Report

By: Shmuel Naaman

Component analysis

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

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. The first principal component will represent a projection of the data that encompass the largest variance in the data.

We can be slightly more precise here, if we calculate the variance of each feature in the data.

```
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
```

We can see that “Fresh” have the largest variance and the next is the “Grocery”. Therefore, we can guess that the first component will contain mostly the “Fresh” and the second component will contain mostly the “Grocery”.

    - What might ICA dimensions look like?

Independent component analysis (ICA) is a technique used to separate a signal into additive subcomponents. In this case the components would not be arranged in a particular hierarchical order. The components in this case might represent independent customer types or other fundamental features of the data.

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

'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents Paper', 'Delicatessen'

```
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]]
```

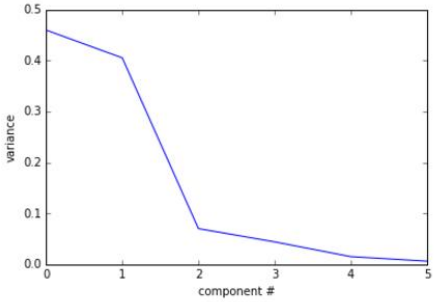
Variance [0.459, 0.405, 0.070, 0.044, 0.015, 0.006]

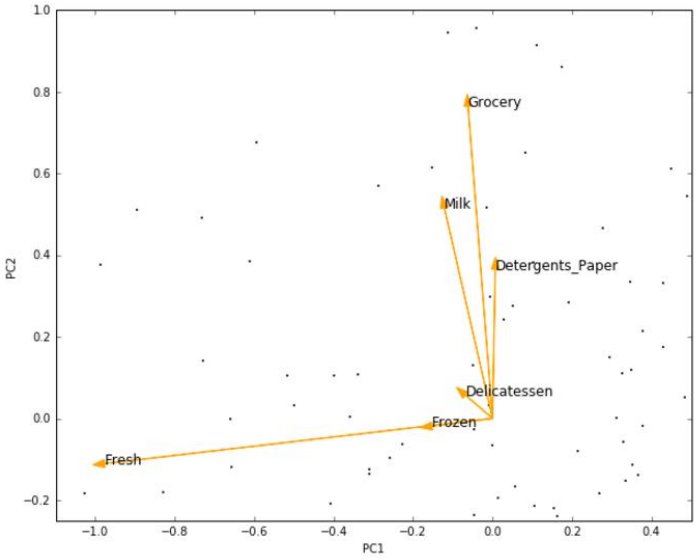
The first 3 components encompass 93% of the variance in the data, as can be seeing in the left figure below and the values above. The first component

encompass ~46% of the variance the second ~40%, the third ~7% and the forth 4%. The other components encompass less than 10% of the variance. I would probably use **the first 4 components**.

The reason: I will want to **include as much variance as possible**, but on the same time there is no much sense to include components that represents 1 or 2 percent of the data. Most probably these **components represent noise or outliers** and would not contribute to the understanding of the data and analysis.

So I would set a **threshold of 2%**, components that explain less than that will not be included.





3. PCA dimensions
    - What are the first few components? What might they represent?
    - How can you use this information?

Answer: The dimensions represent different principal projection of variance in the data. The first dimension represents the projection of the largest variance of the data that might be something like the average buyer. The coefficients values indicate that the first component will include mostly "Fresh" (0.976) and some "Frozen" (-0.15). The second dimension will be orthogonal to the first component, but again at the projection that captures the most of the remained variance. Here the coefficients values indicate that the second component will include mostly 'Grocery' (0.76), less 'Milk' (0.51) and even less 'Detergents Paper' (0.36). The next dimension will be constructed on similar principle, orthogonal and explaining the most of the remained variance. Looking at the value we find that the third component encompass some linear combination that is more or less homogeneous.

This separation of the data into components provides an **orthogonal basis**. The great benefit of such a basis is the possibility to reconstruct the data using only the components that we find interesting or important. This can be useful to remove noise or components that are not relevant.

In this case we can see that the first component include mostly “Fresh”, therefore the distributor need to be aware that the Fresh behave almost as an independent component. For the second component we find that 'Milk', 'Grocery' and 'Detergents Paper’, have similar coefficient value. That indicates that they are dependent components. For example high consumption of Milk will be accompanied by high consumption of Grocery. In other words, customers that consume Milk probably will consume also Grocery and Detergents Paper.

4. ICA

- What are the components that arise?
- How could you use these components?

'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents Paper', 'Delicatessen'

```
[
[ -1.51756140e-07 -9.85444905e-06 5.79190337e-06 3.67884619e-07 -3.24786827e-06 6.06863602e-06]
[ -3.00204048e-07 2.27547278e-06 1.21009590e-05 -1.46010147e-06 -2.82182506e-05 -5.72413627e-06]
[ 3.86495707e-07 2.19194185e-07 6.00140583e-07 5.22173017e-07 -5.07637782e-07 -1.80921288e-05]
[ 8.65238244e-07 1.40265800e-07 -7.73394328e-07 -1.11460940e-05 5.55361952e-07 5.95215636e-06]
[ -2.11897776e-07 1.87788983e-06 -6.36450620e-06 -4.18394781e-07 6.80973263e-07 1.43993091e-06]
[ -3.97586817e-06 8.56673764e-07 6.23090369e-07 6.77817038e-07 -2.05600919e-06 1.04563676e-06]
]
```

Answer: Each vector seems to represents a different buying pattern, where the coefficients in the vector represent above or below average values.

1st component:

Low consumption of **Milk and Detergents Paper**

High consumption of **Grocery and Delicatessen**. This customer might be an old guy with no children’s (low Milk) and lot of money (Delicatessen). (Is this what you mean by “in terms of customers”? If not please be more specific about what the questions asks)

2nd component:

Low consumption of **Detergents Paper and Delicatessen**

High consumption of **Grocery and Milk**. This customer might be families (Grocery) with babies (Milk) with no much money to spoil themselves (low Delicatessen).

3rd component:

Low consumption of **Detergents Paper and Delicatessen**

High consumption of **Grocery and Frozen**. This customer might be an family (Grocery) with adult kids (Frozen) and again no much money to spoil themselves (low Delicatessen).

4rd component:

Low consumption of **Grocery and Frozen**

High consumption of **Fresh and Delicatessen**. This customer might be an office, a start up company or a restaurant (Fresh), as they seem to spend a lot of money on good things (delicatessen).

What could these components be used for? From that we can understand that the data include different type of customers. More than that, we (or the ICA) determined several types of customers in the data set. We can use this information to aggregate similar customers into groups and label these groups. This information can be useful when we want to determine how many independent components encompass in the data set. For example we can decide about the number of clusters for k mean analysis.

Clustering

Decide on K means clustering or Gaussian mixture methods

- What are the advantages and disadvantages of each?
- How will you decide on the number of clusters?

For that section I will use the two algorithms because I want to compare the results. Below I discuss the benefits and flaws of each algorithm.

My choice is to use the **Gaussian Mixture Models,** the reason for that is the fact that the algorithm maximizes only the likelihood and therefore do not assume any specific structure, where K means have several different assumptions about the structure of the data. In this specific case we cannot assume much about the structure of the data, mainly because we do not know the data. The data include different type of customers that might have different size or buying patterns.

Gaussian Mixture Models:

Advantages

1. The fastest algorithm for learning mixture models

2. The algorithm maximizes only the likelihood, therefore it will not bias the means towards zero, or bias the cluster sizes to have specific structures that might or might not apply.

Disadvantages

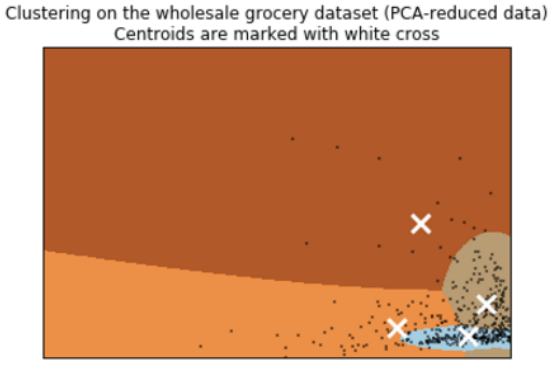
1. When one has insufficiently many points per mixture, estimating the covariance matrices becomes difficult, and the algorithm is known to diverge and find solutions with infinite likelihood unless one regularizes the covariance artificially.
2. This algorithm will always use all the components it has access to, needing held-out data or information theoretical criteria to decide how many components to use in the absence of external cues.

Implement **Gaussian Mixture Models**

1. Sample central points of the clusters

Produce a graphic

2. Visualize important dimensions by reducing with PCA
3. Are there clusters that aren't very well distinguished? How could you improve the visualization?



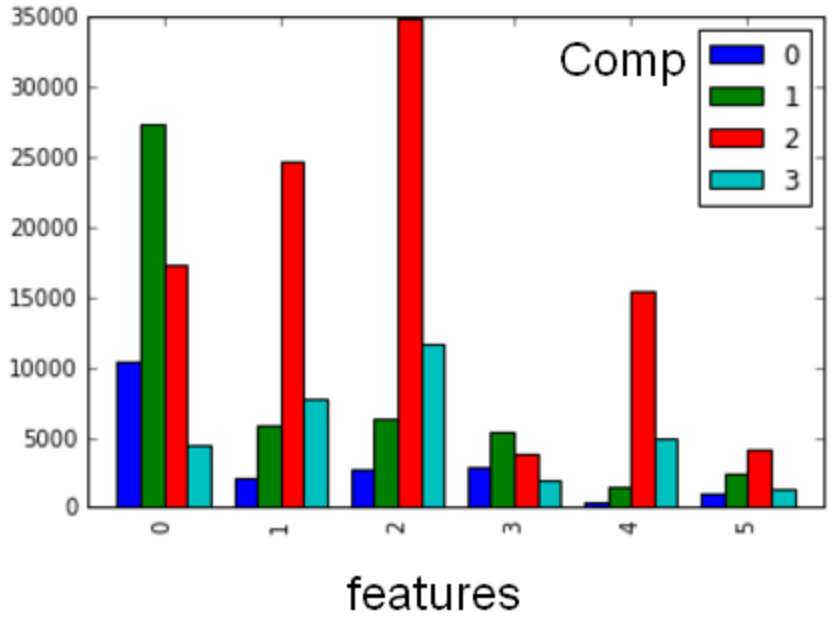
We can see clearly 3(on the left) or 4(on the right) clusters in the figure. The most prominent two clusters are on the bottom right of each figure. The other two clusters (or one cluster for the case of 3) include the data that is more sporadic and it is not clear if that is a real cluster or costumers with no buying pattern.

Sampling: I preformed the inverse transform of the PCA (with 2 components) on the centroids obtained from the GMM with 4 components. That model seems to separate the data in the best way.

The output look like,

'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents Paper', 'Delicatessen'

```
[ 10461.84724846,   2046.24453768,   2668.34762352,   2841.7401855,    442.3560535 ,    982.08245096],
[ 27323.73752904,   5930.11445253,   6341.19086536,   5467.66889137,  1553.81949605,   2377.24316943],
[ 17386.64381576,  24709.29701744,  34874.96082654,   3861.86755204, 15403.00120879,   4132.94357528],
[  4385.95277825,   7725.784475  ,  11659.75895933,   1875.94833342,  4915.38234541,   1345.55492231]
```



As we can see there are 4 different clusters but only 2 different patterns. Components 0 and 1 seem similar and Components 2 and 3 also seems similar. The first component represents customers that buy 'Fresh' and the second component represents customers that buy a combination of mostly Grocery and some milk (and even less from the other products). The interpretation of that is that the data can be separated into 2 type of costumers as describe here above by the 2 centroides.

K Means clustering:

Advantages

1. The problem K means try to solve is computational difficult but there are efficient and fast heuristic algorithms.
2. Simple to implement and run.

3. Easy to understand, therefore might helps to get a better understanding of the problem.

Disadvantages

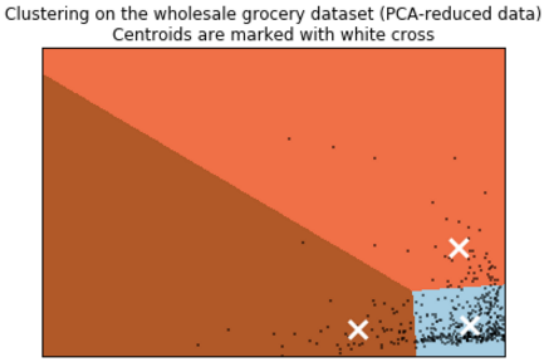
1. The Algorithm might convergence to a local minimum that might be wrong result.
2. The numbers of clusters are given as an input, wrong number might cause poor results, and more or less clusters there really are in the data.
3. The algorithm assumes separable, spherical and similar size clusters that are separable, that might cause a failure to classify when data do not satisfy the assumptions.

Implement **K Means clustering:**

1. Sample central points of the clusters

Produce a graphic

2. Visualize important dimensions by reducing with PCA
3. Are there clusters that aren't very well distinguished? How could you improve the visualization?



For the case of 4 clusters we find similar pattern as we reveal for the Gaussian Mixture Models. However, for the 3 clusters we find different separation, where the lower right corner includes only one cluster. I find this very interesting. It seems that the many assumptions about the data structures for the K means are responsible for this difference.

I find the visualization quite clear, mainly because we are not interested in small details. But we can easily improve the visualization using a log scale or simple zoom in.

Conclusions

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

Answer: This is difficult to answer because each algorithm gave a different perspective about the data.

The **PCA** provide us the ability to find general trends in the data and different variance patterns. We use this information to simplify our data set by removing unwanted or not relevant patterns.

The **ICA** provides us with different type of customers or buyer types. Using this information we could understand better what type or how many types of customers the data set might encompass.

The **clustering algorithms** used the simplified data to separate the customers to different classes and label each customer. This again a very important step in the analysis because we can use this information get plan a better experiments and perhaps even build a prediction model.

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

Answer: The classification reveals additional information about the data set. The buyers does not distribute homogeneously, instead there are different patterns of buying. Fortunately enough we could detect these patterns and labels each client accordingly. One way for a new experiment will be to choose a 'test group' within each label, for example 20% from each class. For this 'test group' we can change small number of features, for example delivery time. That will enable us to compare the response to the change in the 'test group' and separately in the control group (the other 80%). The results can be different in each class, but then we can decide for each class if the change is beneficial or not.

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

Answer: For prediction it is important to have some labels, this data does not include labeling. Other than the customer ID we do not really know what class or type is each customer. One way to overcome this will be to use as labels from the clustering analysis. Obviously the edges of the clusters might be problematic and might introduce error to the training. However, we can choose data points (or customers) that are closer to the center of each cluster and remove the data points on the edges.

The training on these groups will provide us with a model. The model can be tested on a separate test sample (that will not be used for the training procedure).If the model performance is high enough we can use this model for prediction.

Reference

http://wikipedia.org/

http://stackoverflow.com/

http://scikit-learn.org/