**Questions and Report Structure**

**Component analysis**

1. Reflection on PCA/ICA
   * What are likely candidates for early PCA dimensions?

Based on a basic set of statistics, I would say that the first PCA dimensions will correspond to Fresh because it is the feature with the highest variance (in this case standard deviation). Grocery and Milk are up there next.

* + What might ICA dimensions look like?

In the case of ICA we are going to find independent vector represented by a combination of the initial features. I would expect the independent components to represent the buying patterns of different types of customers composed by their spending in different product categories.

1. What proportion of variance is explained by each PCA dimension?

The first two principal components make for about the 85% of the variance (45% and 40% respectively) and then drops to about 7%.

Based on this, I would argue that we should choose two dimensions for PCA because they explain the vast majority of the variance. The next PCs seam not that relevant.

1. PCA dimensions
   * What are the first few components? What might they represent?

In the case we use two dimensions, I would argue that the first Principal Component represent mainly Fresh due to the magnitude of the Fresh component being much higher than the rest. I would say that the second PC is a combination of Grocery, Milk and Detergents\_Paper (in that order of importance) based on the magnitude of their components.

\*From searching the Discussion Board I found an implementation of a biplot (provided by other student, jjinking) that confirmed my findings based on the magnitude / direction of each variable arrow.

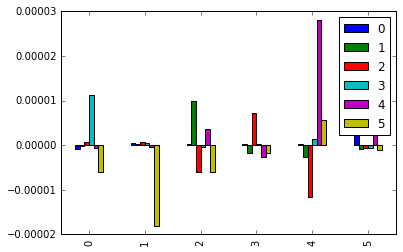
* + How can you use this information?

I would use this information to find out that the most important variable (the one with the most variance) is Fresh and the combination of Grocery, Milk and Detergents\_Paper helps us understand another big portion of the variance. This might be useful to assess the findings in the clustering algorithms and to help us understand the key differences between customers.

1. ICA
   * What are the components that arise?

Explanation of what might each of this vectors represent:

* + - Vector 0 represents Grocery and Delicatessen.
    - Vector 1 represents mainly Delicatessen.
    - Vector 2 represents a combination of Milk, Grocery, Detergents\_Paper and Delicatessen.
    - Vector 3 represents mainly Grocery and a little but of Milk, Detergents\_Paper and Delicatessen.
    - Vector 4 represents mainly Detergents\_Paper and a combination of Milk and Delicatessen.
    - Vector 5 represents Fresh and Detergents\_Paper.



*Figure Nº1: Visual representation of the ICA components after fit*

* + How could you use these components?

The components can be used to transform the original data set into a data set based on independent vectors which will help clustering and classification algorithms to understand trends in data more easily. This is specifically important for algorithm that require independence in between features.

**Clustering**

1. Decide on K means clustering or Gaussian mixture methods
   * What are the advantages and disadvantages of each?

Both clustering algorithms are quite intuitive to run and understand (we just need the specify the number of clusters and run the models) while they also require much less computational time than other clustering algorithms like hierarchical clusters.

The main drawback in this algorithms is determining the number of clusters. As they are required a priori, determining this number could turn into a challenge.

K Means is computationally faster for a small amount of clusters (*k*) than other hierarchical clustering algorithms.

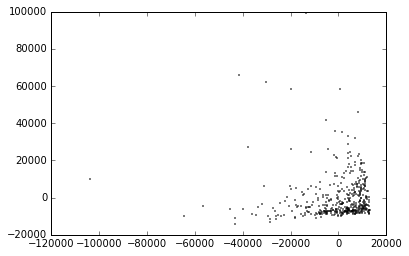
* + How will you decide on the number of clusters?

I will make a visual inspection of the reduced data using PCA on two PCs. Based on this inspection I will try to predict the number of clusters, and will start using both algorithms on this number of clusters. Results will be analyzed on several number of clusters to check their results.

My main goal here is to achieve clusters that are different from each other between clusters, and members of each clusters should be as similar as possible.

1. Implement clusters
   * Sample central points of the clusters
2. Produce a graphic
   * Visualize important dimensions by reducing with PCA

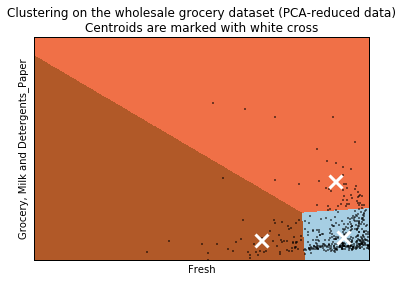
Based on Figure Nº1, we can see that most customers are placed on the bottom right of the graph. In this case the x axis (PC1) is represented mainly by Fresh and the y axis (PC2) is represented by the combination of Grocery, Milk and Detergents\_Paper based on the earlier conclusions.



*Figure Nº2: Visual representation on the reduced data set using PCA with two PCs*

My first appreciation is that there are three clusters. One cluster representing the customers on the bottom right, other cluster representing those who spend more on Fresh while spending the same on Grocery, Milk and Detergents\_Paper and other cluster were customer spend more on Grocery, Milk and Detergents\_Paper and the same on Fresh.

After the visual inspection, I proceeded to test both clustering algorithms with k values between 2 and 4. The best outcome in my opinion came from the k-means clusters with k = 3.

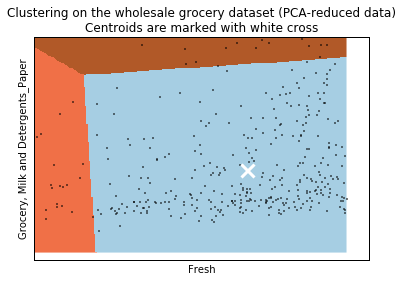


*Figure Nº3: Visual representation of the K-Means results with k = 3.*

After several iterations a chose this number of k based on the visual inspection of the results. The results associated with 4 clusters showed one clusters with very few customers that were quite sparse (the ones at the top) and 2 clusters appeared to group customers that were not very similar between each other.

* + Are there clusters that aren’t very well distinguished? How could you improve the visualization?

The bottom right cluster (light blue) has a high concentration of customers which are hard to visualize.



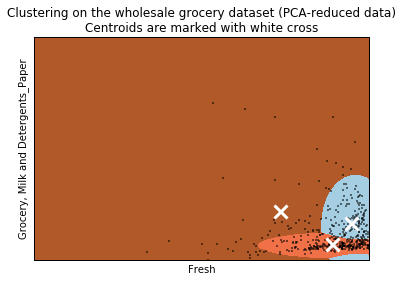
*Figure Nº4: Zoom over on cluster of the visual representation of the K-Means results with k = 3.*

I think this cluster should be further analyzed based on the high density it has and to understand the differences between customers in this cluster. There seams to be a particular group with similar spending’s on Grocery, Milk and Detergents\_Paper but with different spending’s on Fresh.

**Conclusions**

1. Which of these techniques felt like it fit naturally with the data?

I think K-means gives more insights about this data mainly because my results showed clusters that overlapped a lot (mainly the one at the bottom right and top right) on GMM. This overlap resulted in some confusion about how the customers really are. Figure Nº4 shows the results for 3 clusters.



*Figure Nº5: Visual representation of the GMM results with k = 3.*

1. How would you use that technique to assist if the company conducted an experiment?

I would recommend the company to make a more in depth study about which products are customer buying based on this clusters. For example, understand the spending’s on different products, or sub categories, and check if their behavior is truly different.

My intuition would say that customers in each cluster are different from each other, so I would expect to identify new clusters (considering additional information in the dataset) in order to create personalized campaigns to boost sales.

1. How would you use that data to predict future customer needs?

I would try to understand as much as possible the difference between each cluster, and sub clusters using additional information, and by creating customer intelligence I would design new products/campaigns suited for their needs. Also I would search the market for new/related products that appear that could be of interest for them.