 vs. PC1 space. (10 pts)

5)Plot the data of the original accounts separated in K subplots. What can you learn from the accounts that belong to each of the K clusters (see a sample solution with K=4) (10 pts)