Report

In this section you will be using PCA and ICA to start to understand the structure of the data. Before doing any computations, what do you think will show up in your computations? List one or two ideas for what might show up as the first PCA dimensions, or what type of vectors will show up as ICA dimensions.

First component in PCA will capture data with high variance, by looking at data statistics , Fresh and Grocery attributes have the highest std deviation so they might show up the most in the first PCA dimension
ICA will try to extract from the original features new independent non overlapping set of features.

How quickly does the variance drop off by dimension? If you were to use PCA on this dataset, how many dimensions would you choose for your analysis? Why?

From the output of *pca.explained_variance_ratio_* ([0.45961362 0.40517227 0.07003008 0.04402344 0.01502212 0.00613848]), variance first drops off by .5% to reach .40 for the 2nd component then it drops incredibly fast, the remaining four components have variance less than 1.
I'd choose the first two dimensions with variance .45 and .40, as they capture most of the original data, while the other components may be representing some noise due to their very low eigenvalues.

What do the dimensions seem to represent? How can you use this information?

By looking at the weights for each component (*pca.components_)* , High weight for a certain attribute means that there is a strong correlation between that attribute and the principle component.
The first PC is strongly correlated with Fresh attribute (-0.97653685) , other attributes weights are very low.
The second PC represents Mainly Grocery(0.76460638) and Milk(0.51580216)attributes.
By continuing with this sequence, setting a threshold of 0.5, any weight magnitude that exceeds that threshold value well be considered strongly correlated with the corresponding PC .

For each vector in the ICA decomposition, write a sentence or two explaining what sort of object or property it corresponds to. What could these components be used for?

1st vector: This component represents Grocery, Detergents_Paper products,with an inverse relationship between the two
2nd vector: This component doesn't represent much, maybe only a little Grocery spending.
3rd vector: This component represents spending on Delicatessen products.
4th vector: This component represents spending on Fresh products.
5th vector: This component represents spending on Detergents_Paper products.
6th vector: This component represents spending on Milk and Grocery products with an inverse relationship between the two

What are the advantages of using K Means clustering or Gaussian Mixture Models?

Main difference between Kmeans and GMM is that Kmeans performs hard clustering and GMM performs soft clustering, Kmeans is simpler and much more efficient, where GMM is more computationally expensive as it calculates more things during performing the expectation part, i would go with a simpler more efficient model for this problem(Kmeans), no need for the compelxity of GMM

What are the central objects in each cluster? Describe them as customers.

First PC1 corresponds mainly to Fresh Products, and small weights for Milk, Frozen products
Second PC2 corresponds to spending on Grocery,Milk,Detergents_Peper products

From the 3 clusters above, the cluster on the right which is more condensed than the other two, represents High values on PC1 and low PC2 values, which i guess means that this customers segment spends more on Fresh products and slightly less on MIlk and Frozen products
The 2nd cluster on the left, represents values with lower PC1 values and low PC2 values, Which means less spending on Fresh , milk and frozen products than the first customer segment, and ofcourse small spendings on the other products (PC2 products) as well like the first customer segment
The 3rd cluster above the other two, represents High PC1 and PC2 values, which means customers in this segment spend on PC1 products like the first cluster, and spend on PC2 products as well, which are Grocery,MIlk,Detergents_Peper

Which of these techniques did you feel gave you the most insight into the data?

PCA with Kmeans gave me the most information about the data, i can't see where ICA fit with this problem.

How would you use that technique to help the company design new experiments?

By clustering the data the compnay would now have some input on different categories of customers for the company, By analyzing the result of clustering, customers that need Fresh,Milk products most,will need this products in the morning as they open, other products for example detergents are ok to be delivered in the night.
If the major number of customers for the company are from type 1 which needs fresh products, the company must change it's experiment as it won't fit with this category and vice versa.

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

After clustering this data, i can label it, for example i clustered it into 3 clusters, so i have three labels, type1 type2 and type3 and then i can feed my data into a supervised learning algorithm and get a better model.