Clustering in Asset Management: an Original Way to Visualize Client Behaviors

1 Intro

Asset management consists in managing investments on behalf of others. To provide the best service, it's crucial for asset managers to understand clients' behaviors. That's why it's common in private banks to look for patterns within clients' data. Clustering is a common method to achieve this. However, one of the key challenge with such approach is to display and interpret the results. In this article I will briefly give an introduction to clustering and most of all, show a simple and original way to display the results. Below is an insight of the final results.

2 Example

We use a toy example to make the method easier to understand. Say we have a bank with 10 clients. We want to focus on the main characteristics of those clients. In this example we consider only 4 aspects (as you can imagine, this is very simplified):

- Wealth (also called asset under management)
- Number of unique financial instruments
- Number of transactions
- Age

3 Clustering

Clustering is about grouping clients based on similar characteristics. There are numerous ways of performing such tasks and discussing those different methods is not in the scope of this article. Let's use a common clustering method, the Gaussian Mixture. In a nutshell, this method allows to estimate the density of each cluster. For more mathematical details, check out my notes here.

4 Validation

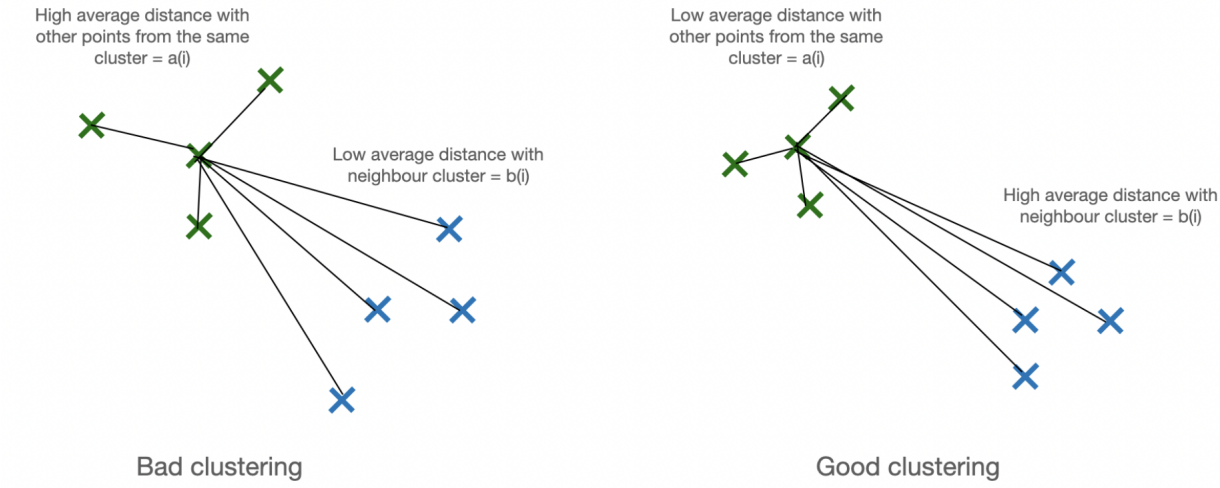
The tricky part of any clustering method is to make sure the clusters are of good quality i.e. compact enough. We call this process the validation step. Validating our results is not the main

topic of this article but I need to dig in few notions so that the visualization part will be clear. In order to validate our model, we need to find a metric that represents the cluster quality to be optimized. The chosen quality metric is the silhouette coefficient i.e. the ratio of the distance intra/inter clusters. It measures how compact is the cluster.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Where:
$a(i)$ = average distance of client i with the other points from the same cluster.
$b(i)$ = smallest average distance of i with the other points from the closest cluster.

The below picture helps to understand better what those distances mean exactly.



The higher the silhouette coefficient, the better is the cluster quality. We can thus test different parameters and chose the combination that gives us the best silhouette coefficient.

Note: the silhouette coefficient is NOT adapted if you use a clustering method that can identify non convex clusters. For more information, I recommend the scikit learn's user guide.

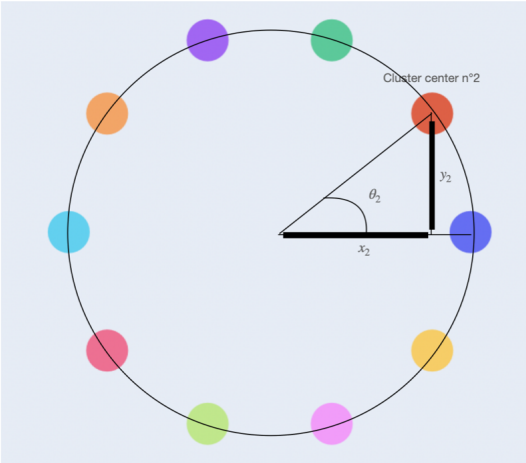
5 Visualization

Now that our clusters are well defined, how do we visualize the results? More specifically, how to communicate our results to the rest of our colleagues in a meaningful way? When researching on the topic on how to best perform cluster visualization, I came across many methods that involve dimension reduction. In short, it consists in transforming our problem to a 2D case. That way we could display results that humans are capable of reading graphically. The problem with this transformation is that **we lose part of the information** contained in the data. One way to overcome this problem is to display only the density. After all, when performing a clustering task the key questions are:
- How precise are the detected patterns?
- How good an observation (i.e. a client) "fit" into a specific cluster?

Displaying the density allows to answer those questions. We will build a solution based on the silhouette coefficient described above.

Step 1: define centers for each cluster

First, we define the center coordinates for each cluster. We want to make sure each cluster is distinct so that it's readable enough. To do so, we find the coordinates of the points that are located on a circle (radius 1) and are equidistant. If we would end up with 10 clusters, the centers would be separated as such:



The coordinates are found using the cosinus/sinus of thetas that correspond to the equidistant points on the circle. The full details can be found in the code [LINK TO BE ADDED].

Step 2: generate coordinates for each cluster

Each cluster's observations are generated with a normal distribution where the center is the center of the cluster from the previous step and the standard deviation is the quality of the cluster. For each cluster, we find the coordinates x and y:

$$x \sim \mathcal{N}(\mu_x, 1 - silhouette)$$

$$y \sim \mathcal{N}(\mu_y, 1 - silhouette)$$

where μ_x is the first coordinate of the cluster center found in the previous step. The silhouette is the coefficient corresponding to the quality of the cluster.

You may ask yourself: why do we generate data? and why do we use the silhouette as standard deviation? We generate data since we focus on density (as explained above), that is, how close are observations with each other. The standard deviation represents the compactness of our clusters - the higher the standard deviation, the less compact are the clusters. We can thus use the opposite of the silhouette coefficient.

Step 3: compute the distance to center

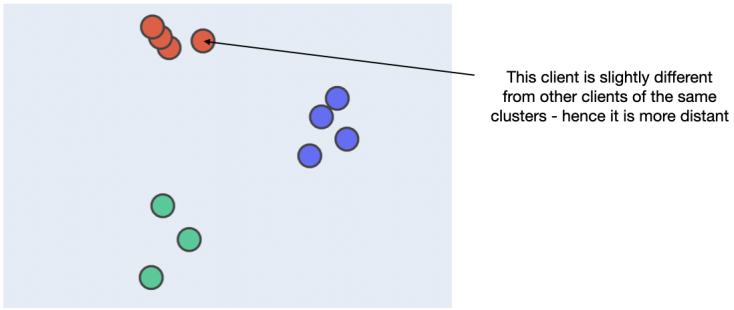
We compute the distance of each generated point to the center of the cluster. A simple Euclidean distance would work fine.

Step 4: map each point to the right generated point

Finally, we sort the data based on their distance to the center. We map observations so that, for each observation, the distance to the cluster center is proportional to silhouette coefficient. That way, we penalize observations that seem to fit poorly to the cluster they belong to (based on the silhouette coefficient).

6 Implementation

For the implementation I use my favorite graphical library: Plotly. It's well maintained and highly customizable package that I highly recommend. From this figure, we can see that three clients from the red cluster seem to have very close behaviors since the dots are very close to each other.



On the contrary, the fourth client of the red cluster fits less good in the cluster. Indeed, the client is the only one in the cluster with a large number of trades.

7 Conclusion

In this article, I have presented a method to display cluster results without reducing dimensions. The method focuses on density since it's the most important information anyone would look at. Using such method, asset managers can quickly see whether there are large/small groups of clients behaving similarly as well as identifying clients with clearest patterns.

8 References

Clustering GMM (1) - https://en.wikipedia.org/wiki/Mixture_model

Clustering GMM (2) - the following blog gives quite a good explanation of the concept: https://jaketae.github.io/
mixture-models/

Silhouette coefficient - https://en.wikipedia.org/wiki/Silhouette_(clustering)

Plotly - https://plotly.com/