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

Component analysis

1. Reflection on PCA/ICA

- **What are likely candidates for early PCA dimensions?**

Some combination of Fresh, Milk, and Grocery. It seems that the bulk of store purchases will come from these 3 categories and it is intuitive that these things would be purchased together by customers, so they will likely have a high degree of collinearity.

- **What might ICA dimensions look like?**

ICA dimensions might look like (1) a dimension that is a combination of Fresh, Milk, and Grocery, (2) Frozen, (3) Detergents_Paper, and (4) Delicatessen. It seems like there would be low collinearity between these 4 features as they are all very different products that intuitively would be purchased at different times and potentially by different customers.

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

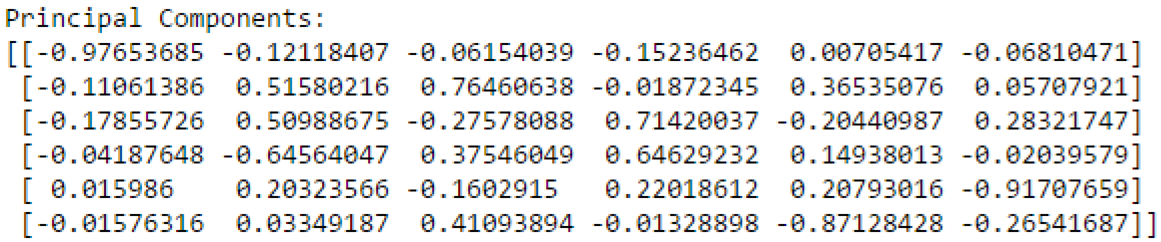
Dimension 1: 0.46
Dimension 2: 0.405
Dimension 3: 0.07
Dimension 4: 0.044
Dimension 5: 0.015
Dimension 6: 0.006

The variance drops off very quickly after the first two dimensions; the first two dimensions account for over 86% of the total variance in the data.

3. PCA dimensions

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

Dimension 1 (first row) has a strong negative correlation with Fresh; it might represent a variable meaning ‘Not Fresh’. Dimension 2 has positive correlations with Milk and Grocery; it might represent a ‘Dairy’ variable.



```
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?**

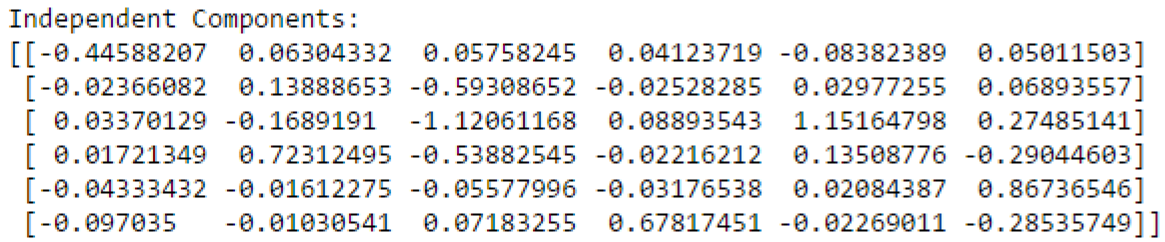
This information can be used for two purposes. The first is to interpret *hidden* underlying variables that are driving deliveries for this business. In this case, it appears there are two hidden variables, a ‘Not Fresh’ variable, and a ‘Dairy’ feature that are might be causing the majority of the variance in the data.

Another thing we can do is use these components, or a subset of them, for building a predictive model. By using a subset of the components, like the first two or three variables, we can eliminate some of the noise from the data and possibly produce more accurate predictions.

4. ICA

- **What are the components that arise?**

Component 1 (first row) has a negative association to Fresh; Component 2 has a negative association with Grocery; Component 3 has a very strong negative association with Grocery but also has a very strong positive relationship with Detergent_Paper; Component 4 has a strong positive relationship with Milk and negative relationship with Grocery; Component 5 has a strong positive association with Delicattessen; Component 6 has a positive relationship with Frozen.



```
Independent Components:
[[-0.44588207  0.06304332  0.05758245  0.04123719 -0.08382389  0.05011503]
 [-0.02366082  0.13888653 -0.59308652 -0.02528285  0.02977255  0.06893557]
 [ 0.03370129 -0.1689191  -1.12061168  0.08893543  1.15164798  0.27485141]
 [ 0.01721349  0.72312495 -0.53882545 -0.02216212  0.13508776 -0.29044603]
 [-0.04333432 -0.01612275 -0.05577996 -0.03176538  0.02084387  0.86736546]
 [-0.097035   -0.01030541  0.07183255  0.67817451 -0.02269011 -0.28535749]]
```

- **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 mutually independent features might represent, and help explain, the diverse attributes and needs of the business's customers. For instance, Component 1 which has a negative relationship to Fresh could potentially refer to gas stations, and Component 5 which has a strong positive association with Delicattessen could be bakeries.

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.

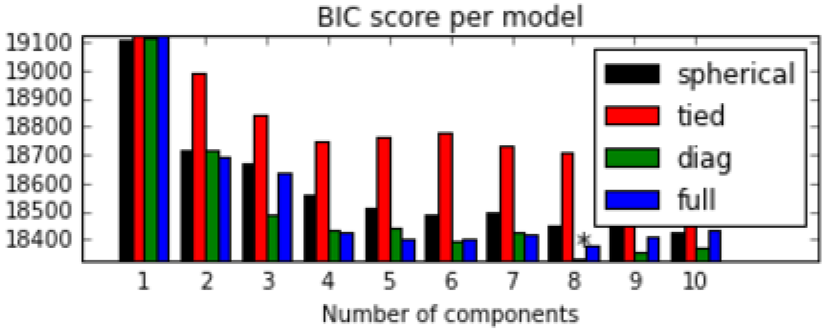
GMM is a *soft* clustering algorithm, meaning that each data point is can belong to multiple clusters based upon the strength of the relationship between the point and the cluster. Each cluster corresponds to a Gaussian, or probabilistic, distribution. The algorithm determines the parameter values of 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.

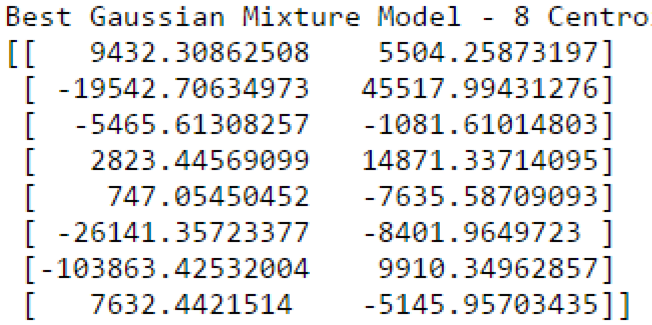
BIC is a metric used for model selection where the model with the lowest BIC is preferred. I created 40 GMM models, with cluster sizes varying from 1 to 10, and with 4 different covariance measures. I plotted the BIC scores of every model in the chart below.



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? Describe them as customers.**

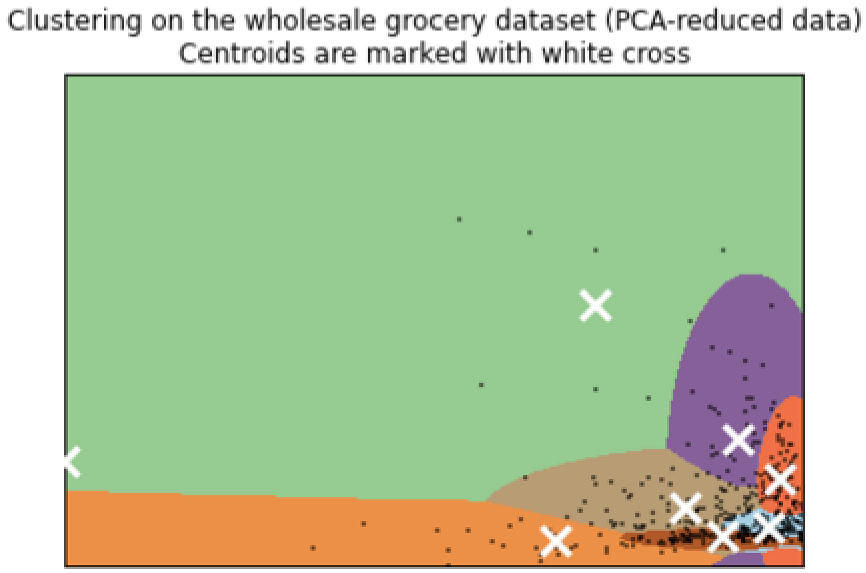


```
Best Gaussian Mixture Model - 8 Centroids:
[[   9432.30862508    5504.25873197]
 [ -19542.70634973   45517.99431276]
 [  -5465.61308257   -1081.61014803]
 [   2823.44569099   14871.33714095]
 [    747.05450452   -7635.58709093]
 [ -26141.35723377   -8401.9649723 ]
 [-103863.42532004    9910.34962857]
 [   7632.4421514    -5145.95703435]]
```

The first column is the first PCA component that I previously interpreted as a variable for 'Not Fresh'; the second column is the second PCA component that I previously named 'Dairy'. The first row in the table above, representing the centroid of a cluster in the data, has high positive values for both. This could possibly be a customer that sells both products that are 'Not Fresh' and also products that are 'Dairy'; a gas station could potentially fit this description. The second row in the table has a high negative value for 'Not Fresh' and a very high positive value for 'Dairy'; this could possibly represent a grocery store that sells both fresh produce and dairy products.

7. Produce a graphic

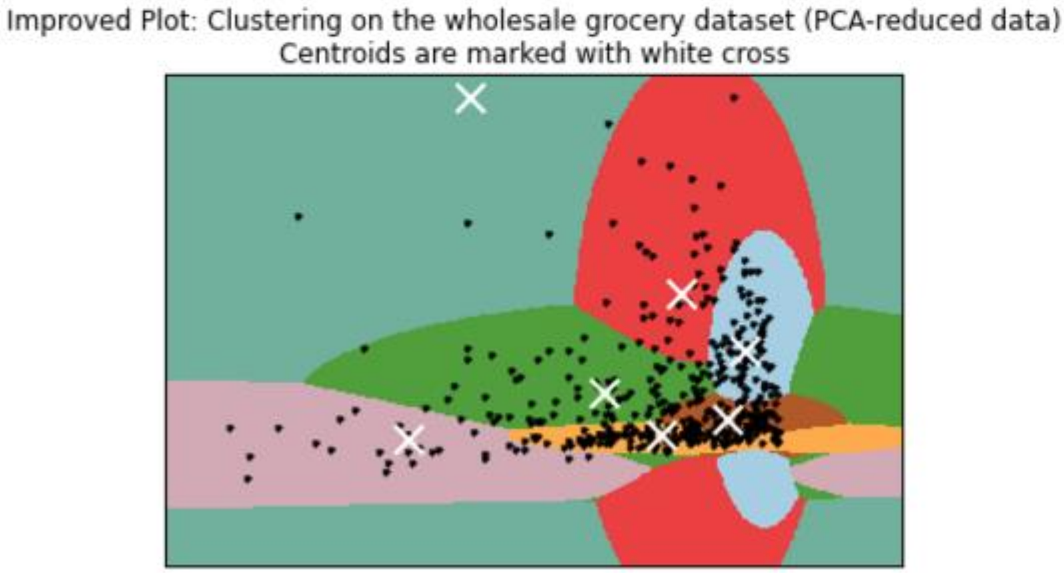
- Visualize important dimensions by reducing with PCA



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

- **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.



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

Conclusions

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

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 in a supervised learning model, 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.