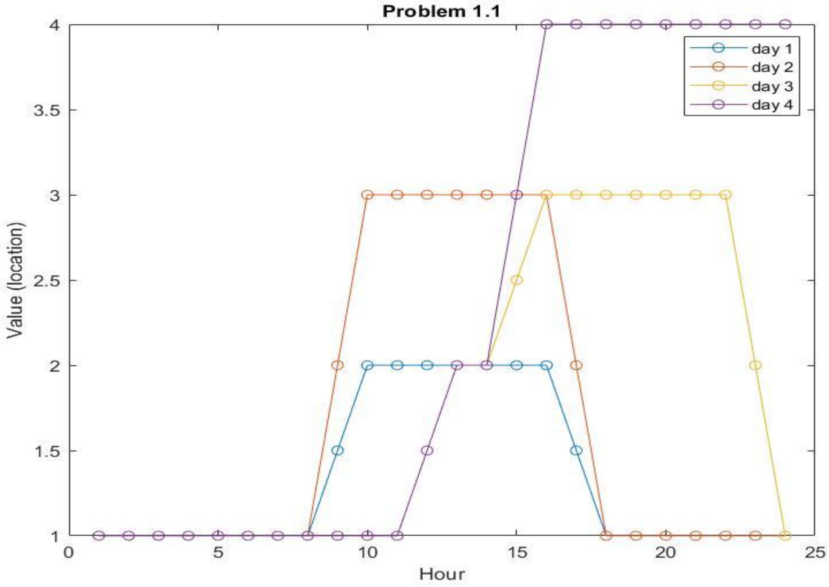
Aaron Scherf 9/23/19

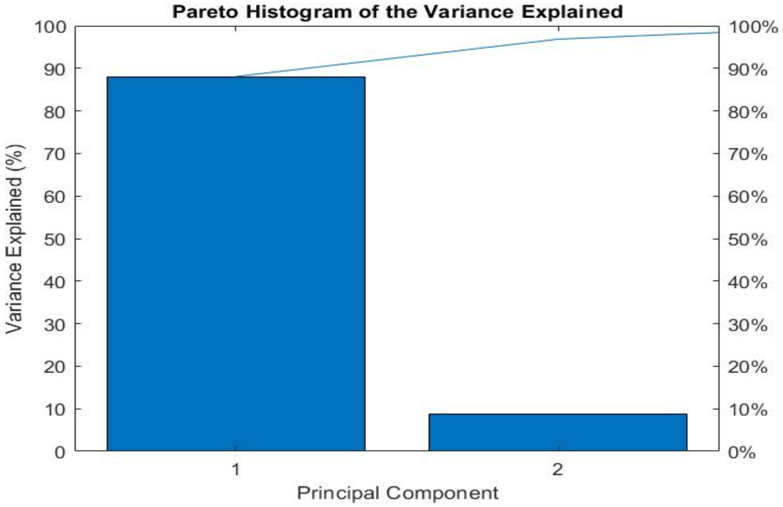
CYPLAN 257: Assignment 1

Problem 1: Basics of PCA

1) Use the script “simple_schedulePr1.m” and plot each day of data of a simple schedule of 4 days (1pt)

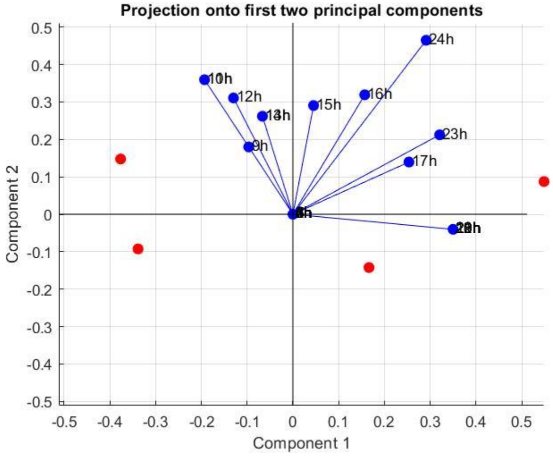


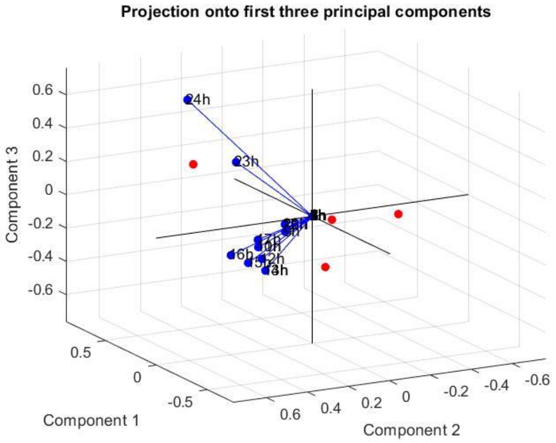
2) Calculate the PCA on this matrix and plot the pareto histogram of the variance explained. How much variance is explained by the first three eigenvectors? (1pt)

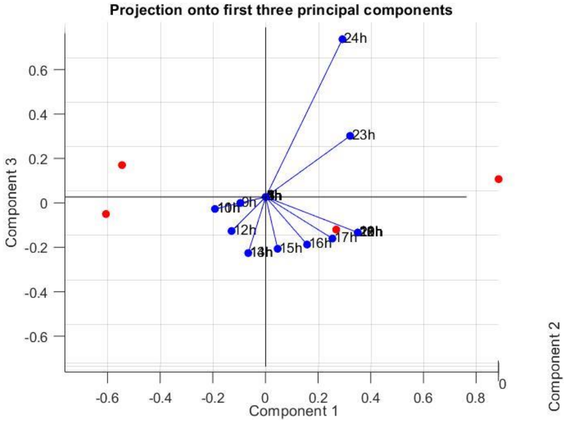


Summing the first three values for the `explained` attribute of the PCA in Matlab indicates that 100% of the variance in the schedule data that was generated is explained by the first three eigenvector components. This is reinforced by the pareto histogram, which shows that the first principal component explains roughly 89% of variance and the second another 9%.

3) Plot the biplots of the scores and component projections of the data. Do they have similar projections? (1pt)

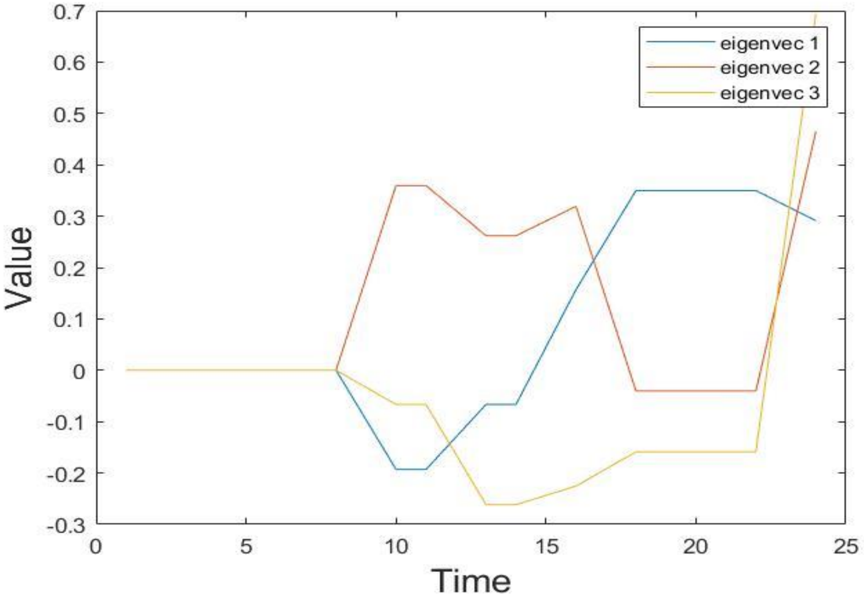




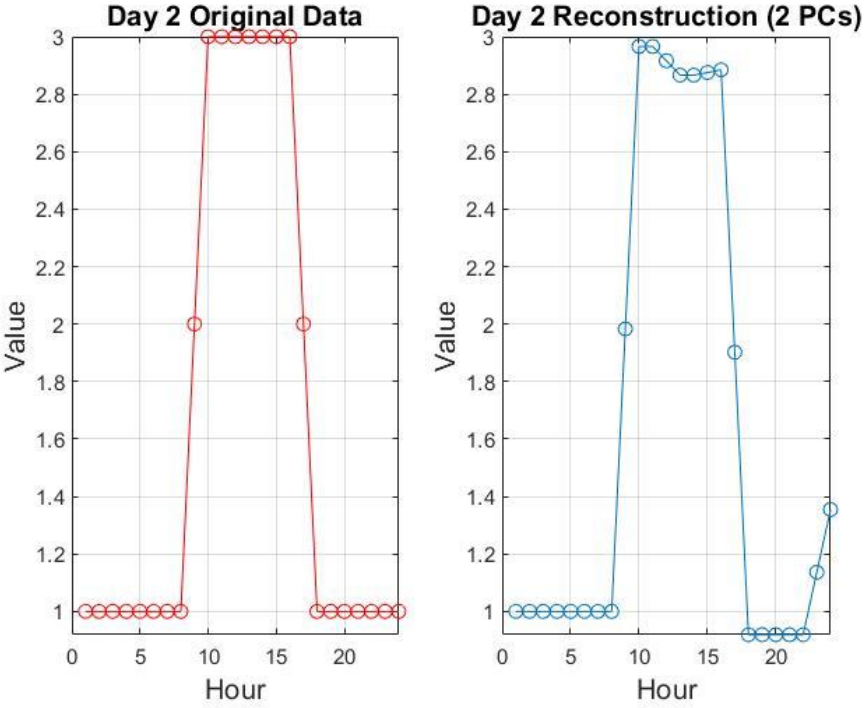


According to the biplots above, the component projections fall within a similar range as the scores, though they are not very proximal to any of the observed scores (in red). The area of the two-dimensional plot of PC 1 and 2 where the component vectors are concentrated seems to be slightly offset from the centroid of the four scores combined. Likewise, in the three-dimensional plot of the first three components, the space where the vectors are concentrated does not seem proximal to the points, thought shifting perspective to the first and third components shows that more vectors are proximal to the scores.

4) Plot the 3 first eigenvectors (1pt)

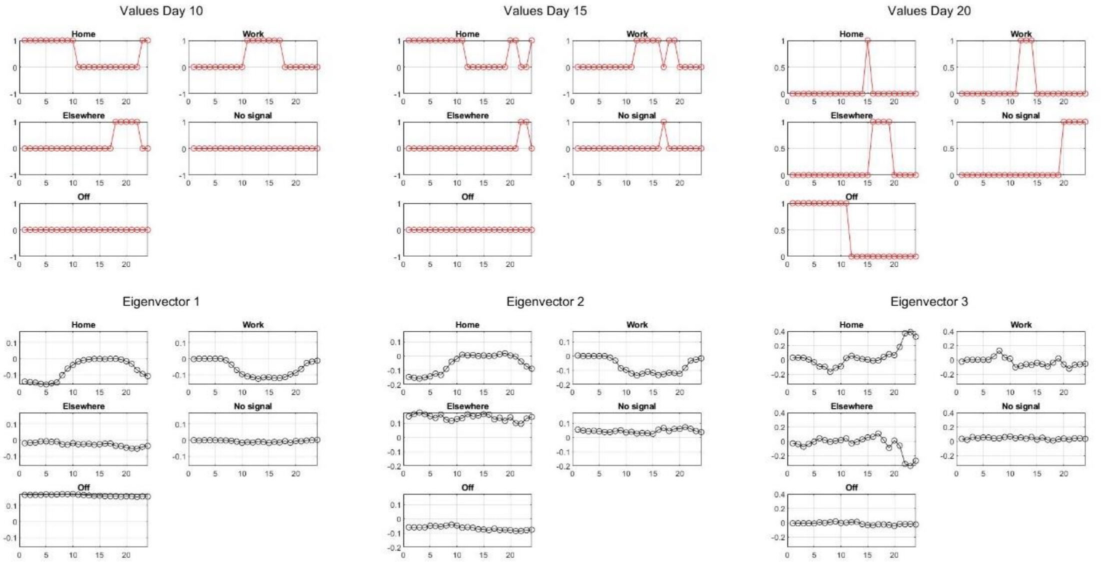


5) Reconstruct day 2 with the first 2 eigenvectors, showing data vs reconstruction (1pt)

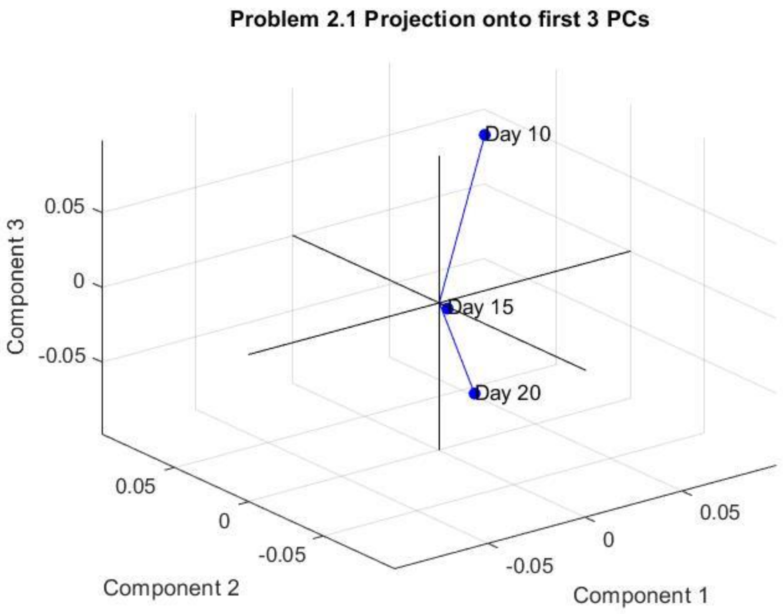


Problem 2: Eigenbehaviors

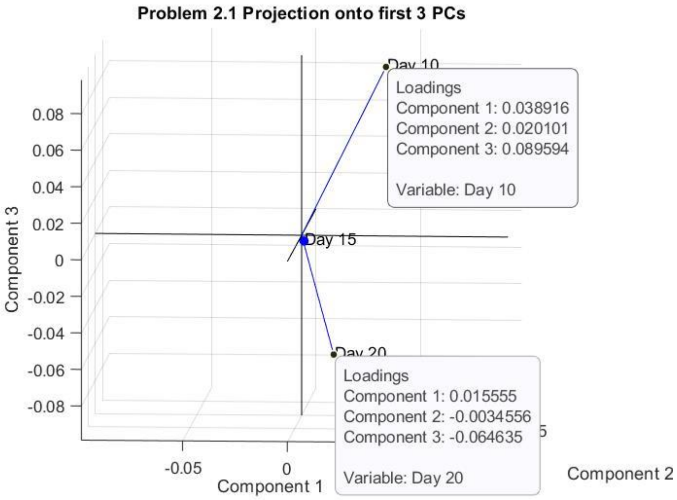
1) How do the first 3 eigenvectors for the chosen subject relate to the behaviors seen in days 10, 15, and 20 of this subject (2 pts)



The behaviors for subject 4 on days 10 and 15 seems inverse to the first two eigenvectors. Day 20 seems to be dominated by unusual "Off" or "No Signal" values, making it difficult to compare predicted behavior to this day. The first two eigenvectors seem to correspond to behaviors for days when subjects are more likely to stay at home during the day than go to a separate workplace, with eigenvector 2 also showing a strong affiliation with the "Elsewhere" category. Neither of these predictions matches the observed behavior for days 10 and 15. Eigenvector 3, on the other hand, seems to predict a propensity to return home at the end of a night, which more closely fits days 10 and 15.

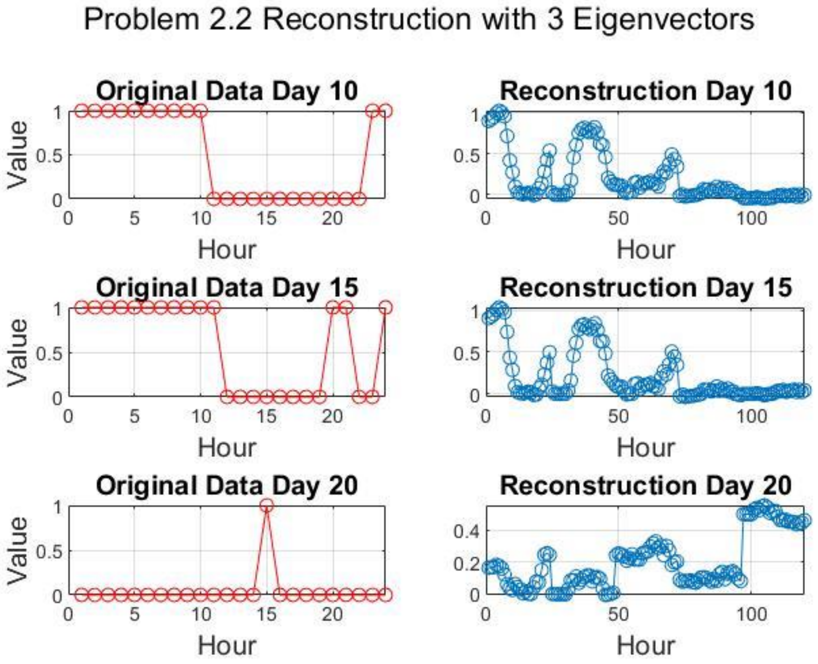


Plotting the principal components against the days in a bi-plot confirms their inability to predict day 15, which has an unusual work-home pattern, hence why it has low values for all 3 PC's.



Inspecting the component loadings for each day in the bi-plot, we see that components 1 and 2 have relatively little association with days 10 and 20, though component 3 has a higher coefficient value. This is likely due to the stability of the "Low-Elsewhere, High-Home" predictions generated by component 3 across both days.

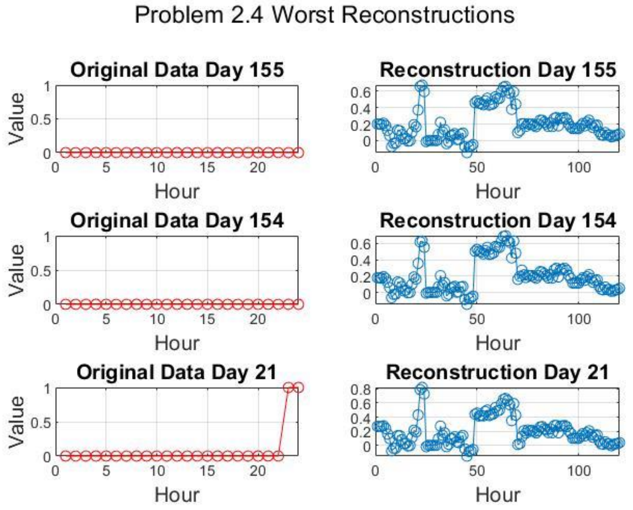
2) Draw the reconstruction of these three sample days with the first three eigenvectors (2pts)



3) What percentage of the variance of the entire data are accounted for by the first three eigenvectors? How many eigenvectors do you need to reconstruct each of the 3 sample days with more than 75% accuracy? (3pts)

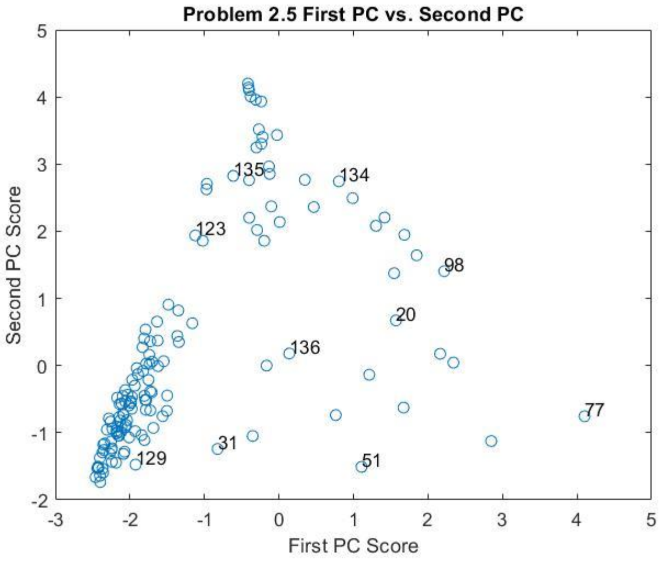
The first three eigenvectors account for approximately 53.84% of the variance in the dataset (code line 307, summing the first three elements of the “explained” object). Using a loop to iterate over eigenvectors and reconstructions until 75% accuracy is achieved, we find it takes 42 eigenvectors to reach this threshold (code lines 313-326).

4) Can you identify a day that is the worst reconstructed by the first 3 eigenbehaviors? (3pts)



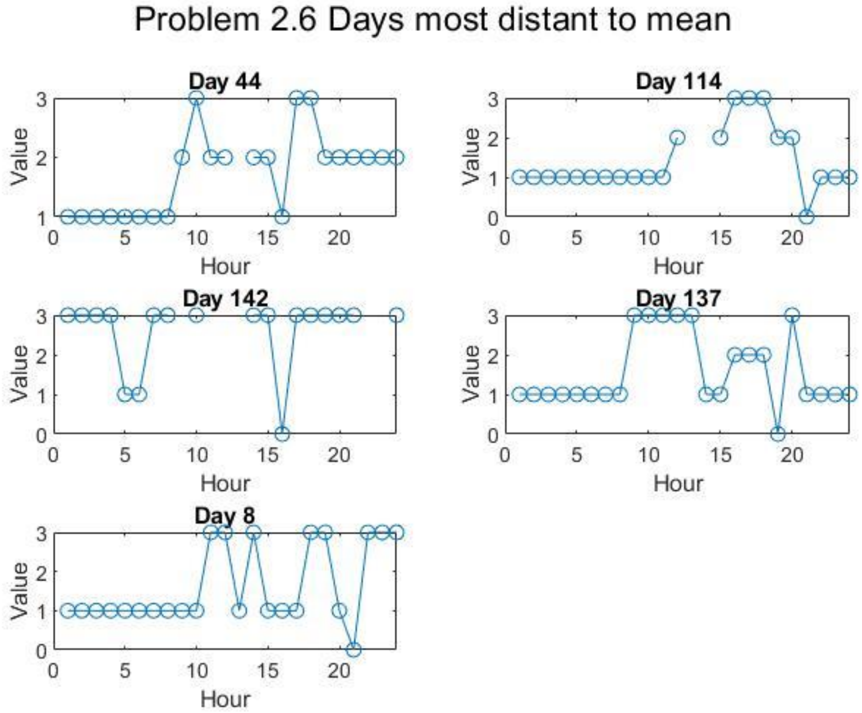
Calculating the error between original days and reconstructions using the norm of the vector, we find that day 155 has the worst reconstruction, likely because the original data has almost no variation, making it difficult to match with the regular signal predictions of the eigenvectors.

5) Plot the first vs second PCA and mark a few days using `gname` (3 pts)

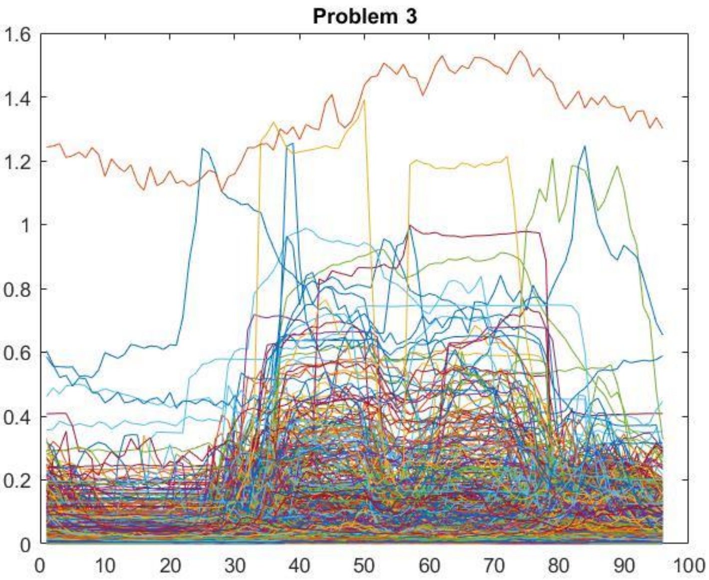


6) Based on the t-squared what are the 5 days most distant to the mean? Plot them as the value of the component vs time (2pts)

   The 5 days most distant to the mean according to their t-squared values are: 44, 114, 142, 137, and 8. There are, however, 8 days with equivalent t-squared values, including the 5 plotted below, meaning any of the following could also be considered most distant: 157, 163, 20.



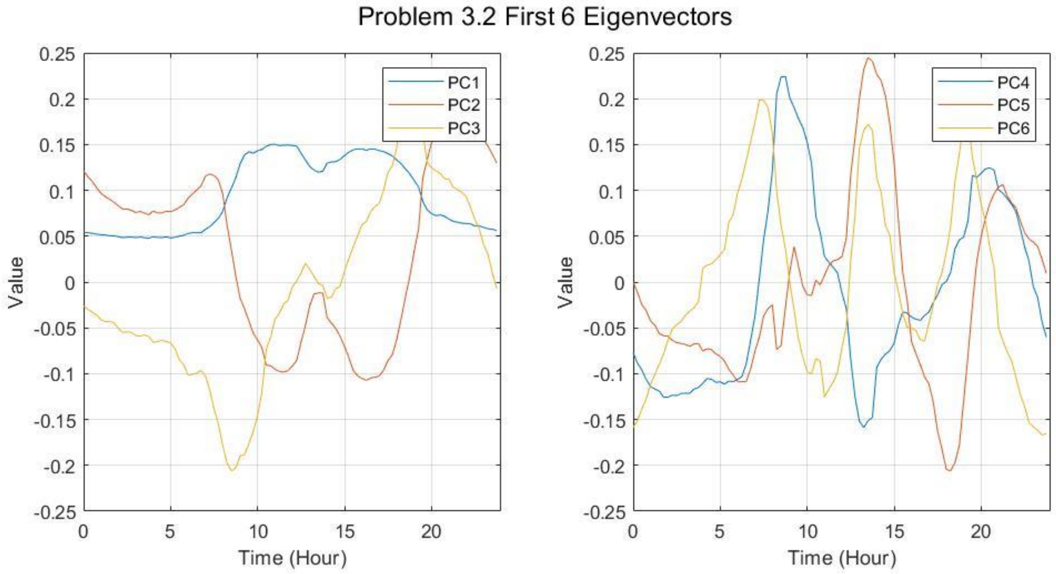
Problem 3: Clustering Electric Consumption with PCA



1) How many accounts are given? What is the dimension of the data? (1pt)

There are 1,255 accounts provided in the data set, corresponding to the 1,255 rows in the data. The dimensions of the data are therefore 1,255 rows by 96 columns, with the columns representing 15-minute segments of time in a 24 hour day.

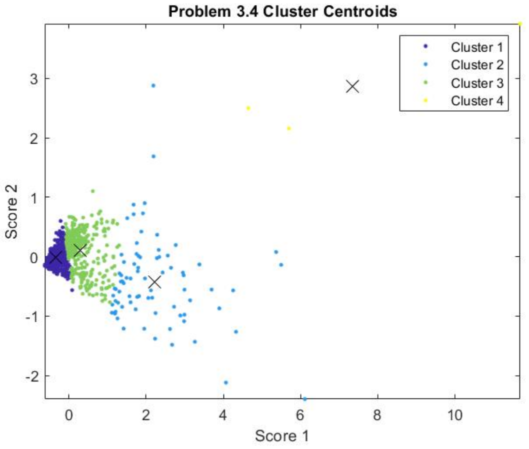
2) Plot the first 6 eigenvectors, convert the x-axis in a range from 1 to 24 (1pt)

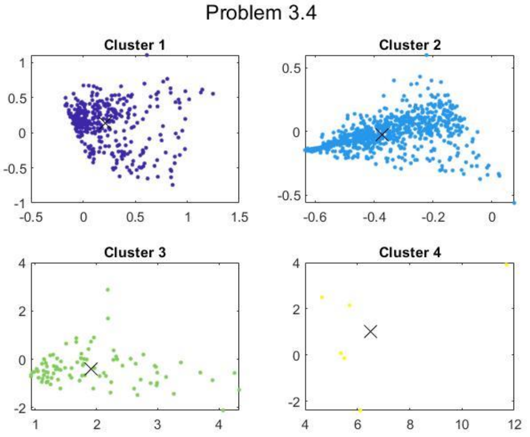


3) How many K eigenvectors are needed to explain at least 92% of the variance?

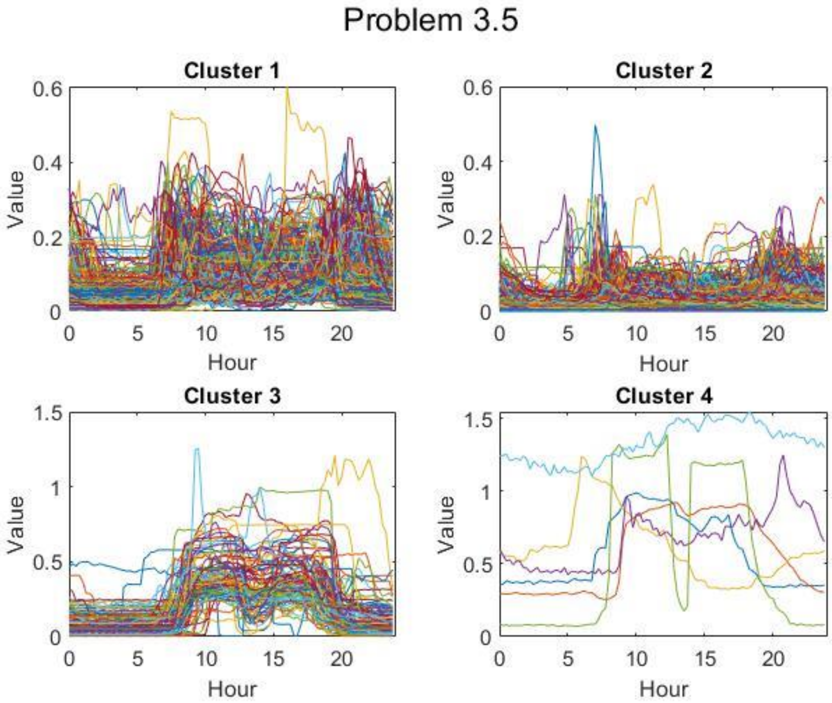
Using another loop with a minimum variance explained condition, we find that 4 eigenvectors are needed to explain 92.43% of the variance. The first eigenvector alone explains 73.87%, the first two 86%, the first three 90.28%, and the first four finally achieves 92.43% (code lines 515-523).

4) Use the previous number to apply K-Means. Show the clusters given by the method and their centroids.





5) Plot the data of the original accounts separated into K subplots. What can you learn from the accounts that belong to each of the K clusters?



The first cluster, being the largest, seems to most closely approximate the original data, excluding the relatively high observations, which are contained mostly in cluster 4. Cluster 2 seems to contain mostly low value accounts, with a few spikes in the morning hours, but otherwise little discernable signal. Cluster 3 shows more usage in the middle of the day and low values at night, perhaps corresponding to retail stores or residential customers who use more electricity during the day. Cluster 4 contains only a few extreme observations, perhaps representing commercial or industrial customers.