Expectation Maximization for Wholesales data

An academic essay on clustering

Eduardo Sánchez López
Escuela Superior de Informática
Universidad De Castilla La Mancha
Ciudad Real España
eduardo.sanchez00@outlook.es

Maira Torres Medina
Escuela Superior de Informática
Universidad De Castilla La Mancha
Ciudad Real España
maria.torres@alu.uclm.es

0 Introduction

This essay is the result of an academic study using a wholesale customers data from an unknown distributor. The objective is to find business insights that could prove useful for the company through unsupervised learning. Not all clustering techniques are valid for this model, for academic purposes it was limited to only Gaussian Mixture Models using Expectation Maximization.

1 Milestone 1

Expectation Maximization does not give a good result with this dataset. Once the clients that spend more were deleted due to them being labelled as outliers, the score lowered more, probably because EM tried to divide the dataset into big and small spenders. In order to give a result that could be useful for the business, the pareto principle was followed to create two groups, one with the top 20%, and one with the bottom 80%.
This approach was unsuccessful since the 20% only represents the 42.9% of the total sales. The pareto rule didn't follow in this case. It was calculated that instead the 58% represents the 80.64% of the total amount of sales, far away from the pareto principle.
For the milestone 2 there was a problem, the clustering algorithm didn't perform well and the pareto principle division was unsuccessful due to the failure in proving the principle in the dataset. To solve this and carry on with the study, the dataset was divided into 6 groups, the 6 possible combinations of channels and regions.

For more information regarding the development for milestone 1 see annex 1.

2 Milestone 2

For this milestone the groups that were obtained in the first milestone conserve all the data points that were dropped during the process. Since the main dataset is divided into 6 depending on both the region and channel, those columns are dropped.

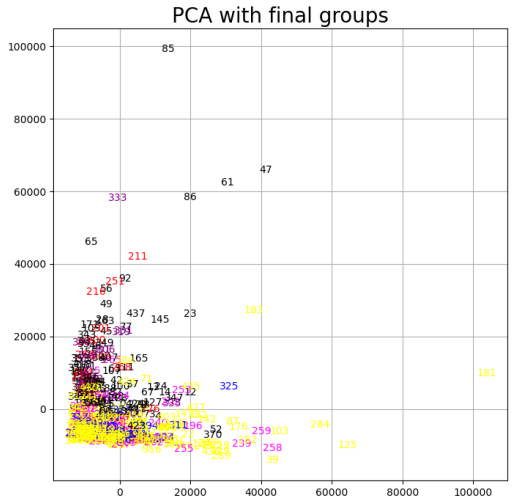

Figure 1 PCA plots with the final groups

This PCA plot shows that the groups are mixed between them, without a clear distinction.

2.1 Representative points

The metric to select the representatives of each group is the sum of the SSE of each feature with the mean of said column in that group dataset. With that calculation done in every sample for every dataset, the one that has the lowest value is the one selected. Every dataset needs to be scaled prior to this, to avoid that features that move bigger quantities make a bigger impact that the other ones.
The representatives are the rows with index **225**, **337**, **404**, **200**, **306** and **9.**

2.2 Value of each group for the company

The most direct metrics for this problem is the total amount of money recollected by each group and the number of clients in each group. The following table expresses this data. In the group name, the letter C represents the channel and R the Region of the group.

	C1R1	C1R2	C1R3	C2R1	C2R2	C2R3
Expenses	1538342	719150	5742077	848471	835938	4935522
Clients	59	28	211	18	19	105
Expense/Clients	26073.59	25683.92	27213.63	47137.27	43996.73	47004.97

Table 1 Group data economic information

The first thing to note is that, as stated in previous sections, the region is not a factor that influences the consumer habits, while the channel in which the client operates is. Note that the second channel represents roughly a third of the total data, but the clients on there spend almost 20000 more on average than those on channel 1. These 6 groups could be summarized into 2 without losing information or potential insights, but since this is an academic study, having a higher number of groups is more interesting in order to continue with the statistical tests.

2.3 Statistical tests

One requisite for the correct selection of the statistical test for this model is check whether the data is related or independent. Assuming that each customer does not have any relationship with the other ones, the data samples for this problem are considered independent and a Kruskal test will be performed. Friedman was taken as a consideration, but since every group has a different number of rows, Kruskal test works without having to make any more transformations. It is said that Kruskal works better with a number of measures per samples bigger than 5, which is the case for this model.

The null hypothesis is that the population median of all the groups are equal. **H0**

The alternative hypothesis is that there are two groups *u* and *v* that does not have the same median. **H1**

With a significant level of 5% the results are as follows.

KruskalResult(statistic=224.95796032866366, pvalue=0.22796077700989464) – **Fresh**

KruskalResult(statistic=162.9440796407792, pvalue=0.9930732001855793) – **Milk**

KruskalResult(statistic=169.2146703117387, pvalue=0.9822571408738162) **– Grocery**

KruskalResult(statistic=220.35032371008182, pvalue=0.2982697881729335) – **Frozen**
KruskalResult(statistic=150.22781320466325, pvalue=0.9993534425258216) – **Detergents_Paper**

KruskalResult(statistic=204.73760665805432, pvalue=0.5895351851018622) – **Delicassen**

KruskalResult(statistic=213.4679706120267, pvalue=0.42043141497627445) – **Total**

The p-value is in every case higher than 0.05, in some cases even being almost 1, the null hypothesis is accepted for all the features.

The default assumption that all data samples were drawn from the same distribution is valid. This means that the groups that are formed are not statistically different from one another.

3 Conclusions

There are two main groups, one for every channel. Channel 1 represents retail markets while channel 2 are restaurants and hotels. For the business the ones that pay more in average are in the second channel, which means that maintaining those clients and attracting more with the same type of business is important. Nonetheless, there are almost as double the number of clients in the retail market. They spend less money than their counterparts, but this information can also be valuable. Perhaps they don't consume more because the offers for them are not as interesting, maybe they are smaller retail markets and the company should focus on bigger ones now that it has a strong baseline.

The 6 groups formed at the end were only for academic purposes to continue with the methodology that was proposed. Another possibility to continue with this study was to separate the data from channel 1 and channel 2 and perform a clustering analysis in each one separately. Due to document size constraints it wasn't followed, but it could throw new insights about the different types of customers of each channel.

REFERENCES

[1] https://stats.stackexchange.com/questions/371333/is-it-important-to-make-a-feature-scaling-before-using-gaussian-mixture-model

All the code can be found in
https://github.com/Edusanc95/SupermarketCluster

Annexes

1 Milestone 1

The first step is finding out which variables are the most appropriate to carry on with the study. For that, let's have a peak at the dataset features and the look of the data that they carry.

Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
2	3	12669	9656	7561	214	2674	1338
2	3	7057	9810	9568	1762	3293	1776
2	3	6353	8808	7684	2405	3516	7844
1	3	13265	1196	4221	6404	507	1788
2	3	22615	5410	7198	3915	1777	5185

Table 2 First 5 data samples

Each row represents certain data about a customer. Each row has columns representing certain variables or features.

Channel and **Region** are categorical features that specifies, as the name says, the channel and region in which those sales were done respectively. Each region and channel have a code assigned to it in order to be able to use data analysis algorithms more easily. The rest of the features, being **Fresh**, **Milk**, **Grocery**, **Frozen**, **Detergents and paper** and **Delicassen** contain continuous numerical data about the shopping habits of the client in question.

Semantically, one can think that the channel and, specially, the region feature are not useful for the problem at hand. This could or could not be true, especially since the goal is to obtain knowledge for the business. As an example, this hypothesis can be rebutted by finding that certain regions spend more on a certain type of product due to the circumstances of their location. Maybe the providers at region 3 spend considerably more on delicassen products than their counterparts on regions 1 and 2. This newfound knowledge could indicate that there are a lot of wealthy people in that area willing to spend money on delicassen products, and with that a possibly great number of retail markets that could be interested in new deals with the company to distribute the product. This is only an example to argue that semantic conclusions such as "it does not matter the origin of the clients" should also be accompanied by data that corroborates that.

1.1 Data cleaning and transformation

Before doing feature, selection or working with any algorithm, there is a need to check that the data that is being handled is ready to be used. Checking for null values, transforming the data structure to be ready for data analysis and evaluating possible scalation are some of the steps that need to be carried.

Data structure

As stated before, region and channel features are already transformed from categorical data to discrete numerical data assigning each region a channel a code. With this transformation there is a problem in that since now the values go from 1 to x, this increment could be misinterpreted by the data analysis and machine learning algorithms. This range of values could be understood as a linear progression and create results that are not as reliable as they should be. To solve this, a new feature with format "is_x" is created for each region or channel. This feature can only take 0 or 1 as a value. 1 means that the attribute that the feature describes is true and vice versa for 0. After that, Channel and Region features are deleted from the dataset. Note that since there are only two possible values for the Channel feature there is no need to do this change. For a better semantic understanding and to be consistent with the changes done to the other categorical data feature, this transformation is performed in the channel feature as well.

Scaling

For GMM there is no need to standardize the data. Gaussian Mixture Modelling explicitly relaxes both the assumption of all clusters having the same variance, and the assumption of no correlation of features within a cluster, and that's why there is no need to standardize the features [1]. This solves the problem on losing data weights by standardizing.

1.2 Feature selection

Correlation heatmap

With all the pre-processing now done, the next step is checking the relevance of all the features for the process. Some characteristics can be less important or even irrelevant.

As stated before, the most suspect variables are Region and Channel. Let's see how correlated are all the variables between them. The ideal result is having correlated variables, but not too much. If two variables have a very high correlation it means that there are possibilities that they add the same information to the model. Having a high number of variables increases the possibility of adding uncertainty and noise to the model input, so if two behave in almost an identical way, one of them could go away after additional data explorations if needed. However, in this problem there are already a low number of variables, so this is not as important as making sure that all the features are useful. The threshold is >0.95 or <-0.95. If it's surpassed, a deeper analysis would be carried to check if some characteristic must be dropped from the dataset.

A feature having no correlation or low correlation with all or most of the rest is also a case to study. It could mean that it is an isolated characteristic that is not useful in the model. The threshold is between -0.05 and 0.05.

With these bases settled, the following correlation heatmap was created (Figure 1).

As suspected, we can see that the region features are almost no correlated with all the other ones. Specifying the region does not matter for this problem and one can safely assume that the origin of the customers is not a factor in their consumer habits. The channel variables have a degree of relationship between other variables, so they seem useful for the model.

Note the high correlation between **Detergents_paper** and **Groceries**. While being over 0.9 it does not mean that they offer the same information, it only shows that they are usually bought in similar proportions, there could be room for their own impact on the model.

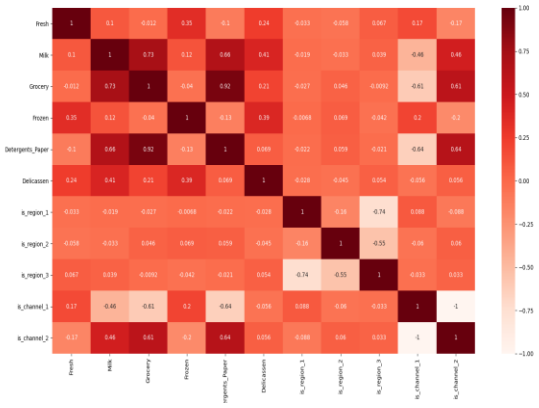

Figure 2 Correlation heatmap of the features

PCA

After the feature selection, let's look at the data with a PCA with two components projected over a 2d plot. The sum of the variances is about 0.85, which seems good enough for data visualization.

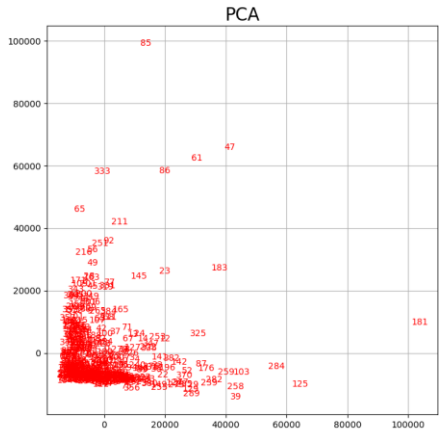

Figure 3 PCA plot of the dataset

This data shows a big blob of customers and several points that are spread out. The number represents the index of the customer in the dataset, which will help to identify outliers. Before clustering let's plot again, but now assigning each variable belonging to channel 1 a magenta color and a blue color for those on the channel 2.

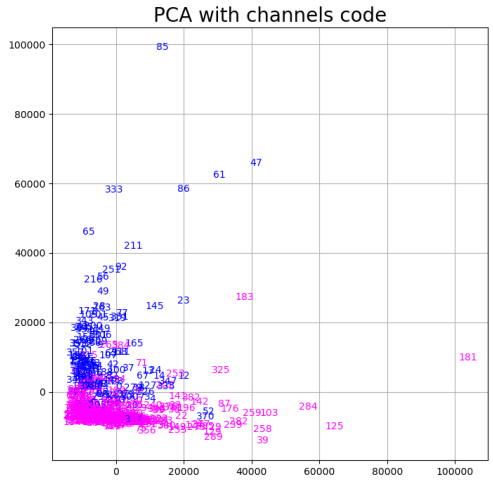

Figure 4 PCA plot discriminating between channels

As shown, the two groups converge on the bottom left, but they grow into different directions. The magenta group stretch to the right and the blue group to the top of the plot. This indicates that while most of the points converge into a common section, each channel has a characteristic or set of characteristics that behaves differently from the other one. Before continuing with this idea, let's apply Expectation Maximization and observe what is the result.

Clustering

The clustering process will be divided into multiple cycles. Each cycle will contain both parametrization and an algorithm execution with conclusions attached to them. It's possible that new adjustments are extracted from those conclusions. In that case, a new cycle would start with those suggested changes until the conclusions obtained are final.

1.3 First cycle

Parametrization

The silhouette score is the metric that is used to determine the number of components of the EM algorithm. For that let's plot the score from using 1 component up to 30 components.
There are two conclusions that can be extracted from the graph:

• The best value for the number of components is two.

• For the current data, the maximum silhouette score is low at around 0.4. This could indicate that the algorithm being use is not suitable for this type of data or that the data itself is hard to work with in a cluster. There are indications that there are some outliers by looking at the plots of the past section, they could be the reason behind the poor result.

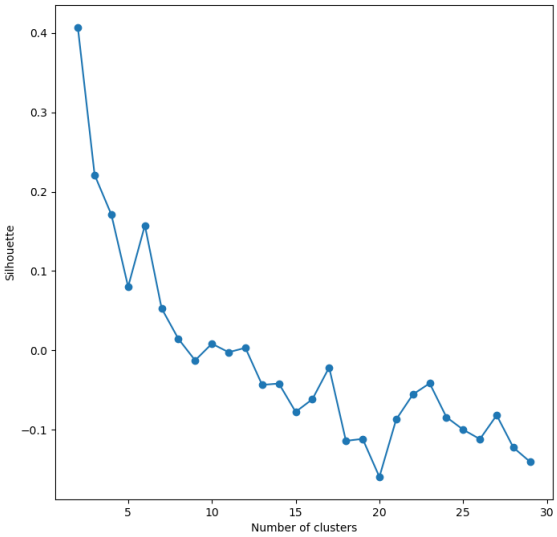

Figure 5 Silhouette score according to the number of components

Algorithm execution

With the number of components decided, the following plot is a representation of the dataset colored by the cluster they belong.

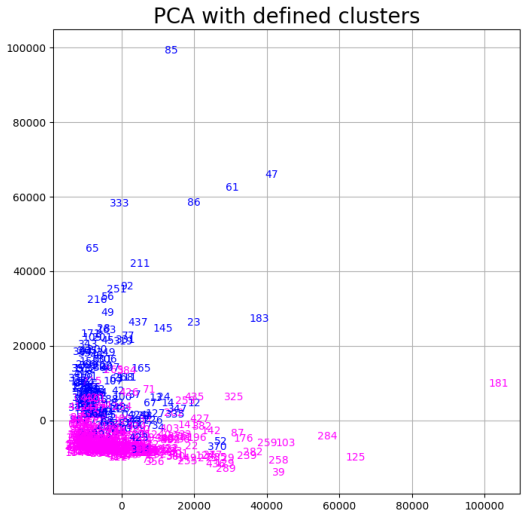

Figure 6 PCA plot discriminating between clusters

This graph is certainly like the one which colored the variables according to which channel they belong. A quick calculation shows us that there is over a 99% coincidence that the values of the channel 1 belong in the first cluster and vice versa with channel 2. This result is obtained no matter how many times the code is executed since the algorithm fits the dataset 300 times searching for the solution that has the higher likelihood.

This high coincidence could mean that the is_channel features are skewing the data and that all the other characteristics are not used for the determination of the clustering. This could happen because the easiest way to divide the dataset into two groups it's by discriminating by the channel. This solution has a high probability that's the one that gives the most likelihood to the model since there's a variable that directly correlates to the probability that the data belongs to this cluster.

In fact, looking at the frequency of the values in the clusters likelihood matrix there is only two values, 0 or 1. For an algorithm that is based on the probability that one data point belongs to a cluster, having only "yes" or "no" values is unusual. After this initial analysis, the conclusion was to repeat the clustering process without the channel features.

1.4 Second cycle

Parametrization

Same process as the first cycle.

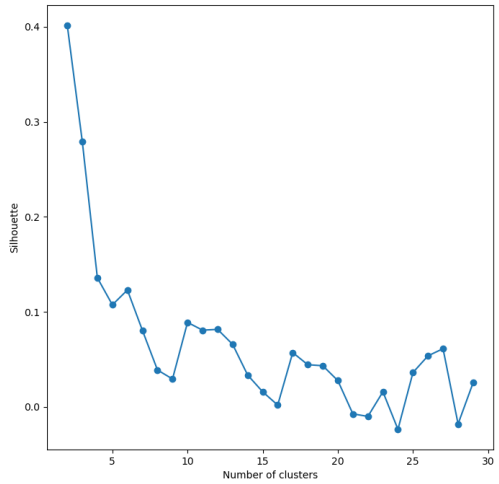

Figure 7 Silhouette score for the second cycle

Again, best number of components is two. Note that the silhouette score is still low.

· Algorithm execution

Now the amount of coincidence between the clusters and the channels is down to 0.7, denoting that the consumer habits of each channel tend to be different, but not in all of them. The likelihoods are still one-sided, but now there are some elements that EM considers that don't belong 100% to either cluster. This is plotted in the figure (Figure 7) in which the pink points belong to the first cluster 100% and the celeste ones to the second cluster. The ones that are not as likely are plotted with a variation of those two colors. All of them are near the big data blob in the bottom right.

Outliers

By the multiple plots done in this document one can see that there are some of possible outliers. Those are the customers number **333**, **86**, **61**, **47**, **85** and **181**. It does not mean that there could not be more outliers or that even they are outliers, after all this is done by eye. If there are more outliers, it will be studied in the next cycle.

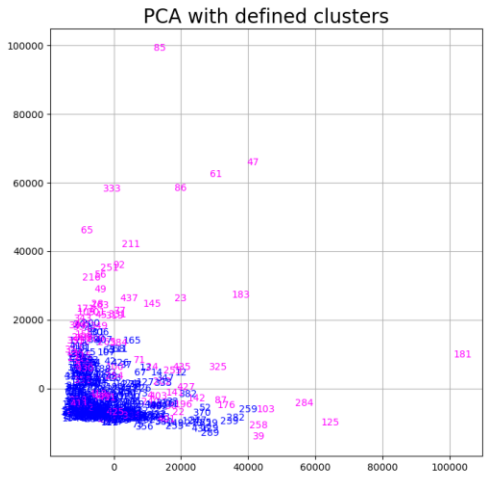

Figure 8 PCA plot of the second cycle

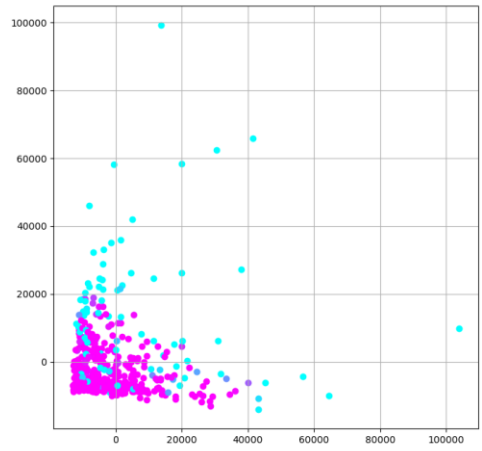

Figure 9 PCA plot of the second cycle with the likelihoods of each point

To make sure that they are outliers, let's look at some basic data of each of the features and the total amount spent, such as the mean, median, percentile 90 and standard deviation.

	Mean	Median	P90	Standard deviation
Fresh	12000.29	8504	27090.50	12647.32
Milk	5796.26	3627	12229.90	7380.37
Grocery	7951.27	4755.50	18910.10	9503.16
Frozen	3071.93	1526	7545.30	4854.67
Detergents_Paper	2881.49	816.50	7438.30	4767.85
Delicassen	1524.87	965.50	2945.90	2820.10
Total	33226.13	27492	57818.70	26356.30

Table 3 Basic data of each of the features

And the data of each point.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Total
333	8565	4980	67298	131	38102	1215	120291
86	22925	73498	32114	987	20070	903	150497
61	35942	38369	59598	3254	26701	2017	165881
47	44466	54259	55571	7782	24171	6465	192714
85	16117	46197	92780	1026	40827	2944	199891
181	112151	29627	18148	16745	4948	8550	190169

Table 4 Second cycle outliers' information

We can see that all of them surpass the percentile 90 of the total amount spent and that the points that are the furthest from the data agglomeration are the bigger spenders. The points 86 and 333 spend less and thus are close to the blob. The element 181 is the one that's furthest from all the data points due to not only having big expenses but because it spends more than 5 times the percentile 90 in fresh, which is a pattern that no other customer follows. Similar thing happens with the point 85, which spends almost 6 times than the percentile 90 in detergents and paper. One thing that all of them have in common, apart from 181, is that they spend more than the percentile 90 in groceries. 181 is from channel 1 while all of them are from channel 2, so it could be a pattern that belongs to the channels.

All these outliers are extracted from the dataset and the analysis continues with the third cycle.

1.5 Third cycle

Parametrization

Method unchanged from previous steps.
Note that the score is even lower than before. That indicates that the deleted elements were important for the differentiation into groups. One possibility is that this dataset is divided into two type of buyers, and they are not divided into which channel they belong to, but by the amount of money that they spend.

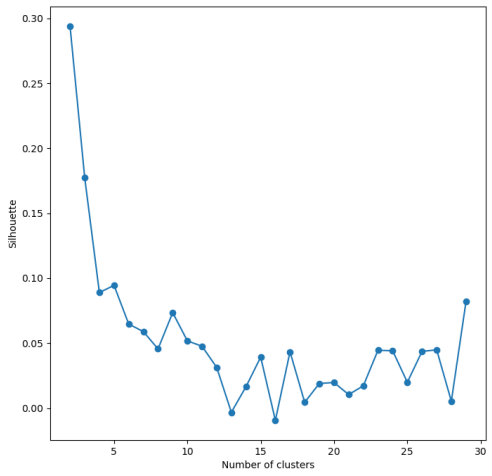

Figure 10 Silhouette score for the third cycle

Algorithm execution

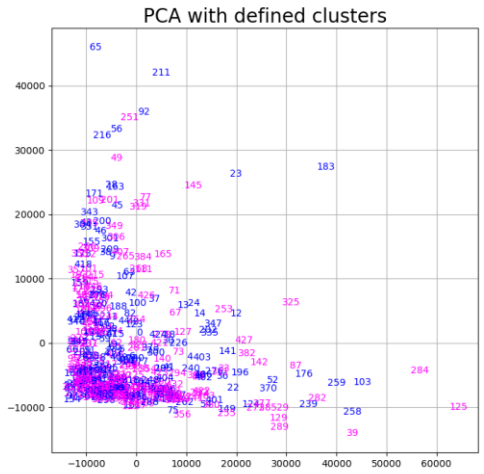

Figure 11 PCA plot of the third cycle

The PCA plots show that the two clusters are mixed in the dataset agglomerate. The quality of this PCA with 2 features is quite high, as demonstrated in a previous section, so it looks like the cluster division is getting worse and starts to feel like a random selection.

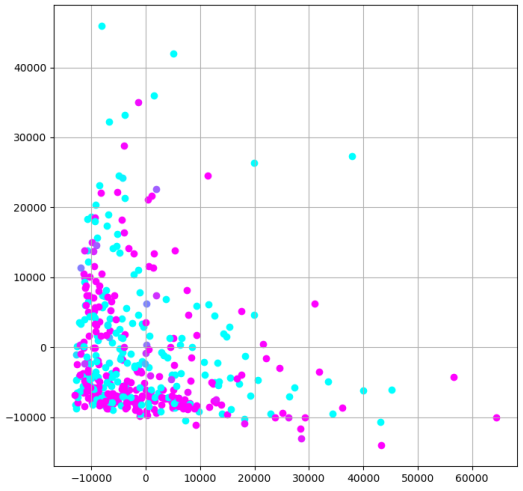

Figure 12 PCA plot with likelihoods of the third cycle

The points that are getting selected for a possible outlier selection are **65**, **221**, **183**, **284** and **125**.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Total
65	85	20959	45828	36	24231	1423	92562
221	5396	7503	10646	91	4167	239	28042
183	36847	43950	20170	36534	239	47943	185683
284	68951	4411	12609	8692	751	2406	97820
125	76237	3473	7102	16538	778	918	105046

Table 5 Third cycle outliers' information

Except for sample 221 which has values not near the percentile 90, the rest of the samples behave like the already identified outliers. Most of them spend less than those outliers, which is expected because they are closer to where the rest of the data is.
The next cycle would proceed with the drop of rows 65, 183, 284 and 125. If the silhouette result is lower, the clustering process will stop.

1.6 Fourth Cycle

Parametrization

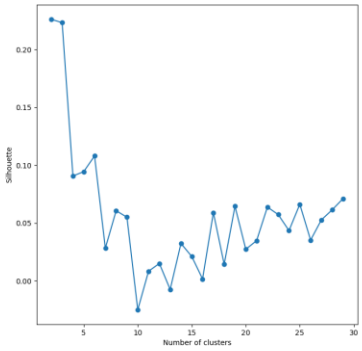

Figure 13 Silhouette score for the fourth cycle

With a silhouette less than 0.3 for the optimal number of components, which is lower than in cycle 3, it is now demonstrated that dropping what could be considered outlier samples is worse for the model. The gaussian mixture is probably trying to make the division between big expenders and the rest of customers. Since more and more big expenders are being deleted from the dataset, it has a harder time finding a way to separate the data into more than one big cluster.

Notes

• No formal outlier methodology was followed during the clustering process because it was deemed as unnecessary since the clustering algorithm that was imposed wasn't giving good results. This was even clearer once the first set of obvious outliers were dropped, when the silhouette scored lower each time it was done. If the groups were going to be created using a clustering algorithm, a statistical way to identify outliers would be followed.