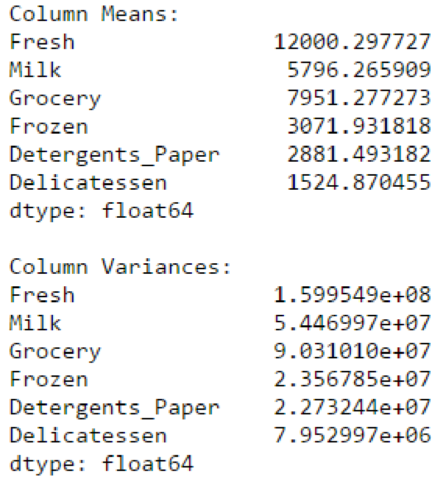
Thomas Hepner
2/19/2015
P3 - Creating Customer Segments

Component analysis

1. Reflection on PCA/ICA

- What are likely candidates for early PCA dimensions?

PCA uses an orthogonal transformation to convert a set of potentially correlated variables into a set of linearly uncorrelated variables. The variables in our data, *Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicatessen,* may have a high degree of collinearity. Early PCA dimensions will represent the variables that capture the largest proportion of variation in the data. The variables *Fresh, Grocery,* and *Milk* seem like good candidates for early PCA dimensions because they intuitively constitute the bulk of consumers' retail purchases. A quick examination of the feature means and variances shows these variables have the largest means and greatest variances suggesting that they will make ideal PCA dimensions.



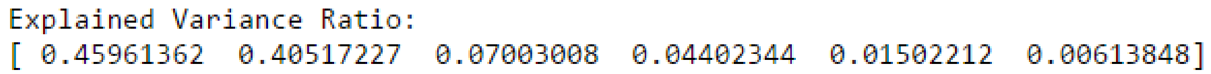
```
Column Means:
Fresh               12000.297727
Milk                 5796.265909
Grocery              7951.277273
Frozen               3071.931818
Detergents_Paper     2881.493182
Delicatessen         1524.870455
dtype: float64

Column Variances:
Fresh               1.599549e+08
Milk                5.446997e+07
Grocery             9.031010e+07
Frozen              2.356785e+07
Detergents_Paper    2.273244e+07
Delicatessen        7.952997e+06
dtype: float64
```

- What might ICA dimensions look like?

ICA unveils possibly hidden features that underlie our observed variables. ICA dimensions might represent the types of customers makes purchase from the company. They might be supermarkets, grocery stores, convenience stores, gas stations, discount retailers, and E-commerce stores. Each of these types of businesses have their own set of needs that are hidden from us; ICA might be able to reveal the types of businesses to which the company caters.

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



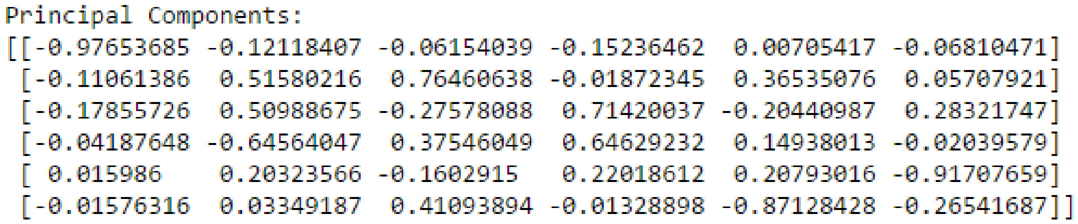
```
Explained Variance Ratio:
[ 0.45961362  0.40517227  0.07003008  0.04402344  0.01502212  0.00613848]
```

Using 80% of the explained variance as a cutoff point, the first two dimensions are selected as they account for over 86% of the total variance in the data. The proportion of variance explained falls for each dimension falls sharply after the first two.

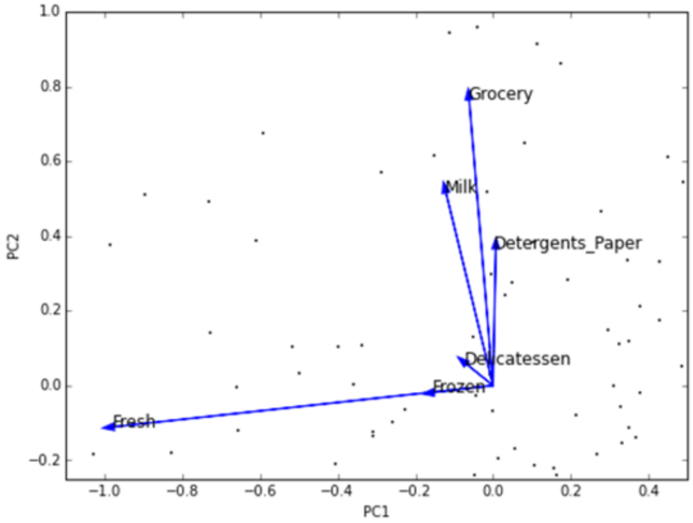
3. PCA dimensions

- What are the first few components? What might they represent?

Component 1 (first row) has a strong relationship with *Fresh*; it might mostly represent the original *Fresh* variable. Component 2 has strong associations with *Milk* and *Grocery*; it might represent a combination of the two variables which we will call *Dairy*.



```
Principal Components:
[[-0.97653685 -0.12118407 -0.06154039 -0.15236462  0.00705417 -0.06810471]
 [-0.11061386  0.51580216  0.76460638 -0.01872345  0.36535076  0.05707921]
 [-0.17855726  0.50988675 -0.27578088  0.71420037 -0.20440987  0.28321747]
 [-0.04187648 -0.64564047  0.37546049  0.64629232  0.14938013 -0.02039579]
 [ 0.015986    0.20323566 -0.1602915   0.22018612  0.20793016 -0.91707659]
 [-0.01576316  0.03349187  0.41093894 -0.01328898 -0.87128428 -0.26541687]]
```



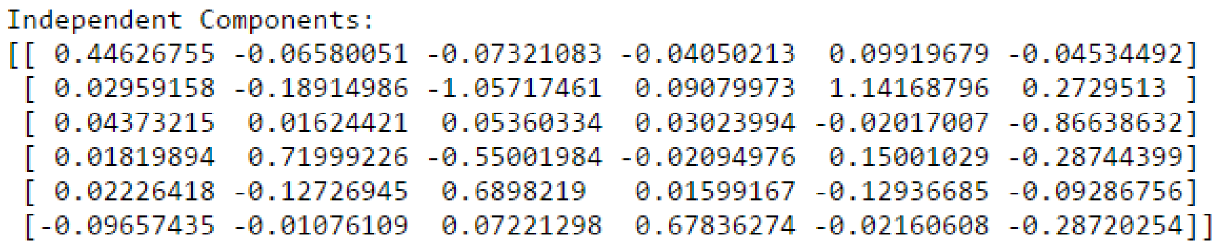
- How can you use this information?

We can use the principal components, or rather a subset of them, to capture as much of the variance in the data as possible while reducing the number of features; it is a feature selection and reduction technique. By doing this, and selecting the first *n* variables that explain a cutoff variance in the data, such as 80% or 90%, we can keep the majority of the signal in the data while reducing the feature set and eliminating noise.

4. ICA

- What are the components that arise?

Component 1 (first row) has an association to Fresh; Component 2 has a strong association with *Grocery* and also with *Detergents_Paper*; Component 3 has a strong association with *Delicattessen*; Component 4 has strong relationships with *Milk* and *Grocery*; Component 5 has a strong relationship with *Grocery*; Component 6 has a relationship with *Frozen*.



```
Independent Components:
[[ 0.44626755 -0.06580051 -0.07321083 -0.04050213  0.09919679 -0.04534492]
 [ 0.02959158 -0.18914986 -1.05717461  0.09079973  1.14168796  0.2729513 ]
 [ 0.04373215  0.01624421  0.05360334  0.03023994 -0.02017007 -0.86638632]
 [ 0.01819894  0.71999226 -0.55001984 -0.02094976  0.15001029 -0.28744399]
 [ 0.02226418 -0.12726945  0.6898219   0.01599167 -0.12936685 -0.09286756]
 [-0.09657435 -0.01076109  0.07221298  0.67836274 -0.02160608 -0.28720254]]
```

- How could you use these components?

These components can be used to understand the hidden mutually independent features that might exist to explain the data. These hidden features might represent, and help explain, the diverse types and needs of business customers. For instance, Component 1 which has a strong relationship with *Fresh* could potentially refer to grocery store which sells a lot of fresh produce, and Component 3 which has a strong association with *Delicattessen* could represent stores selling fine foods like specialty cheeses and meats.

Clustering

5. Decide on K means clustering or Gaussian mixture methods (GMM)

- What are the advantages and disadvantages of each?

K means is a *hard* clustering algorithm, meaning that each data point fits into exactly one cluster, the cluster to which it is closer to than any other. The advantages of K means are that it is (1) computationally faster than GMM and (2) its clusters are tighter than GMM; its disadvantages are that it (1) is difficult to compare the quality of the clusters it produces, (2) does not work well with non-globular clusters, (3) is sensitive to outliers and noise, and (4) produces different results based on starting initialization of centroids.

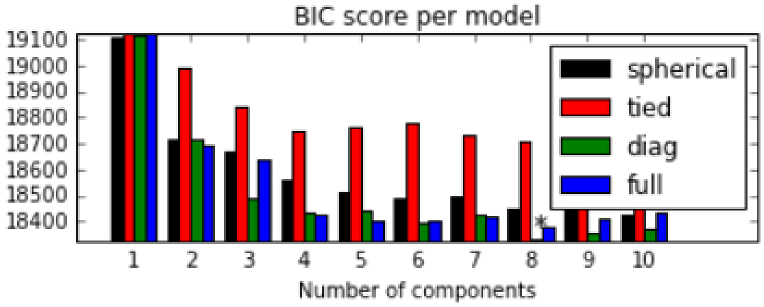
GMM is a *soft* clustering algorithm, meaning that each data point is can belong to multiple clusters based upon the strength of the relationship between the point and the cluster. Each cluster corresponds to a Gaussian, or probabilistic, distribution. The algorithm determines the parameter values of the Gaussian distribution for each cluster, and assigns probabilities to each data point that that particular point belongs to a specific cluster. The advantages of the GMM algorithm are that it (1) creates flexible probability distributions and clusters, and (2) works well with non-globular clusters and is not sensitive to outliers and noise; its disadvantages are that it is (1) computationally more expensive than K means, and (2) its clusters are not as tight as clusters produced by K means.

I will use the GMM clustering algorithm for this particular problem because our data size and number of features are both small and therefore performance is not a top concern, the *soft* clustering approach seems like a better fit for interpretation of the clustering, and it seems like there will be non-globular clusters in the data.

- How will you decide on the number of clusters?

I will calculate the Bayesian Information Criterion (BIC) score for different combinations of GMM models with varying number of components and covariance parameters. The GMM model with the lowest BIC will be used to determine for the final clustering model.

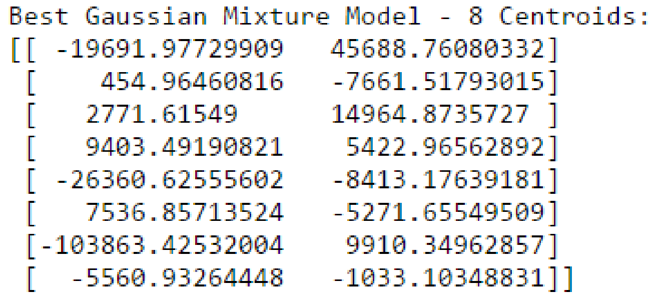
BIC is a metric used for model selection where the model with the lowest BIC is preferred. I created 40 GMM models, with cluster sizes varying from 1 to 10, and with 4 different covariance measures. I plotted the BIC scores of every model in the chart below.



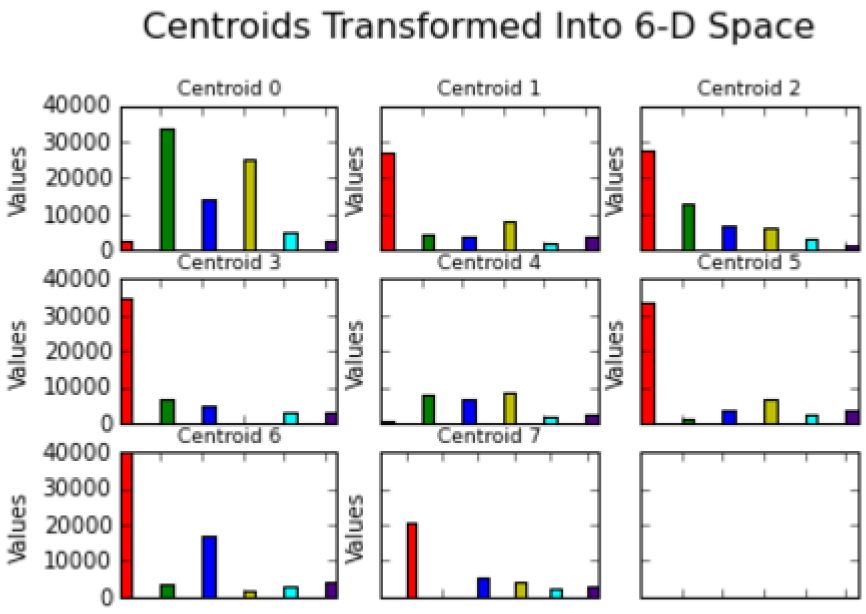
The model with the lowest BIC has 8 clusters, or components, and has *diag* covariance measure. Therefore, 8 clusters are chosen.

6. Implement clusters

- Sample central points of the clusters. What are the central objects in each cluster?



```
Best Gaussian Mixture Model - 8 Centroids:
[[ -19691.97729909   45688.76080332]
 [    454.96460816   -7661.51793015]
 [   2771.61549      14964.8735727 ]
 [   9403.49190821    5422.96562892]
 [ -26360.62555602   -8413.17639181]
 [   7536.85713524   -5271.65549509]
 [-103863.42532004    9910.34962857]
 [  -5560.93264448   -1033.10348831]]
```



The bars in the above charts represent the variables in sequential order: *Fresh, Milk, Grocery, Frozen, Detergents_Paper,* and *Delicatessen.* The centroids represent the following types of customers:

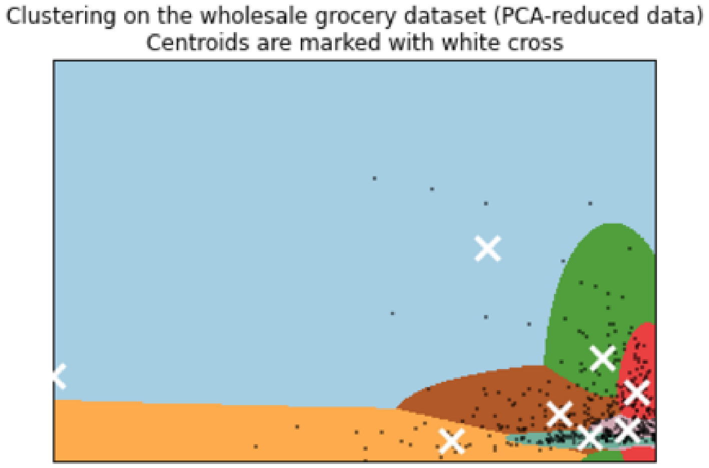
- Centroid 0: Purchases relatively more from *Milk, Frozen,* and *Grocery.* (Ex: Grocery store)

- Centroid 1: Mostly purchases *Fresh.* (Ex: Outdoor Produce Stand)
- Centroid 2: Buys mostly *Fresh* but also *Milk.*
- Centroid 3: Procures much more *Fresh* than any other category.
- Centroid 4: Purchases little of any products, but purchases more of *Frozen, Milk,* and *Grocery.*
- Centroid 5: Mostly buys *Fresh,* but also purchases *Frozen.*
- Centroid 6: Purchases a lot of *Fresh*, but also *Grocery.*
- Centroid 7: Purchases *Milk*, and little from other categories.

Additional Commentary: *Fresh* is the strongest feature in 5 of the 8 centroids; *Milk* and *Grocery* are also important, but to a lesser degree than *Fresh* in the majority of the centroids. These variables appear to distinguish the company's customers more than the other features.

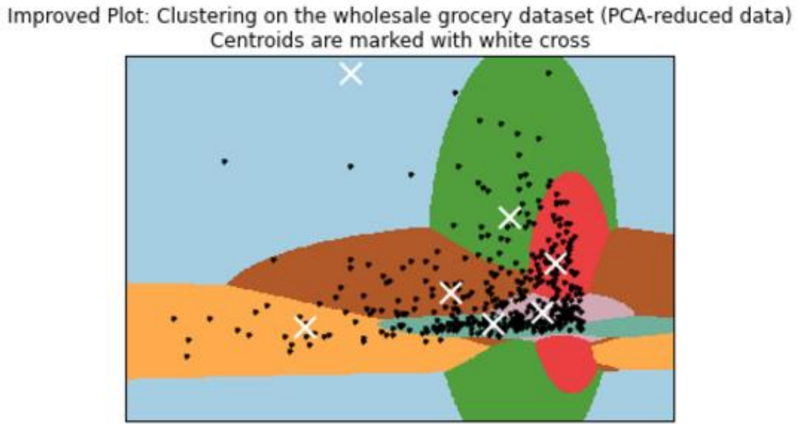
7. Produce a graphic

- Visualize important dimensions by reducing with PCA



- Are there clusters that aren't very well distinguished? How could you improve the visualization?

The clusters in the bottom right corner are densely populated and likely have a lot of overlap. The corners in the top and left halves of the plot are sparsely populated and take up the majority of the plot; I could improve the visualization by focusing on the data in the bottom right corner since it contains the majority of the data points and centroids.



Conclusions

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

PCA seemed like a better approach than ICA for feature selection and reduction before applying K Means or GMM to the reduced data. PCA revealed that there was a high degree of collinearity in the data; the first 2 PCA components created from the original 6 features in the data was able to capture over 86% of the variation. Intuitively this makes sense, as many retailers sell a variety of different products like fresh produce, frozen food, and home supplies like detergents and toilet paper.

The Gaussian Mixture Model (GMM) felt like it naturally fit with the data. K means is a *hard* clustering algorithm and it didn't seem appropriate to cluster businesses that sell a diverse combination of products into single, non-overlapping categories. The *soft* clustering provided by GMM seemed more appropriate given that a supermarket can also be a grocery store, and a convenience store or gas station sells groceries as well as its other products. In addition, the data, as you can see in the plots above, was non-globular in many parts of the graph, and K means is ill-suited for clustering these points. For these reasons, I chose GMM since it seemed to naturally fit with the data and business context.

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

Let's say the company is planning on doing an A/B test, or randomized controlled trial, or a new business practice. For example, perhaps the company is switching deliveries of milk and cheese from Sunday evenings to Monday mornings to improve freshness and the quality of the products delivered to customers. We could use our clusters created by our GMM to make sure that when we do our A/B test that we are comparing customers within the same clusters; for instance, this new business practice might be beneficial for one type of customer, like supermarkets, but terrible for convenience stores. The clusters from our GMM would prevent us from making false inferences about our customer data, and make necessary changes to improve the business.

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

We could take the existing customer data and clusters, using the clusters as categories to create a supervised learning model, such as K Nearest Neighbors, and use the model to predict customer needs within each cluster, and potentially even between clusters. For instance, one cluster might represent supermarkets, and these customers on average might need 1,000 containers of Tide detergent delivered biweekly on Wednesday evenings and Monday mornings, whereas another cluster might represent gas stations, and the typical gas station needs 50 liters of milk delivered every Monday morning. We could use our past data about customers, and which group they belong to, to anticipate the future needs of our existing customers, and new customers we expect to acquire.