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?

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 .

How can you use this information?

Since the first two vectors capture most of the data, I can reduce the number of features to only two new features, one feature is a combination of fresh foods, milk, frozen foods and the other feature is a combination of milk, groceries and detergent_paper
This will help to better visualize the data in 2d space, and will reduce curse of dimensionality.

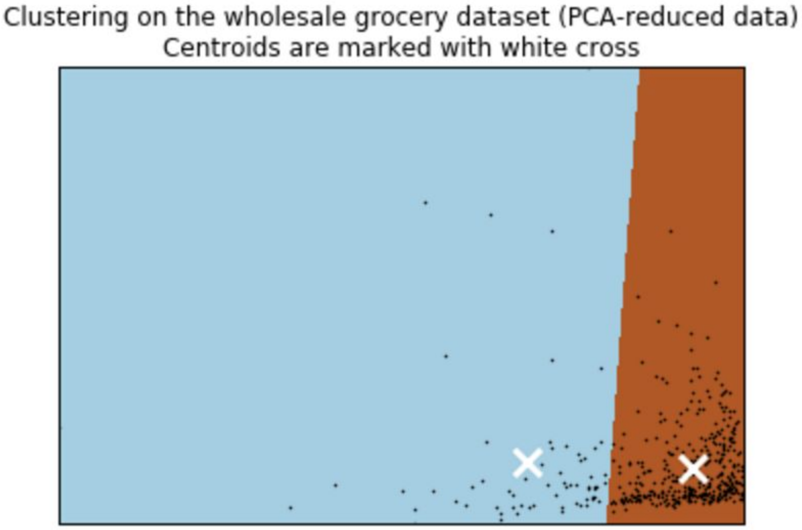
For each vector in the ICA decomposition, write a sentence or two explaining what sort of object or property it corresponds to. What could these components be used for?

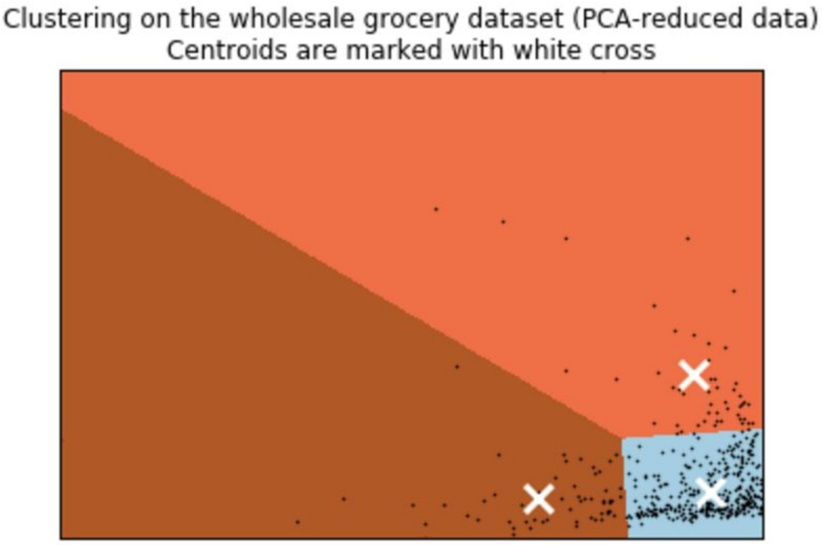
1st vector: This component represents Grocery, Detergents_Paper products,with an inverse relationship between the two
2nd vector: This component doesn't represent much, maybe only a little Grocery spending.
3rd vector: This component represents spending on Delicatessen products.
4th vector: This component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

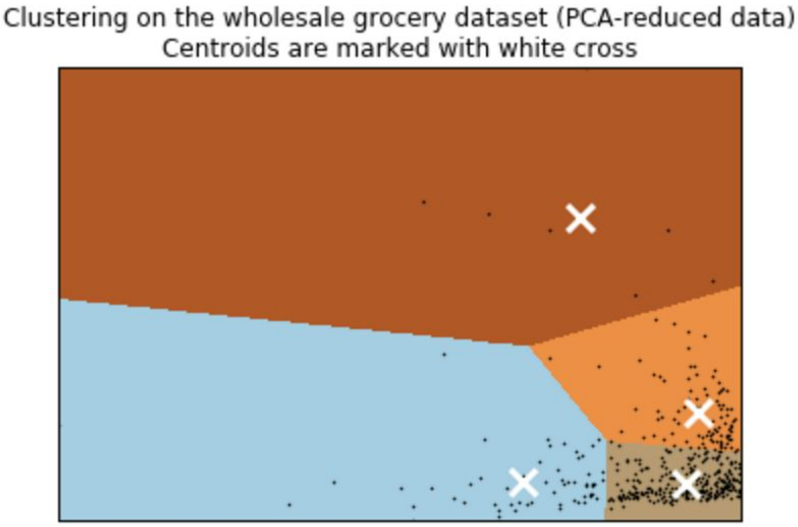
Main difference between Kmeans and GMM is that Kmeans performs hard clustering and GMM performs soft clustering, Kmeans is simpler and much more efficient, where GMM is more computationally expensive as it calculates more things during performing the expectation part, i would go with a simpler more efficient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

Choosing K in Kmeans





Clustering on the wholesale grocery dataset (PCA-reduced data)
Centroids are marked with white cross



Clustering on the wholesale grocery dataset (PCA-reduced data)
Centroids are marked with white cross

After trying different number of clusters 2 - 3 - 4 ,after looking at the scatter plot of the data, and knowing i have two PCs representing my data, I didn’t choose 2 as the number of clusters, I assumed that each PC is either high(H)or low(L) , so probably i will have four clusters (HH)(HL)(LH)(LL), after trying 3 & 4 as the number of clusters, i settled with 3 as it represents (HH)(HL)(LL), I didn’t choose 4 clusters as I don’t like the brown cluster, it doesn’t distinguish the customers well, as people in this cluster may have high or low values for PC1.

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 means that this customers segment spends less on Fresh products and slightly more on Milk and Frozen products, and spends less on PC2 products (milk, groceries and detergents_paper).

The 2nd cluster on the left, represents values with lower PC1 values and low PC2 values, Which means more spending on Fresh products, less spending on milk and frozen products than the first customer segment, and of course small spendings on PC2 products.

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 well on PC2 products as well.

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, PCA main advantage was the ability to downsize the number of features to two features only representing the whole data and capturing all the original 6 features, this appeared first when analyzing the vecotrs of the 6 PCs, the first two vectors had the highest weights, when i have two features the data could easily be plotted in scatter plots, and i could apply kmeans on it and choose the amount of K i need according to the data distribution on the scatter plot.
Kmeans represents different types of customers, after clustering the data i could now understand that i have three different type of customers buying from the wholesale store, and each type is interested in different things according to their PC1 and PC2 values.

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

After clustering the data, I now have 3 types of customers (clusters), if the company wanted to try a new experiment, it could split first each cluster to two groups A and B, and try the new experiment on part of a cluster and the other part of the same cluster will continue using the old technique, then the company could monitor the effect of the new experiment on each type of customers, and whether the new experiment was better (sales increased, complaints decreased) or should it stick with the old technique for that specific cluster.
This A/B testing should be applied on the same type of customers, that’s why it should be applied to each cluster, as each cluster represents a different type of customers than the other one.

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.